

Behavior As A Modality: A Framework To Enable Automated Persuasion

By

Yaman Kumar
PhD19024

A dissertation

submitted to the

faculty of the Graduate School of

Indraprastha Institute of Information Technology

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

Department of Computer Science and Engineering

Indraprastha Institute of Information Technology Delhi

July 2025

Copyright

Yaman Kumar

2025

All Rights Reserved

THESIS CERTIFICATE

The work contained in this thesis entitled, **Behavior As A Modality: A Framework To Enable Automated Persuasion**, has also been submitted to the PhD program of the Indraprastha Institute of Information Technology, Delhi. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.



Yaman Kumar





Supervisors:

1. Dr. Rajiv Ratn Shah, Associate Professor, Department of Computer Science and Engineering, IIIT Delhi
2. Dr. Changyou Chen, Associate Professor, Department of Computer Science and Engineering, State University of New York at Buffalo

Committee Members:

1. Dr. Dinesh Manocha, Paul Chrisman Iribe Professor of Computer Science and Electrical and Computer Engineering, Distinguished University Professor, University of Maryland
2. Dr. Nalini Ratha, SUNY Empire Innovation Professor, Department of Computer Science and Engineering, State University of New York at Buffalo
3. Dr. Aniket Bera, Associate Professor, Department of Computer Science, Purdue University

Place: New Delhi, India and Buffalo, USA

Date: Jul 2025

Preface

We're actually much better at planning the flight path of an interplanetary rocket (rocket science) than we are at managing the economy, merging two corporations, or even predicting how many copies of a book will sell (behavior prediction). So why is it that rocket science ***seems*** *hard, whereas problems having to do with people - which arguably are much harder - seem like they ought to be* ***just*** *a matter of common sense (easily predictable)?* - Duncan J. Watts

Also,

If the brain were so simple we could understand (predict) it, we would be so simple we couldn't. - Emerson Pugh

But,

Nothing in Nature is random (unpredictable). A thing appears random only through the incompleteness of our knowledge (ignorance). - Baruch Spinoza

While,

Ignorance is bliss. - Thomas Gray

but,

Timendi causa est nescire. (Ignorance is (also) the cause of fear.) - Seneca

And,

What would life be if we had (only fear and) no courage to attempt anything? - Vincent Van Gogh

As a computer scientist working on problems related to human behavior, I am often asked why I chose this particular domain. The questions come from various perspectives - whether such problems are better suited for psychologists and marketers, why these problems are interesting at all, whether human behavior contains too much randomness to be mathematically tractable, if the problems are ill-defined, and if they are even objectively solvable. For the ones interested in the

art of diction, I try to lay out my motivations in the quotes above. But for others, I try to explain through a simple narrative and answer the questions more scientifically in Chapter-1.

Behavior is unsolved. Let me tell you a little story. Kelin is an eager advertiser who releases a campaign on Facebook one Friday evening, paying $1000 to run an ad across California. With each launch, she silently sends a prayer that the ad resonates, draws clicks, leads to purchases of her product, ensures her campaign's success, and lands her the long-awaited promotion. Come Monday morning, Kelin witnesses human behavior in all its varied glory (within the platform's constraints, obviously): she has received 28 comments on her post, 867 likes, 9045 views, 349 clicks, and 28 purchases. Satisfied but seeking improvement, she tweaks a few words she feels might better appeal to Californians and relaunches. This time, her metrics jump by 10.8%. Puzzled by the reasons, but pleased with the outcome, she presses on.

From my perspective, Kelin and countless others like her are replicating what the pioneering botanist Gregor Mendel did in the 1800s. The difference is that the subjects for Mendel were peas and for Kelin, it is humans. Mendel was trying to solve the puzzle of why some pea plants are tall and some small, some green while some yellow, and some pea seeds round while some wrinkled. The modern-day Kelins are trying to solve what makes people click, comment, like, and purchase, why certain words perform better in California while others in Texas, how behavior can be modified, and so on. Mendel's laboratory was his 2-acre Moravian monastery farm. Kelin's laboratory is the digital landscape of Facebook, Twitter, YouTube, TikTok, Google, and her website.

Before Mendel, the general understanding of heredity was one of: *Spontaneous Generation* (organisms could arise spontaneously from non-living matter), *Lamarckism* (traits acquired by an organism during its lifetime could be passed on to its offspring), *Blending Inheritance* (traits of offspring were a blend of the traits of their parents), and *Preformationism* (miniature versions of organisms existed within the reproductive cells of parents). Today, 150 years later, we know to a very high degree of certainty, how traits in organisms arise and their mechanism of inheritance, to the point that we can calculate the probability of a certain type

of rare cancer in the offspring of two given parents. However, Kelin's problem of who will click on her ad and how to maximize it remains unsolved and is often considered not worthy enough to be solved by science. Before the heroics of Mendel and Darwin, even the *science* of heredity was considered a domain of philosophy and not worth the *seriousness* of science.

Rocket science is considered the hardest of sciences. *It is solved.* It is solved to the extent that interplanetary launches over millions of kilometers can be planned to an accuracy of a few meters. Yet human conduct stays inscrutable, quirky, maddingly difficult to forecast and optimize for. It is unsolved to the extent that even opinion polls conducted right before the day of the election give opposite results to what is the actual outcome the next day. In my opinion, if behavior is not the problem to be worth solving, then what is!

Acknowledgments

कार्पण्यदोषोपहतस्वभाव :
पृच्छामि त्वां धर्मसम्मूढचेता : |
यच्छ्रेय : स्यान्निश्चितं ब्रूहि तन्मे
शिष्यस्तेऽहं शाधि मां त्वां प्रपन्नम् | |BG 2 :7| |

मयि सर्वाणि कर्माणि संन्यस्याध्यात्मचेतसा |
निराशीर्निर्ममो भूत्वा युध्यस्व विगतज्वर : | |BG 3 :30| |

नैव किञ्चित्करोमीति युक्तो मन्येत तत्त्ववित् |
पश्यञ्शृण्वन्स्पृशञ्जिघ्रन्नश्नन्गच्छन्स्वपञ्श्वसन् | |
प्रलपन्विसृजन्गृह्ण्न्नुन्मिषन्निमिषन्नपि |
इन्द्रियाणीन्द्रियार्थेषु वर्तन्त इति धारयन् | |BG 5 :8-9| |

This is an attempt to capture and thank those who have shaped my journey. Albeit, due to indirect and latent relations, this list will remain non-exhaustive despite my best attempts, it is still an attempt worth making.

In no particular order (with names of the organizations where I met these giants): Rajiv Ratn Shah (IIIT-D), Changyou Chen (SUNY at Buffalo), Ranjeeta Rani (GMPS), Roger Zimmerman (National University of Singapore), Debanjan Mahata (Bloomberg), Jessy Junyi Li (University of Texas at Austin), Balaji Krishnamurthy (Adobe MDSR), Sridhar Gantimahapatruni (Adobe), Jayakumar Subramanian (Adobe MDSR), Amanda Stent (Bloomberg), Anil Seth (FIITJEE), Dhruva Sahrawat (IIIT-D), Yifang Yin (National University of Singapore), Mika Hama (Second Language Testing Institute), Payman Vafaee (Columbia University), Pankaj Bansal (Adobe), Mohit Srivastava (Adobe), Gaurav Jain (Adobe), Shubham Yadav (NSIT), Rohit Jain (NSIT), Mohd Khwaja Salik (NSIT), Pratham Nawal (NSIT), Mayank Singh (NSIT), Somesh Singh (BITS-Pilani Goa), Aanisha Bhattacharyya (Adobe MDSR), Varun Khurana (IIIT-D), Rita Yadav (GMPS), Prabha Sinha (GMPS), Neeta Pandit (GMPS), Geetha Nair (GMPS), Swami Sarvapriyananda (Ramakrishna Mission), and finally my parents, Sushil Singla and Neena Singla, and my brother, Aman Singla.

Hopefully, I can return whatever I have gathered from these individuals back to society.

Abstract

Communication, as a system of messages, symbols, and cultural exchanges, is ubiquitous across all species. Scholars have argued that communication represents one of the most transformative evolutionary transitions in life's history (Smith and Szathmary, 1997), alongside pivotal developments like chromosomal mechanisms, eukaryotic formation, sexual reproduction, and multicellular life. Its unique capacity to enable cooperation and facilitate the unlimited transmission of cultural information grants species an unprecedented form of adaptive flexibility (Kirby *et al.*, 2008).

Because of the critical role communication plays in the survival and advancement of the species, communication has been studied since the ancient times. The earliest known work on communication, called Precepts by Ptah-Hotep appeared more than 4500 years ago (2300 BCE) (Gray, 1946). Since then, communication has seen three distinct waves of intensified interest: the first one in Ancient Greece with great Sophists like Aristotle, Isocrates, and Plato producing seminal works like Rhetoric, Phaedrus, and Antidosis (Hackforth, 1972; Rapp, 2002; Norlin *et al.*, 1928), the second one with the rise of print, the reformation, the Renaissance, and the European colonial pursuits (Mack, 2011), the third and most recent one during the Second World War (Briñol and Petty, 2012). We currently stand at the cusp of a fourth such phase, precipitated not by political upheaval (like the ideas of democracy or world war) or mechanical innovation (like the printing press and the steam engine), but by the unprecedented accumulation of digital content and behavioral data. This data now serves as the foundation for developing large language and diffusion models, which hold transformative potential for behavioral scientific inquiry. We will show in this thesis that these tools, while still in their infancy, have the potential to solve many problems considered ambitious in behavioral sciences.

Communication is composed of seven modalities: the communicator, message,

channel, time of receipt, receiver, time of behavior, and receiver's behavior (Shannon and Weaver, 1949; Lasswell, 1948, 1971). Critically, each communication turn's behavior becomes the subsequent turn's message, rendering communication a strategic interaction between sender and receiver aimed at optimizing shared or individual objectives (Smith and Harper, 2003). Examples like legal defense and prosecution, scientific discourse, mating, organizational communication, diplomacy, political propaganda, and culture (like folk songs and maxims), present different types of goals.

This thesis explores behavioral sciences' enduring mission—first articulated by Aristotle 2,500 years ago—of identifying and leveraging persuasive mechanisms (Rapp, 2002). The field has traditionally bifurcated into two epistemological approaches: explanation and prediction. Historically, behavioral scientists have sought explanations that can provide interpretable causal mechanisms behind human and societal functioning. However, societies and humans do not render themselves to clean-cut equations and formulas, as is evidenced by the limited success of behavioral explanations in predicting behavior. The emergence of extensive digital behavioral repositories has consequently shifted focus towards more robust predictive methodologies.

In this thesis, we start with the more traditional approach of behavior explanation, where we cover persuasion strategies in advertising images and videos. We construct the largest set of generic persuasion strategies based on theoretical and empirical studies in marketing, social psychology, and machine learning literature. We introduce the first dataset for studying persuasion strategies in advertisements. Further, we also introduce methods called universal adversarial triggers to mine behavioral models to understand what they learn. While persuasion strategies help a human correlate and understand content and behavior, universal adversarial triggers help understand what models learn, which makes them successful in predicting behavior.

Next, we turn attention towards behavior prediction by constructing general behavior models. These models, similar to large language models, aim to understand behavior *in general*, as opposed to designed for a specific behavioral task. We use the large repositories of digital analytics to train these models. The format

of this data is the general communication model consisting of the communicator, message, time of message, channel, receiver, time of effect, and effect. We call these models, Large Content and Behavior Models (LCBMs). We further show that large language models, while being used as general purpose models for a variety of tasks in different domains, are unable to solve behavioral problems. We investigate the reason for this and find that while training LLMs, behavioral data is removed as noise due to which they lose the behavioral capabilities.

We also show that after including the behavioral training data back leads to other positive side effects. Namely, we show that since behavior is an after effect of content (message), therefore, we can make inferences about content by looking at the receiver behavior. An example for this is blood pressure or eye dilation levels upon watching the movie Jurassic Park indicates the excitement level of different scenes. We show results for this hypothesis on more than 40 content understanding tasks across all four modalities of text, image, video, and audio.

Finally, we make initial strides towards solving the problem of generating performant content. We show this both for performant text generation, by taking the illustrative case of the behavior of memorability, and images, by generating images that are more engaging. We also develop mechanisms to measure the engagement potential of text to image generation models. We show that existing metrics to benchmark the quality of text to image generation models are not correlated with engagement. We develop a model to measure the engagement potential of an image. We release the first automated arena to benchmark the engagement of text-to-image models. We rank several popular text-to-image models on their ability to generate engaging images and further encourage the community to submit their models to the arena.

Contents

Preface iv

Acknowledgments vii

Abstract ix

Contents xii

Detailed Table of Contents

1 What is Behavior? Introducing The Two Cultures of Behavioral Sciences 1

1.1 Research Questions and Thesis Contributions 16

2 Explaining Behavior: Persuasion Strategies and Universal Adversarial Triggers 20

2.1 Persuasion Strategies In Advertisements 21

2.1.1 Related Work 26

2.1.2 Generic Taxonomy of Persuasion Strategies 29

2.1.3 Persuasion Strategy Corpus Creation 31

2.1.3.1 Persuasion Strategy Dataset For Image Advertisements 31

2.1.3.2 Persuasion Strategy Dataset For Video Advertisements 35

2.1.4 Modeling: Persuasion Strategy Prediction 37

2.1.4.1 Modelling Persuasion Strategy For Image Advertisements 37

2.1.4.1.1 Feature Extractors 39

2.1.4.1.2 Cross-Modal Attention 41
2.1.4.1.3 Persuasion Strategy Predictor 44
2.1.4.1.4 Multi Task Learning 45
2.1.4.1.5 Active Learning 46
2.1.4.2 Modelling Persuasion Strategy For Video Advertisements 48
2.1.4.2.1 Video Verbalization 50
2.1.4.2.2 Prompt format 55
2.1.4.2.3 Results 56
2.1.4.2.4 Ablation 59
2.1.4.2.5 A few examples of the stories generated using our method 59
2.1.4.2.6 Hallucinations Present In the Automatically Generated Stories 62
2.2 MINIMAL: Mining models for universal adversarial triggers 65
2.2.1 Abstract 65
2.2.2 Introduction 66
2.2.3 Related Work 69
2.2.4 The Proposed Approach 70
2.2.4.1 Class-Impressions Generation (CIG) Algorithm 70
2.2.4.2 The Universal Trigger Generation (UTG) Algorithm . 72
2.2.5 Experiments 73
2.2.5.1 Sentiment Analysis 74
2.2.5.2 Natural Language Inference 77
2.2.5.3 Paraphrase Identification 78
2.2.6 Analyzing the Class Impressions 80
2.2.7 Conclusion and Future Work 84

2.3 Chapter Conclusion 85

3 Modeling Behavior: A Case For Large Content and Behavior Models 86

3.1 Introduction 90

3.1.1 Related Work 97

3.2 Setup 99

3.2.1 The Content Behavior Corpus (CBC) 99

3.2.2 Model 103

3.2.3 Content Behavior Test Benchmark 107

3.2.4 Behavior Instruction Fine-Tuning (BFT) 109

3.3 Results and Discussion 112

3.3.1 Verbalization Listings 113

3.3.2 Cross Domain Generalization 117

3.4 Failure Cases and Model Limitations 119

3.4.1 Scale vs. Specialization Trade-offs 119

3.4.2 Behavior Domain Specificity and Transfer Limitations 120

3.4.3 Context Length and Complexity Limitations 121

3.4.4 Behavioral Signal Quality Dependencies 121

3.5 Conclusion 122

3.6 Ethical Considerations: Bias Mitigation and Privacy-Preserving Strategies in Large-Scale Behavioral Modeling 122

3.6.1 Privacy-Preserving Data Collection and Processing 123

3.6.2 Bias Mitigation and Fairness-Aware Learning 123

3.6.3 Opportunities for Future Work 124

4 Analyzing Behavior: Teaching Behavior Improves Content Understanding 126

4.1 Synthesizing Human Gaze Feedback for Improved NLP Performance . 127
4.1.1 Introduction . . . 128
4.1.2 Related Work . . . 130
4.1.3 Proposed Model . . . 134
4.1.3.1 ScanTextGAN Model Architecture . . . 134
4.1.3.2 Dataset . . . 137
4.1.3.3 Parameter Settings . . . 138
4.1.4 Performance Evaluation . . . 138
4.1.4.1 Evaluation Datasets . . . 139
4.1.4.2 Evaluation of Scanpath Generation . . . 140
4.1.4.2.1 Scanpath Evaluation Metrics . . . 140
4.1.4.3 Application to NLP Tasks . . . 144
4.1.5 Intent-Aware Scanpaths . . . 147
4.1.6 Conclusion . . . 147
4.1.7 Limitations . . . 149
4.2 Teaching Human Behavior Improves Content Understanding Abilities Of VLMs . . . 150
4.2.1 Introduction . . . 151
4.2.2 Methodology . . . 156
4.2.2.1 BLIFT Dataset . . . 156
4.2.2.1.1 Data from Reddit . . . 156
4.2.2.1.2 Data from YouTube . . . 157
4.2.2.2 Instruction Fine-Tuning LLaMA-Vid . . . 159
4.2.3 Results and Discussion . . . 165
4.2.3.1 Evaluation . . . 165
4.2.4 Discussion . . . 171

4.2.5 Conclusion . . . 178
4.2.6 Appendix . . . 179
4.2.6.1 Listings . . . 179
4.2.6.2 Dataset Descriptions . . . 180
4.2.6.3 Memorability Metrics: Detailed Explanation and Limitations . . . 181
4.2.6.4 Failure Case Analysis and Model Limitations . . . 184
4.2.6.5 Limitations . . . 186
4.2.7 Broader Impacts . . . 186
4.2.7.1 Ethical Implications . . . 186

5 Optimizing Behavior: Generating Content to Optimize Behavior 188

5.1 Long-Term Ad Memorability: Understanding And Generating Memorable Ads . . . 188
5.1.1 Introduction . . . 189
5.1.2 LAMBDA Protocol, Study & Insights . . . 196
5.1.2.1 Video Collection . . . 196
5.1.2.2 Annotation Protocol . . . 197
5.1.2.3 What makes an Ad memorable? . . . 198
5.1.3 Predicting Ad Memorability . . . 201
5.1.3.1 Encoding Multimodal Content . . . 204
5.1.4 Two-Stage Training . . . 205
5.1.4.1 Results and Discussion . . . 205
5.1.5 Generating Memorable Ads . . . 206
5.1.5.1 Evaluation . . . 207
5.1.5.2 Results . . . 209
5.1.6 Conclusion . . . 212

5.1.7 Generation of Ads using Henry-SEED 212
5.1.8 Extraction And Use Of Cognitive And Perceptual Signals In Advertisements 222
5.1.9 Ablation Experiments 225
5.1.10 Questionnaire to Gather Human Preferences over Generated Ads 228
5.1.10.1 Expert Feedback Collected For Generated Ads 230
5.1.11 Perplexity evaluation 231
5.1.12 Annotation Protocol and Participant Details for the LTM Study 232
5.1.12.1 Memorability Questionnaire 234
5.1.12.1.1 Introductory Questionnaire (to be filled before the study starts) 234
5.1.12.1.2 Checks (to be answered during the experiment) 235
5.1.12.1.3 Recognition Questions (asked after a few days after watching the videos) 235
5.1.13 Collection of all the Prompts used in the Paper 235
5.1.13.1 GPT-4 Prompts 235
5.1.13.2 Henry Prompts 237
5.1.13.3 Mistral prompt for filtering marketing ads 238
5.1.14 Computing Infrastructure and Hyperparameters 238
5.1.14.1 Modeling Memorability 238
5.1.14.1.1 Images 238
5.1.14.2 Videos 239
5.1.14.3 Generating Memorable Ads 239
5.1.14.4 Inference hyperparameters 239
5.1.15 License and Terms of Release 239
5.1.16 Limitations and Potential Risks 240

5.2 Measuring And Improving Engagement Of Text-to-Image Generation Models . . . 240
5.2.1 Introduction . . . 241
5.2.2 *EngagingImageNet*: Dataset With In-The-Wild User Engagement 246
5.2.3 *EngageNet*: Measuring Image Engagement . . . 247
5.2.3.1 Alignment of Existing Models with Viewer Engagement 247
5.2.3.2 EngageNet Model To Align With Viewer Engagement 248
5.2.4 Methods to Improve Image Engagement . . . 252
5.2.4.1 Conditioning Stable Diffusion on More Engaging Prompts 252
5.2.4.2 Preferred Finetuning on High-Engagement Images . . 254
5.2.4.3 Aligning Stable Diffusion With Engagement . . . 254
5.2.4.4 Evaluating the Methods Adopted for Engagement-Optimization 256
5.2.5 *Engagement Arena*: Measuring Engagement Capabilities of Text-to-Image Models . . . 258
5.2.6 Conclusion . . . 260
5.2.7 Appendix . . . 260
5.2.7.1 Related Work . . . 260
5.2.7.2 Study with marketers . . . 262
5.2.7.3 Analysing Visual Aspects that Drive Engagement . . 263
5.2.7.4 EngagingImageNet Filtering Steps . . . 263
5.2.7.5 EngagingImageNet Additional Details . . . 264
5.2.7.6 Prompts for Instruction Finetuning . . . 264
5.2.7.7 Repurposing EngageNet for Design Specification Generation (DSG) . . . 268
5.2.7.7.1 Training For Design Specification Prediction Task 268
5.2.7.7.2 Results for Engagement-conditioned Design Specification Prediction Task . . . 272

5.2.7.7.3 Evaluation Metrics for Design Specification Prediction . . . 272
5.2.7.8 Performance Alignment of Stable Diffusion using Design Specification Generation (DSG) Reward . . . 275
5.2.7.8.1 DDPO Additional Details . . . 275
5.2.7.8.2 Design Specification Generation (DSG) reward . . . 276
5.2.7.9 Broader Impacts and Limitations . . . 276
5.2.7.9.1 Limitations . . . 278
5.2.7.10 Additional Figures . . . 278

6 Conclusion and an Outlook for Future Work **285**

6.1 Limitations and Future Work . . . 286
6.2 Societal and Ethical Implications of Behavior Optimized AI . . . 290
6.2.1 Dual-Use Nature and Manipulation Risks . . . 291
6.2.2 Privacy and Data Protection Challenges . . . 292
6.2.3 Bias, Fairness, and Representation . . . 292
6.2.4 Autonomy, Consent, and Human Agency . . . 293
6.2.5 Accountability and Governance Frameworks . . . 294
6.2.6 Long-term Societal Implications . . . 294
6.2.7 Toward Responsible Development . . . 295
6.3 Open Research Questions . . . 295

Publications covered as part of thesis **299**

Dataset contributions covered as part of this thesis **300**

Other Publications During PhD **301**

Patents During PhD **304**

Awards **306**

List of Tables **307**

List of Figures **321**

Bibliography **337**

Chapter 1

What is Behavior? Introducing The Two Cultures of Behavioral Sciences

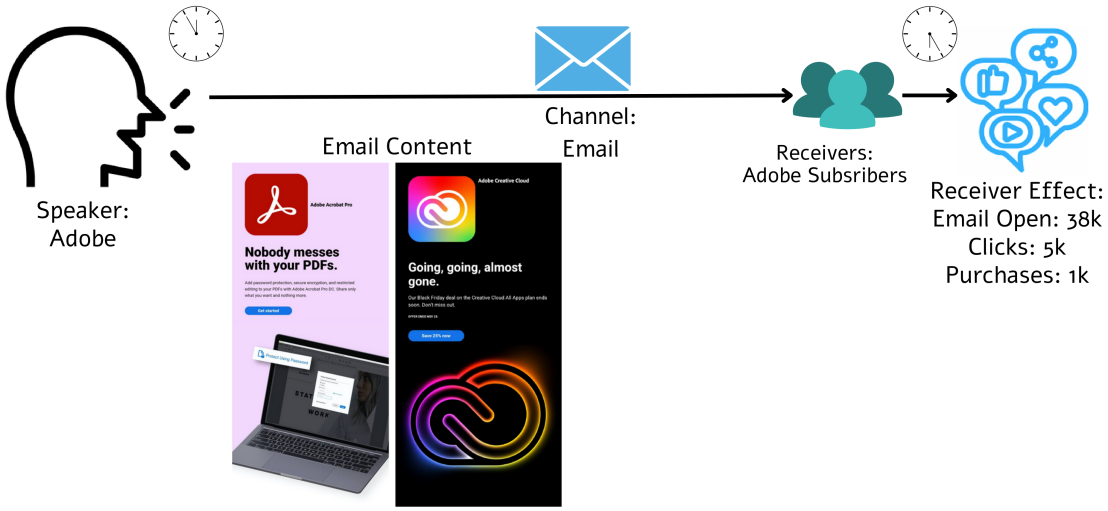


Figure 1.1: Communication process can be defined by seven factors: Communicator, Message, Time of message, Channel, Receiver, Time of effect, and Effect. Any message is created to serve an end goal. For marketers, the end goal is to bring in the desired receiver effect (behavior) (like clicks, purchases, likes, and customer retention). The figure presents the key elements in the communication pipeline - the marketer, message, channel, receivers, and finally, the receiver effect.

Behavior as a modality[*] occurs in the process of communication. Communication includes all of the procedures by which one mind may affect another (Shannon and Weaver, 1949). This includes all forms of expression, such as words, gestures, speech, pictures, and musical sounds. Communication can be seen as being composed of seven modalities (Fig. 1.1): (the communicator, message, time of message (or time of receipt), channel, receiver, time of effect, and effect).

The seven modalities of communication interact in different ways depending on the context. We illustrate this with four examples:

[*] A *modality* is defined in terms of information, such that a modality is a medium through which information is conveyed (Liang *et al.*, 2022; Grifoni, 2009; Martin *et al.*, 2001). Similarly, a multimodal distribution is defined as having more than one peak in the probability distribution describing the nature of information.

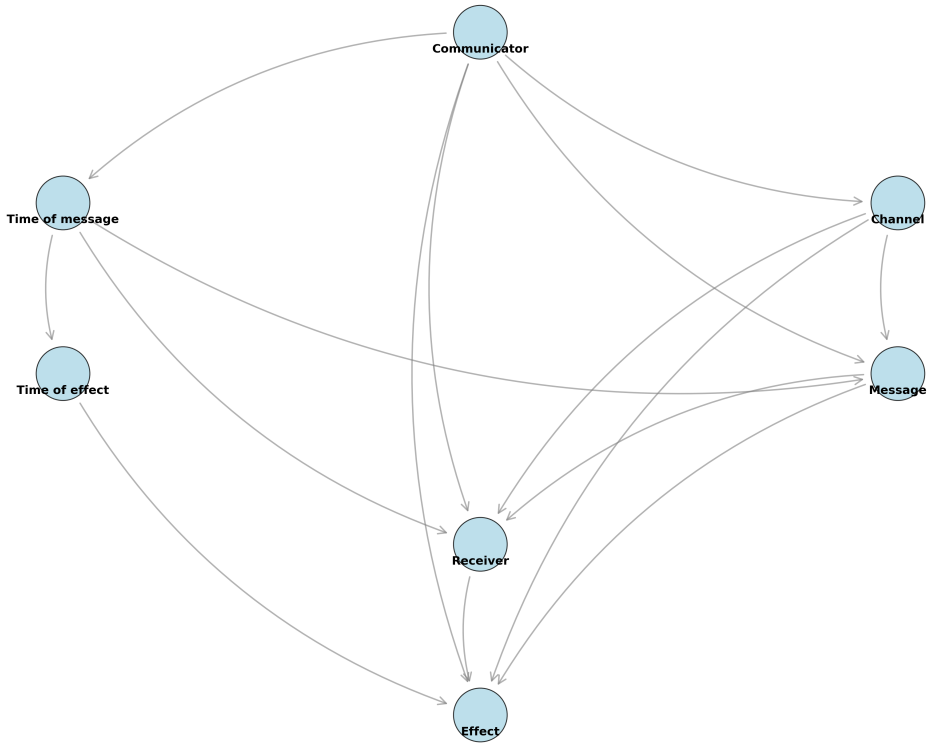


Figure 1.2: Interaction of the communication modalities in a mass communication scenario like a speech.

1. **Mass Communication, such as a Speech**: In mass communication, the communicator might be an individual—like a CEO or politician—or an organization, such as a media outlet or political party. The message is the content being conveyed, while the channel refers to the medium through which it is delivered (e.g., radio, television, or the internet). The receiver is the audience, the time of effect is when the audience receives the message, and the effect is their subsequent behavior.

   Fig. 1.2 illustrates how the seven modalities interact in the context of a speech. The communicator typically controls three key factors: the message, the channel, and the timing of the message. These choices are influenced by the communicator's goals, the characteristics and constraints of the chosen channel, and the relevance of the timing.

   For example, a politician giving a speech on "Artificial Intelligence" might tailor their message depending on whether it's broadcast over public radio or delivered live at a rally. Likewise, the timing can shape the content: a speech delivered near Independence Day might include different themes than one given around New Year's.

   The time of effect is when the audience engages with the message—often simultaneous with its delivery in live communication. The audience's behavior—the ultimate effect—is shaped by the message, the channel, the timing, the communicator, and their own characteristics as receivers.

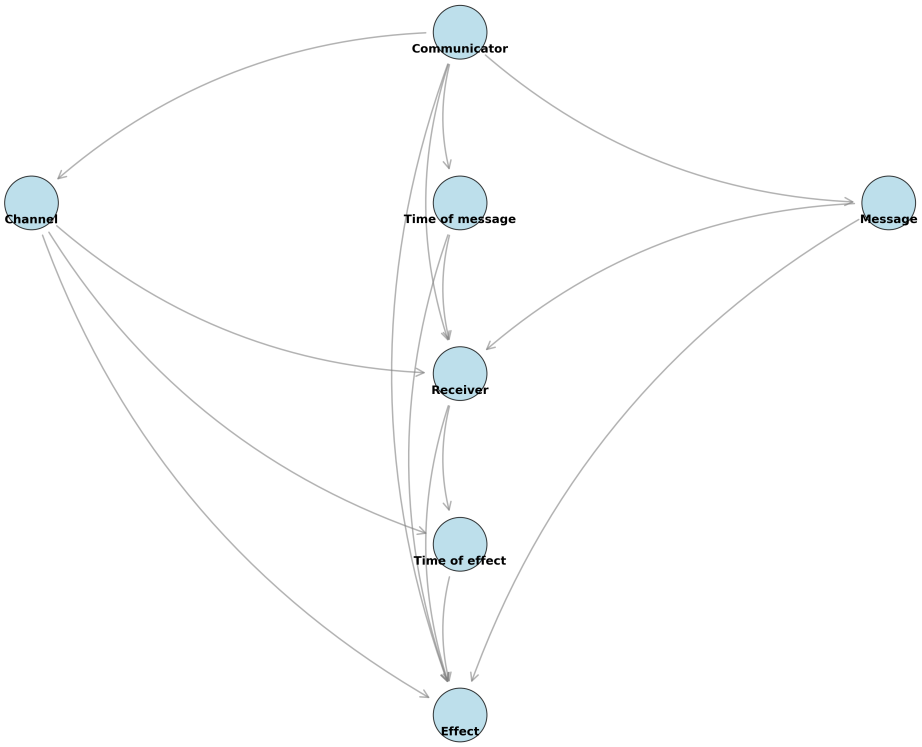


Figure 1.3: Interaction of the communication modalities in a mass communication scenario on a social media platform like Twitter.

2. **Social Media**: In mass communication via social media, the communicator can be an individual (such as a journalist, influencer, or politician) or an organization (like a news outlet or a company). The message represents the content being shared, while the channel refers to the medium—such as Twitter, Facebook, or another social media platform.

   Fig. 1.3 illustrates how the seven modalities interact in the context of a tweet. Unlike traditional speeches, social media communication has an asynchronous nature. The communicator still controls three key factors: the message, the channel, and the timing of the message. However, the reach and effect of the message can be influenced by platform (channel) algorithms and the content.

   For example, a tech company announcing a product launch on Twitter might craft the message differently than if they were making a live keynote speech. The choice of timing is also crucial—posting during peak engagement hours may enhance visibility.

   The time of effect is when the audience engages with the message, which may not be immediate, as tweets can be shared and rediscovered over time. The ultimate effect on the audience—whether it be a change in opinion, an action taken, or a broader societal reaction—depends on factors such as the communicator's credibility, the content's relevance, and the digital platform's dynamics.

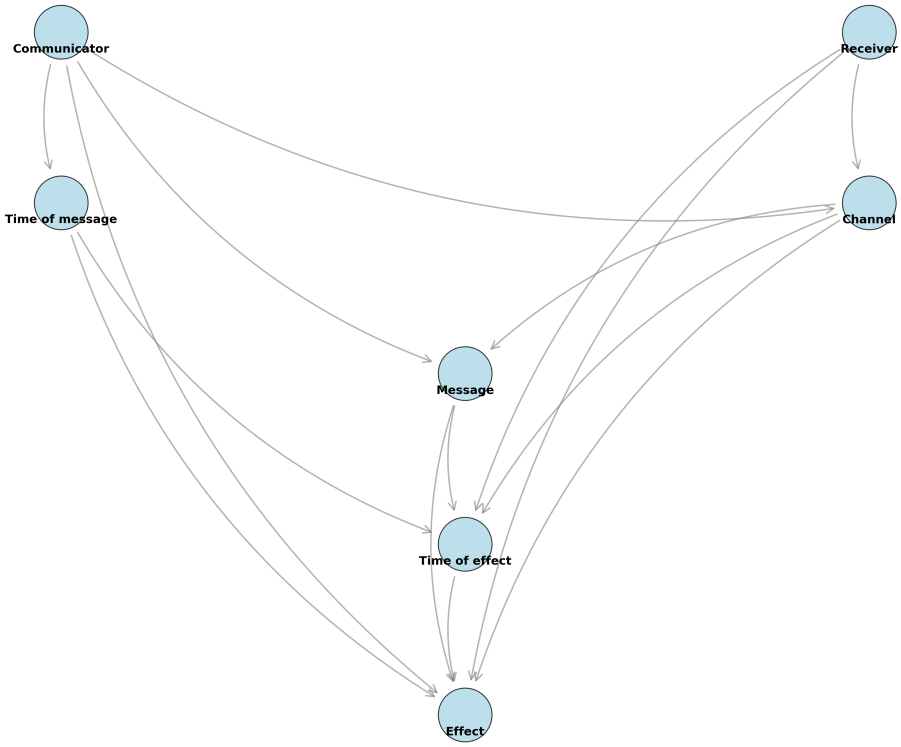


Figure 1.4: Interaction of the communication modalities in a person-to-person communication scenario on a channel like Whatsapp.

3. In peer-to-peer (P2P) communication, two individuals engage in direct exchanges, typically through mutually agreed-upon channels such as messaging apps (e.g., WhatsApp, Signal), emails, or phone calls. Unlike mass communication, where the communicator chooses the audience, in P2P communication, the sender does not determine the receiver—the interaction occurs only when both parties establish a connection, such as exchanging contact information or joining the same platform.

   Fig. 1.4 illustrates the interaction of seven key modalities in a P2P communication setting. The channel is chosen by mutual agreement between the sender and receiver, based on factors like convenience, urgency, and formality. The message is shaped by the sender's intent channel. The time of effect depends on when the receiver reads and processes the message, which may vary based on availability and responsiveness, and also the message content, and when the message was sent.

   Unlike mass communication, P2P messaging typically has a controlled effect, as it is intended for a specific recipient. However, the effectiveness of the message depends on the communication, timing, the receiver's interpretation, the sender, and the channel.

4. The next setting is a continuation of the previous one, where the communicator and receiver swap their roles. The channel remains the same. The effect of the previous iteration becomes the message of the next iteration,

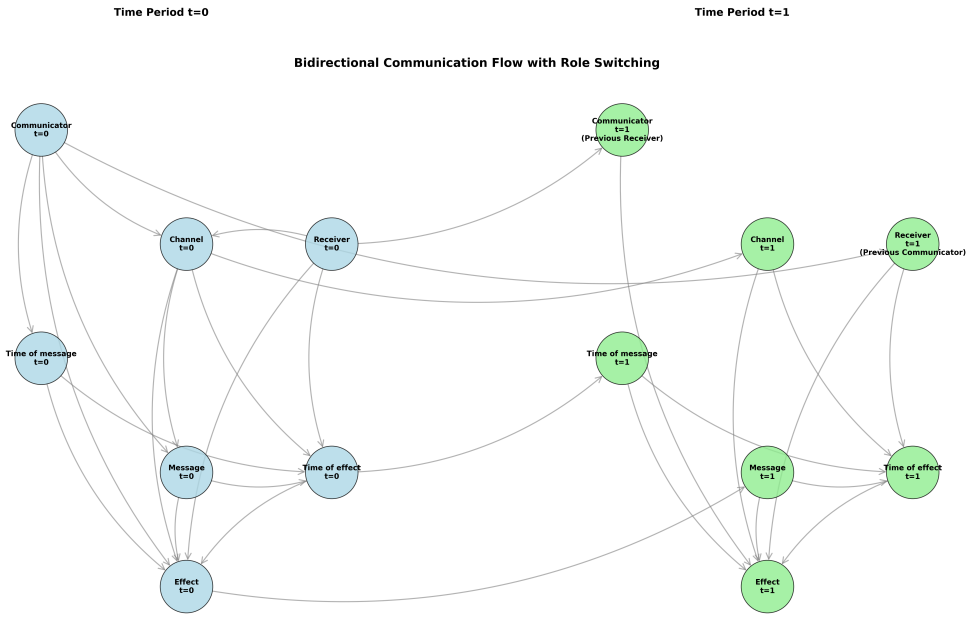


Figure 1.5: Interaction of the communication modalities in person-to-person communication with bidirectional communication.

with its target receiver as the communicator of the previous iteration. The new receiver's effect follows the same dependencies as the last iteration, namely, the new message, the time of the new message, the time of the new effect, the new communicator and the receiver, and the channel. Fig. 1.5 presents this case.

These modalities can vary independently of each other (Khandelwal *et al.*, 2024; Khurana *et al.*, 2023; SI *et al.*, 2025; Khurana *et al.*, 2024) and carry signals about each other (Khurana *et al.*, 2023; Bhattacharyya *et al.*, 2023). The message as a modality carries information from the communicator to receiver and encodes information generated by the communicator. Similarly, behavior (*aka* effect) as a modality carries information from the receiver and encodes information generated by the receiver. This is often a continuous cycle, where behavior generated in the previous cycle becomes the message of the next cycle, thus forming a (continuous) conversation.

Different fields of behavioral sciences deal with different parts of behavior. we will give a broad overview of these fields in the upcoming paragraphs, but two streams have emerged broadly in behavioral sciences: explanation and prediction of behavior (receiver effect) (Breiman, 2001; Hofman *et al.*, 2017; Shmueli, 2010).

Historically, behavioral social scientists have sought explanations of human

behavior that can provide interpretable causal mechanisms behind human functioning. A few prominent examples are Milgram's (Milgram and Gudehus, 1978) and Asch's (Asch, 1948) experiments on persuasion, explaining the causal mechanism of obedience to authority. Similarly, Cialdini's (**?**) identification of six principles of persuasion (reciprocity, commitment/consistency, social proof, authority, liking, and scarcity) provides a theoretical framework that explains *why* certain messaging strategies are effective. In economics, prospect theory by Kahneman and Tversky (**?**) explains how people make decisions under uncertainty by identifying cognitive biases like loss aversion and the endowment effect.

The approach of theorizing has worked remarkably well in physical sciences where the data is plentiful, and theories make unambiguous predictions. For instance, Newton's laws of motion precisely predict planetary orbits, allowing us to calculate exactly when Halley's comet will return (every 76 years). Einstein's theory of relativity predicted the bending of light around massive objects, which was empirically confirmed during the 1919 solar eclipse. In chemistry, the periodic table not only organized known elements but successfully predicted the properties of undiscovered elements like gallium and germanium. Similarly, Maxwell's electromagnetic theory predicted the existence of radio waves decades before they were experimentally demonstrated by Hertz.

However, such theoretical success has not been replicated in *predicting* social outcomes in behavioral sciences (Collaboration, 2015; Tetlock, 2017; Collaborative, 2023). Unlike physical phenomena where F=ma reliably predicts motion, human behavior is far more complex and context-dependent. For example, while Cialdini's persuasion principles explain why authority figures influence behavior, they cannot predict with Newton-like precision whether a specific endorsement will increase sales by 15% or 50%. In fact, many studies have shown that expert human opinions fare similar to non-experts (*e.g.*, predicting economic and political trends (Tetlock, 2017), societal change: (Collaborative, 2023), and advertising success: (Singh *et al.*, 2024*b*)), and the opinion of non-expert population is roughly the same as a random coin toss in predicting behavior (*e.g.*, predicting cascades (Tan *et al.*, 2014) or image memorability (Isola *et al.*, 2013)). At the same time, causal mechanisms proposed by experts and theorists have their own merits; most notably, they help decision-makers (often humans) to make intuitive sense of the situation and make

their next decision based on it.

In parallel, due to the availability of human behavior data at scale, researchers in machine learning are showing a growing interest in traditionally behavioral science topics, such as messaging strategies leading to persuasion (Habernal and Gurevych, 2016; Kumar *et al.*, 2023*a*; Luu *et al.*, 2019; Bhattacharyya *et al.*, 2023), information diffusion (Cheng *et al.*, 2014*a*; Martin *et al.*, 2016), and prediction and predictability of human behavior (Choi and Varian, 2012; Song *et al.*, 2010). This prediction-oriented approach mirrors the success of machine learning in other domains. For example, in computer vision, deep learning models can classify images with superhuman accuracy (achieving 97.8% accuracy on ImageNet compared to human performance of 94.9%) without necessarily understanding *why* an image contains a particular object. Similarly, in natural language processing, BERT and GPT models achieve remarkable performance on language tasks through pattern recognition in massive datasets, even though they lack explicit understanding of grammar rules or semantic theories. In recommender systems, collaborative filtering algorithms successfully predict user preferences (Netflix's recommendation system was estimated to save the company $1 billion annually) by finding patterns in user behavior data, without needing to understand the psychological mechanisms behind why users like certain content.

In the prediction community, different subfields have emerged dealing with the different parts of the problem of optimization of human behavior. For instance, advertisement personalization studies how to optimize (choose) *receiver* for a given message (Chandra *et al.*, 2022), and recommendation systems study how to *choose content* from a set of pre-decided contents for a given receiver to elicit a certain effect (Herlocker *et al.*, 2004). A popular problem within the prediction community is the effect prediction problems, for example, clickthrough (CTR) prediction (McMahan *et al.*, 2013), Twitter cascade prediction (Cheng *et al.*, 2014*a*; Martin *et al.*, 2016), sales prediction (Choi and Varian, 2012; Pryzant *et al.*, 2017), content memorability prediction (Isola *et al.*, 2011; Khosla *et al.*, 2015; SI *et al.*, 2025), *etc.* There are also works to optimize the time of the message to elicit certain effect (Newstead and Romaniuk, 2010; SI *et al.*, 2025). Some of the major problems studied in behavioral sciences are given below. Through this list, one can observe that all the factors of communication are studied independently

in their own light without relying on the underlying unity and continuity of the communication process.

1. **Problems related to optimization in the sender space**:
   (a) **Source Optimization**: *Who* should send a particular message over a channel to a specific audience to get the desired behavior? This includes selecting between different brand voices, influencers, spokespersons, or organizational entities based on authority, trustworthiness, and audience affinity.

2. **Problems related to optimization in the receiver space**:
   (a) **Personalization**: Identifying the optimal receiver for a specific content-channel-time combination to maximize engagement and conversion probability.
   (b) **Customer Segmentation**: Strategic division of the customer base into distinct, actionable groups based on shared characteristics, behaviors, or value propositions to enable differentiated marketing approaches.
   (c) **Social Network Analysis**: Modelling the interconnectedness of receivers (and senders) together in a graph to describe social phenomena like contagion and homophily.
   (d) **Lookalike Modeling**: Identifying and targeting prospective customers who share similar characteristics, behaviors, and propensities with existing high-value customers or target audiences.
   (e) **Market surveys**: Systematic collection and analysis of primary data about target markets, customer preferences, and competitive landscape to inform marketing decisions.
   (f) **Identity stitching**: Probabilistic and deterministic matching of cross-channel, cross-device customer actions to create unified customer profiles and journey maps.
   (g) **Behavior Explanation**: Discovering causal mechanisms behind a receiver action.

3. **Problems related to optimization in the content space**:
   (a) **Recommender Systems**: Identifying the optimal content that should be delivered next to a certain receiver given a fixed channel, time, and a repository of contents (and their corresponding senders).
   (b) **A/B Testing**: A randomized experiment involving two or more variants with the goal of discovering which variant of a message performs better with a certain audience.
   (c) **Customer Targeting**: Determining what message should be delivered to a target audience segment, targeted based on multi-dimensional attributes (geographic, demographic, behavioral, and psychographic) through particular channels.
   (d) **Propensity Modelling or Engagement Modelling**: Modeling probability of engagement in terms of actions like Clickthrough, social media actions such as likes and shares for a certain audience, sender, and campaign.

(e) **Transsuasion**: Conversion of a content from low-performing to high-performing for a given audience, sender, and time, while maintaining the content's intent, style, and emotional impact.

(f) **Transcreation**: Conversion of a content designed for one audience (like a particular culture) to another audience, while maintaining the content's intent, style, and emotional impact.

(g) **Search Engine Optimization**: Improving the quality and quantity of website traffic to a website from search engines by doing content optimization (like adding keywords, backlinks, *etc*).

(h) **Performant Content Generation**: Generate content that can perform better for a given audience, sender, time, and goal.

(i) **Argument Mining**: Automatic extraction and identification of argumentative structures from natural language text.

(j) **Persuasion Strategies**: Use of rhetorical devices (such as emotion, social identity, and scarcity) to optimize the effect of a message on a certain audience.

4. **Problems related to optimization in the channel space**:
   (a) **Channel Optimization**: Optimizing channels for a particular audience, sender, time, and goal.
   (b) **Marketing Mix Modeling**: Measuring and attributing the impact of various decisions like channel investments, discounts, promotional campaigns in their contribution to engagement and sales.
   (c) **Auction Design** and **Bidding**: Mechanisms to discover the cost of attention of a certain receiver to a particular sender, time, and campaign goal.

5. **Problem related to optimization in the time space**:
   (a) **Send Time Optimization**: Determining the optimal timing for message delivery for a certain receiver, sender, content combination.
   (b) **Trend Forecasting**: Projecting marketing and social trends in the future.

A common theme that runs through behavioral sciences is the intent to control behavior. For example, marketers want to increase click-through rates from 2% to 5%, political campaigns aim to boost voter turnout by 15%, and content creators seek to double their video engagement metrics. From this intent, explanation and prediction serve as intermediate steps toward control and optimization. For instance, understanding why certain ad formats generate higher conversion rates (explanation) enables predicting which content will perform best (prediction), ultimately allowing marketers to systematically increase sales by 20-30%.

Optimizing behavior means fulfilling the communicator's objectives by strategically controlling the other six parts of the communication process (Fig. 1.1): selecting the right spokesperson (communicator), crafting compelling messages (content), choosing optimal platforms (channel), timing releases strategically (time), targeting specific audiences (receiver), and measuring outcomes (effect). However, current approaches are fragmented—a model trained to predict Twitter engagement cannot predict YouTube video views, and an advertising optimization system for fashion brands fails when applied to food products. Due to this limitation, the solution requires a general understanding of human behavior that can transfer across domains, platforms, and contexts. In this thesis, our aim is to develop such models that can understand behavioral patterns universally, whether predicting social media engagement, optimizing email marketing campaigns, or improving video content retention rates.

The characteristic that marks the digital age is the prevalence of human behavioral data in huge repositories. This data is *big* (allowing to model heterogeneity), *always-on* (allowing to look in the past as well as live measurements), observational (as opposed to reactive), but also *incomplete* (does not capture all that is happening everywhere everytime in a single repository) and *algorithmically confounded* (generated as a byproduct of an engineering process with a goal) (Salganik, 2019). While the predictive culture has tried to make use of some of this data in the form of social media datasets like Twitter (Tumasjan *et al.*, 2010; Asur and Huberman, 2010) and Instagram (Kim *et al.*, 2020), Google trends (Choi and Varian, 2012; Carrière-Swallow and Labbé, 2013), Wikipedia (Generous *et al.*, 2014; De Toni *et al.*, 2021; Mestyán *et al.*, 2013), shopping websites (Krumme *et al.*, 2013; De Montjoye *et al.*, 2015) and other data sources (Brockmann *et al.*, 2006; Song *et al.*, 2010; Miritello *et al.*, 2013), these efforts are limited, in the sense of being dependent on one or a few chosen platforms, able to answer a limited set of questions, and restricted by access to private data. We want a model that can understand (predict and explain) *human behavior in general* as opposed to modeling a particular effect (retweet prediction) on a particular platform (*e.g.* Twitter) for a certain type of users. This problem carries parallels with the problem being solved in the natural language processing (NLP) community, where supervised models in NLP are limited by the amount of supervision available and being able

to answer one question (for which the supervised model was trained). The problem was solved by developing Large Language Models (LLMs), which are general purpose models capable of *understanding language*, and hence can solve natural language tasks like sentiment analysis, question answering, email generation, and language translation in zero-shot (*i.e.* without needing any explicit training for that task) (Devlin *et al.*, 2019; Brown *et al.*, 2020; Radford *et al.*, 2018; Raffel *et al.*, 2020; Radford *et al.*, 2019).



Figure 1.6: Levels of content analysis. The figure lists tasks and their sample outputs arranged in a hierarchy (Shannon and Weaver, 1949). This is roughly based on levels of language. Notably, humans are good at predicting the first three levels but not the last level (Tetlock, 2017; Collaborative, 2023; Tan *et al.*, 2014; Isola *et al.*, 2013).

Similarly, how do we develop a model capable of understanding behavior *in general*? With the intent to answer this question, we take motivation from LLMs, where the idea is to train a model on a data-rich task. The task chosen to train LLMs is the next-word prediction, and the dataset is the text collected from the entire internet. The next-word prediction task is a data-rich task that can be trained on the huge text repositories from the internet. The intuition is that two approaches have always worked for neural networks: larger model sizes and more data for training (Mikolov *et al.*, 2013; Devlin *et al.*, 2019; Radford *et al.*, 2018; Raffel *et al.*, 2020). Going from a few million tokens of text (Mikolov *et al.*, 2013; Radford *et al.*, 2018) to a trillion tokens (Touvron *et al.*, 2023; Brown *et al.*, 2020) leads to an increase in the transfer learning capability leading to performance improvements over a wide variety of natural language tasks.

The digital revolution has provided us with huge repositories of data. We leverage the human behavior repositories available on the internet for this general-purpose human behavior model. The format of this data is the general communication model shown in Fig. 1.1 consisting of communicator, message, time of message, channel, receiver, time of effect, and effect. Due to the incomplete nature of behavioral repositories, all the factors are usually not always available.

However, a subset is always available, and we show that the data scale, along with a large model, helps make a general behavior understanding model (Khandelwal *et al.*, 2024). We call this model, Large Content and Behavior Model (LCBM). We show that LCBM can predict behavior, explain it, and generate a message to bring about certain behavior (SI *et al.*, 2025; Khandelwal *et al.*, 2024; Khurana *et al.*, 2024).

Are general LLMs unable to solve behavioral problems? A question that arises is whether LLMs, which already learn trillions of text tokens, are able to understand and predict behavior. We investigate that question over several large models, including GPT-3.5 (Brown *et al.*, 2020), GPT-4 (OpenAI, 2023), Llama-13 B and LLama-7B (Touvron *et al.*, 2023), and find that they are unable to solve the behavioral problems listed before. The reason for this is that large language models only include one factor (message) out of the 7-factor communication model (Fig. 1.1) while considering other parts as "noise" (for instance, see (Biderman *et al.*, 2022; Penedo *et al.*, 2023)). This systematic purge of communicator, receiver, channel, time, and, most importantly, behavior causes the models not to develop any behavioral capabilities (Level-C of Shannon and Weaver (Shannon and Weaver, 1949)). As an example, Llava (Liu *et al.*, 2023*a*), a recent large language and vision model (VLM) trained by connecting a vision encoder with a language model, shows that after training on a few hundred thousand instructions, the language model can now "see", and is able to answer questions on the images. However, the questions all lie in the first two levels of content analysis shown in Fig. 1.6. The reason is that the instructions used to align the image encoder with the downstream LLM all lie in the first two levels (sender and message) while ignoring the last two (receiver and behavior). In the upcoming chapters, we explore how we can train a general behavior model and how including the other factors of communication back in training data helps in understanding human behavior.

Outline for the upcoming chapters: Following the two traditions of behavioral sciences, we delve into both explanation and prediction. Figure 1.7 gives a visual description of the various chapters and how they link with each other. In Chapter-2, we start with a more traditional approach to behavior explanation, where we cover the first works on extracting persuasion strategies in advertisements (both images and videos) (Kumar *et al.*, 2023*a*; Bhattacharyya *et al.*, 2023). The con-

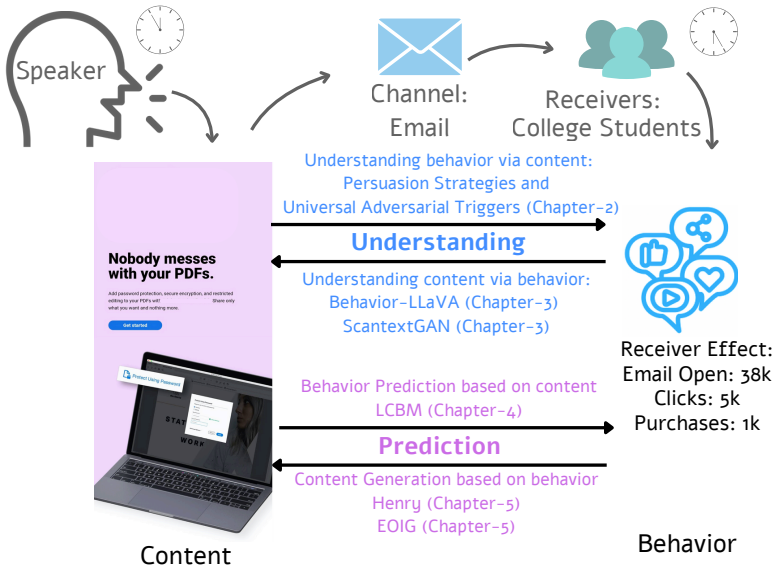


Figure 1.7: Communication process can be defined by seven factors: Communicator, Message, Time of message, Channel, Receiver, Time of effect, and Effect. Any message is created to serve an end goal. In this thesis, we explore the two main concerns of behavioral sciences: understanding (or explanation) and prediction. The figure shows the links between the different chapters and how they link together to form the two core pillars of understanding and explanation.

tributions of these works include constructing the largest set of generic persuasion strategies based on theoretical and empirical studies in marketing, social psychology, and machine learning literature and releasing the first datasets to enable the study and model development for the same. These works have been deployed to understand the correlation between the kinds of marketing campaigns and customer behavior measured by clicks, views, and other marketing key performance indicators (KPIs). Further, we also introduce methods called universal adversarial triggers (UATs) to mine behavioral models to understand what they learn. This approach provides a converse approach for human behavior understanding. While persuasion strategies help a human correlate and understand content (message) and behavior, universal adversarial triggers help understand what models learn, which makes them successful in predicting behavior.

Following this, in Chapter-3, we delve into the question of modeling behavior. The key insight behind this chapter is that behavior is always produced by a receiver in response to a content sent by a sender at a time. We model behavior together with the pieces of sender, receiver, time, and content. We show that while large language models already model content, they do not model the

other pieces of sender, receiver, and time. We model these factors together and show emergent abilities in understanding behavior. We observe that teaching the Large Content and Behavior Models (LCBM) behavior and content simulation improves its capabilities on them (expected), but the model also shows signs of domain-adaptation in behavior modality (few-shot capability, unexpected) and improvements in behavior understanding (zero-shot capability, unexpected). To spur research on the topic of large content and behavior models, we release our generated behavior instruction fine-tuning data from over 40,000 public domain YouTube videos and 168 million Twitter posts. The data contains: 1) YouTube video links, automatically extracted key scenes, scene verbalizations, replay graph data, video views, likes, comments, channel name, and subscriber count at the time of collection, and 2) Twitter extracted account names, tweet text, associated media (image and video) verbalizations (including image captions, keywords, colors, and tones), tweet timestamps, and like counts. We also release a benchmark to test performance on the joint content behavior space introducing two types of tasks in this space: predictive and descriptive. In the predictive benchmark, we test the model's ability to predict behavior given the content and predict content given the behavior. In the descriptive benchmark, we validate its explanation of human behavior by comparing it with ground-truth annotations we obtain from human annotators that try to explain human behavior.

Next, in Chapter 4, we analyze the communication process in more detail. As behavior is the signal a receiver emits when a sender sends a content; similarly, one can see this behavior emitted as a content in the next cycle, where the receiver becomes the sender, and the sender becomes the receiver. Therefore, we ask if we can understand the content better by modeling behavior. For example, a person's heightened state of emotional response, like dilated pupils and sweat, while watching an action scene from the movie Jurassic Park gives us much information about the scene itself. Today's models are built only on content (the Jurassic Park movie itself) while ignoring the human behavioral responses over the content. Behavioral responses like likes, shares, comments, replay graphs, and upvotes are freely available and waiting to be integrated into the workflow to understand content better. We show evidence for this hypothesis by improving LLMs across 46 different tasks over 23 benchmark datasets across all four modal-

ities of language, audio, text, and video. The hypothesis of extracting and using signals from behavior is lately getting attention in the fields of human alignment and reinforcement learning with human feedback (RLHF) where researchers try to use human behavioral signals of likes, upvotes, and annotations of a response's helpfulness to improve content generation (Kreutzer *et al.*, 2018; Stiennon *et al.*, 2020; Ziegler *et al.*, 2019; Nakano *et al.*, 2021; SI *et al.*, 2025; Lee *et al.*, 2023; Wu *et al.*, 2023*b*; Khurana *et al.*, 2024, 2023). In our work, we propose a scalable approach to increase the content understanding abilities of VLMs, requiring minimal cost and no architectural changes.

Communication serves as a fundamental mechanism for achieving shared goals between senders and receivers (Smith and Harper, 2003). Humans possess a remarkable capacity to cooperate with strangers, enabled by language that has allowed our ancestors to exchange information, resolve conflicts, and create shared constructs like fictions, social structures, and cultural frameworks (Misyak *et al.*, 2016; McCroskey, 2015; Smith and Szathmary, 1997). This ability emerges early in human development, with children demonstrating communication and persuasion skills from a young age (Perner and Wimmer, 1985). Notably, strategic communication extends beyond human species, manifesting in both conspecific (Hare *et al.*, 2000; Smith and Harper, 2003) and interspecific (Krebs and Dawkins, 1984; Fouts *et al.*, 2002) interactions. A compelling example is the "broken wing display" observed across various bird genera, where adults feign injury to appear vulnerable, strategically luring predators away from their offspring (Griffin, 2001). Building on this foundation of strategic communication, the final chapter of this thesis (Chapter-5) demonstrates how modeling the complete communication workflow enables the generation of messages designed to elicit specific behavioral outcomes. We explore this concept across two modalities:

1. Text Domain: Through the illustrative case of memorability, we develop methods to generate content that demonstrates enhanced long-term retention (SI *et al.*, 2025).

2. Visual Domain: We advance techniques for generating images that achieve higher performance metrics, specifically focusing on engagement through social media likes (Khurana *et al.*, 2024).

In addressing these challenges, we make several key contributions: First, we in-

troduce UltraLAMBDA, the first large-scale advertisement dataset, comprising 5 million ads with automatically extracted content labels, including ASR transcriptions, captions, OCR text, emotion indicators, and memorability scores assigned by our model. Our analysis reveals that current large language models (LLMs) like GPT-3.5 and GPT-4 struggle to generate inherently memorable content. In response, we developed Henry, which demonstrates a 44% average improvement in memorability scores through progressive generation techniques. This work represents the first successful application of synthetic data to a domain previously lacking large-scale training resources.

Second, we address the critical need for engagement-optimized image generation, particularly relevant to industries such as advertising, fashion, and e-commerce, where user engagement metrics (clicks, likes, purchases) directly measure success. We present EngagingImageNet, a comprehensive dataset containing 168 million tweets collected from 10,135 enterprise accounts (2007-2023). This dataset includes rich metadata: account information, tweet text, media content, image captions, keywords, color analysis, posting timestamps, and engagement metrics.

Our analysis reveals that traditional image generation metrics (fidelity, aesthetics) show no correlation with actual engagement. To bridge this gap, we developed EngageNet, an engagement-aware vision language model (VLM) capable of predicting user engagement levels for images. Building on EngageNet's capabilities, we release Engagement Arena, the first automated benchmark for assessing the engagement potential of text-to-image models. This platform not only enables systematic comparison of existing models but also provides an open framework for the research community to evaluate and improve engagement-oriented image generation techniques.

1.1 Research Questions and Thesis Contributions

This thesis addresses a fundamental challenge in behavioral sciences: the fragmentation between explanation-focused and prediction-focused approaches to understanding human behavior. Current methods are limited in scope—models trained

for specific platforms, domains, or behavioral outcomes cannot generalize across different contexts. Our central research question is:

How can we develop general-purpose models capable of understanding, predicting, and optimizing human behavior that transfer across domains, platforms, and contexts?

This overarching question leads to several specific research questions that guide our investigation:

1. **Explanation**: How can we systematically extract and understand the mechanisms behind human behavioral responses to content? What strategies and triggers influence behavior, and how can we make these mechanisms interpretable to human decision-makers? We address this by constructing comprehensive taxonomies of persuasion strategies from theoretical literature, developing automated extraction methods for advertisements across image and video modalities, and introducing universal adversarial triggers (UATs) to probe what behavioral models learn, making their decision processes interpretable.
2. **Prediction**: How can we build models that predict human behavior by jointly modeling content and behavioral responses? What happens when we train models on the complete communication process rather than isolated components? We develop Large Content and Behavior Models (LCBM) that jointly model all seven communication factors, demonstrate emergent few-shot and zero-shot capabilities, and show that behavioral instruction fine-tuning on large-scale datasets leads to domain adaptation and improved generalization across platforms and contexts.
3. **Understanding**: Can information present in one modality (*viz.*, behavior) improve information extraction of another modality (*viz.*, content)? How can we leverage human behavioral responses (likes, shares, engagement patterns) to enhance models' comprehension of content across different modalities? We systematically integrate behavioral signals into content understanding pipelines, demonstrating improvements across more than 50 tasks spanning more than 25 benchmark datasets in text, audio, image, and video domains through scalable approaches.
4. **Generation**: Can we change one modality (*viz.*, content) to control another modality (*viz.*, behavior)? How can we generate content optimized for specific behavioral outcomes? Specifically, can we create memorable advertisements, engaging social media posts, or persuasive communications by understanding the content-behavior relationship? We develop progressive generation techniques for memorable text content, create engagement-aware vision-language models for image generation, and establish automated evaluation frameworks that measure content effectiveness based on predicted behavioral outcomes rather than traditional aesthetic metrics.

Our work makes several key contributions across four main areas:

Interpretable Behavior Analysis:

- We construct the largest taxonomy of persuasion strategies grounded in marketing and psychology literature, enabling systematic analysis of behavioral mechanisms across visual and textual content.
- We introduce universal adversarial triggers as a novel method for probing behavioral models, making their learned representations interpretable to human decision-makers.

Behavioral Modality Integration:

- We demonstrate that behavioral signals can serve as a universal modality and a core modality which always occurs together with content. We show that by understanding the behavior modality, we can improve content understanding; we show consistent improvements across diverse tasks in natural language processing, computer vision, and multimodal domains.
- We develop scalable methods for integrating human behavioral responses (gaze patterns, engagement metrics, social signals) into existing model architectures without requiring significant computational overhead.

Joint Content-Behavior Modeling:

- We show that jointly modeling content and behavioral modalities leads to emergent capabilities including few-shot behavior prediction, zero-shot domain adaptation, and improved generalization across platforms and contexts.
- We release comprehensive datasets spanning social media posts, video content, and advertisement corpora to enable systematic research in content-behavior relationships.

Behavior-Optimized Content Generation:

- We develop the first systematic approaches for generating content optimized for specific behavioral outcomes, including memorable advertisements and engaging visual content.
- We establish automated evaluation frameworks that assess content effectiveness based on predicted behavioral responses rather than traditional aesthetic or linguistic metrics.
- We demonstrate that synthetic content generation can be successfully applied to domains with limited training data by leveraging behavioral understanding.

Through this work, we aim to move beyond fragmented, platform-specific solutions toward a general understanding of human behavior that can be applied across diverse contexts, ultimately enabling more effective communication and content optimization.

Therefore, we will cover explanation, analysis, prediction, and generation aspects of behavior. We will cover the following works in this thesis:

1. MINIMAL: Mining models for data-free universal adversarial triggers. AAAI, 2022, (covered in Chapter-2)

2. Persuasion Strategies in Advertisements, AAAI, 2023, (covered in Chapter-2)

3. A Video Is Worth 4096 Tokens: Verbalize Videos To Understand Them In Zero Shot, EMNLP, 2023, **Nominated for best paper award** (covered in Chapter-2)

4. Large Content And Behavior Models To Understand, Simulate, And Optimize Content And Behavior, ICLR, 2024, **Nominated for best paper award** (covered in Chapter-3)

5. Synthesizing Human Gaze Feedback for Improved NLP Performance, EACL, 2023 (covered in Chapter-4)

6. Teaching Human Behavior Improves Content Understanding Abilities Of VLMs, ICLR, 2025 (covered in Chapter-4)

7. Long-Term Ad Memorability: Understanding and Generating Memorable Ads, WACV, 2025 (covered in Chapter-5)

8. Measuring And Improving Engagement of Text-to-Image Generation Models, ICLR, 2025 (covered in Chapter-5)

Chapter 2

Explaining Behavior: Persuasion Strategies and Universal Adversarial Triggers

Modeling what makes an advertisement persuasive, *i.e.*, eliciting the desired response from consumer, is critical to the study of propaganda, social psychology, and marketing. Despite its importance, computational modeling of persuasion in computer vision is still in its infancy, primarily due to the lack of benchmark datasets that can provide persuasion-strategy labels associated with ads. Motivated by persuasion literature in social psychology and marketing, as the first contribution, we introduce an extensive vocabulary of persuasion strategies and build the first ad corpus (both image and video) annotated with persuasion strategies (Section 2.1). We then formulate the task of persuasion strategy prediction with multi-modal learning. The image dataset also provides image segmentation masks, which labels persuasion strategies in the corresponding ad images on the test split. We publicly release our code and dataset at `https://midas-research.github.io/persuasion-advertisements/`. This section is based on two papers (Kumar *et al.*, 2023*a*; Bhattacharyya *et al.*, 2023).

Persuasion strategies offer a human-interpretable framework for understanding persuasion and its influence on behavior. Another focus of my work is to uncover what models trained to predict behavior actually learn. Advances in deep learning and large language models have significantly improved the accuracy of behavior prediction (Khandelwal *et al.*, 2024), but they have also increased the models' opacity, making them more like black boxes. Investigating what these *behavior prediction models understand about behavior* can provide insights that may be challenging to derive from first principles. This motivation underpins our other work (Singla *et al.*, 2022), discussed in Section 2.2. In this work, we mine behavioral prediction models to understand what they learn about behavior which makes them so successful at behavior prediction. This approach is complementary to the approach afforded by persuasion strategies. While persuasion strategies help

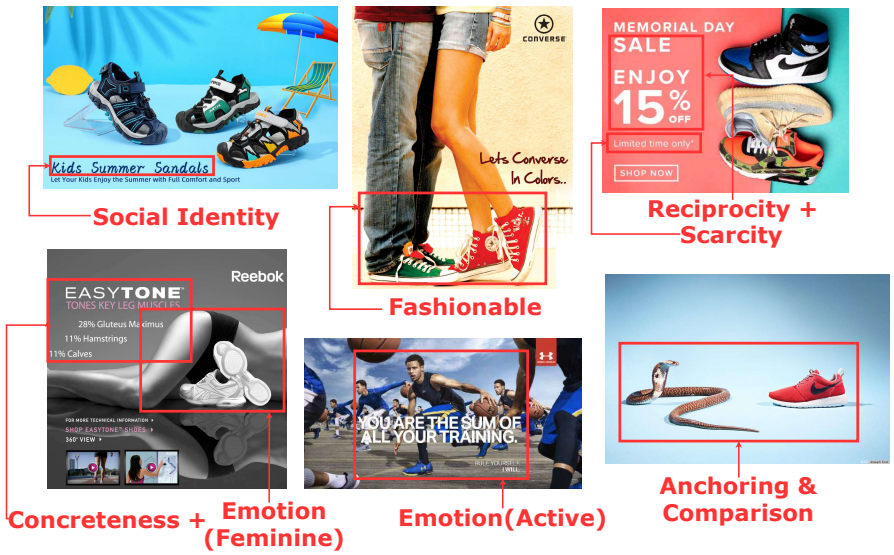


Figure 2.1: Different persuasion strategies are used for marketing the same product (footwear in this example). The strategies are in red words and to be defined by us in the paper.



Figure 2.2: Examples of videos with their annotated persuasion strategies. Relevant keyframes and ASR captions are shown in the figure, along with the annotated strategies. These two videos can be watched at https://bit.ly/3Ie3JG0, https://bit.ly/3OgtLwj .

humans summarize all messages to predict and understand behavior, our method, called universal adversarial triggers (UATs) help us understand a model's global behavior. We observe that content words with the lowest entropy appear as UATs. We explain this idea with the help of several models and datasets.

2.1 Persuasion Strategies In Advertisements

Marketing communications is the mode by which companies and governments inform, remind, and persuade their consumers about the products they sell. They are the primary means of connecting brands with consumers through which the

consumer can know what the product is about, what it stands for, who makes it, and can be motivated to try it out. To introduce meaning into their communication, marketers use various rhetorical devices in the form of persuasion strategies such as **emotions** (*e.g.*, Oreo's "Celebrate the Kid Inside", humor by showing Ronald McDonald sneaking into the competitor Burger King's store to buy a burger), **reasoning** (*e.g.*, "One glass of Florida orange juice contains 75% of your daily vitamin C needs"), **social identity** (*e.g.*, Old Spice's "Smell like a Man"), and **impact** (*e.g.*, Airbnb showing a mother with her child with the headline "My home is funding her future") (Refer to Fig. 2.3 to see these ads). Similarly, even for marketing the same product, marketers use different persuasion strategies to target different demographies (see Fig. 2.1). Therefore, recognizing and understanding persuasion strategies in ad campaigns is vitally important to decipher viral marketing campaigns, propaganda, and enable ad-recommendation.

Studying rhetorics of this form of communication is an essential part of understanding visual communication in marketing. Aristotle, in his seminal work on rhetoric, underlining the importance of persuasion, equated studying rhetorics with the study of persuasion* (Rapp, 2002). While persuasion is studied extensively in behavioral sciences, such as marketing (Meyers-Levy and Malaviya, 1999; Keller *et al.*, 2003) and psychology (Hovland *et al.*, 1953; Petty and Cacioppo, 1986), computational modeling of persuasion in computer vision is still in its infancy, primarily due to the lack of benchmark datasets that can provide representative corpus to facilitate this line of research. In the limited work that has happened on persuasion in computer vision, researchers have tried to address the question of which image is more persuasive (Bai *et al.*, 2021) or extracted low-level features (such as emotion, gestures, and facial displays), which indirectly help in identifying persuasion strategies without explicitly extracting the strategies themselves (Joo *et al.*, 2014). On the other hand, decoding persuasion in textual content has been extensively studied in natural language processing from both extractive, and generative contexts (Habernal and Gurevych, 2016; Chen and Yang, 2021*a*; Luu *et al.*, 2019). This forms the motivation of our work, where we aim to identify the persuasion strategies used in visual content such as advertisements.

*"Rhetoric may be defined as the faculty of discovering in any particular case all of the available means of *persuasion*" (Rapp, 2002)

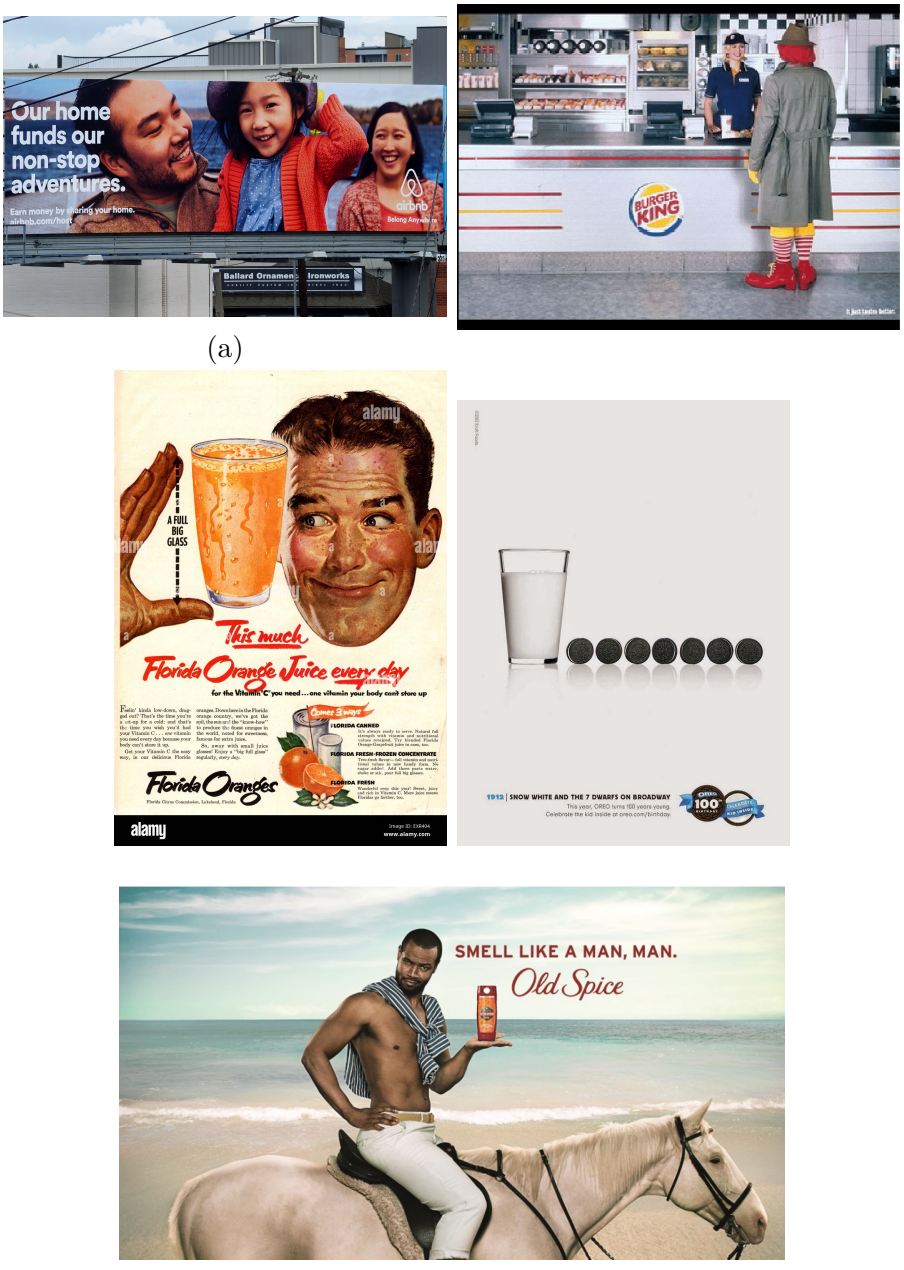


Figure 2.3: Example of various rhetoric strategies used in advertisements

The systematic study of persuasion began in the 1920s with the media-effects research by Lasswell (Lasswell, 1971), which was used as the basis for developing popular models of persuasion, like the Elaboration Likelihood Model (ELM) (Petty and Cacioppo, 1986), Heuristic Systematic Model (HSM) (Chaiken, 1980), and Hovland's attitude change approach (Hovland *et al.*, 1953). Laswell in this research broke down communication into five factors by defining communication as an act of *who* said it, *what* was said, in *what* channel it was said, to *whom* it was said, and with what *effect* it was said. Later, this model was used as the basis for developing popular models of persuasion, like Elaboration Likelihood Model (Petty and Cacioppo, 1986), Heuristic Systematic Model (Chaiken, 1980), and Hovland's attitude change approach (Hovland *et al.*, 1953). Amongst these, the most widely accepted model of persuasion theory is the Elaboration Likelihood Model (ELM).

These models of persuasion posit a dual process theory that explains attitude and behavior change (persuasion) in terms of the following major factors: stimuli (messages), personal motivation (the desire to process the message), capability of critical evaluation, and cognitive busyness. These factors could be divided into cognitive, behavioral, and affective processes of attitude change. Thus, a person may begin liking a new political candidate because she just donated $100 to the campaign (behavior-initiated change), because the theme music in a recently heard commercial induced a general pleasantness (affect-initiated change), or because the person was impressed with the candidate's issue positions (cognitive initiated change). Similarly, if a person already likes a political candidate he may agree to donate money to the campaign (behavioral influence), may feel happiness upon meeting the candidate (affective influence), and may selectively encode the candidate's issue positions (cognitive influence) (Petty and Cacioppo, 1986).

ELM posits that when facing a message from a persuader, the persuadee reacts by using the two information processing channels: central processing or peripheral processing. When the persuadee processes information centrally, the cognitive responses, or elaborations, will be much more relevant to the information, whereas when processing peripherally, the individual may rely on heuristics and other rules of thumb when elaborating on a message. The factors which influence how and how much one will elaborate the persuasive message is given by the message type, personal motivation, and other factors presented in the ELM. Being at the high

end of the elaboration continuum, people assess object-relevant information in relation to schemas that they already possess, and arrive at a reasoned attitude that is supported by information (Van Lange *et al.*, 2011).

In this chapter, we build on these psychological insights from persuasion models in sociology and marketing and study the message strategies that lead to persuasion. We codify, extend, and unify persuasion strategies studied in the psychology and marketing literature into a set of 20 strategies divided into 9 groups (see Fig. 2.4, Table 2.1.2): *Authority and Credibility*, *Social Identity and Proof*, where cognitive indirection in the form of group decisioning and expert authority is used for decisions, *Value and Impact Formulation* where logic is used to explain details and comparisons are made, *Reciprocity*, *Foot in the door*, *Overcoming Resistance* where social and cognitive consistency norms are harnessed to aid decision-making, *Scarcity*, *Anthropomorphism* and *Emotion* where information is evaluated from the lenses of feelings and emotions. In addition to introducing the most extensive vocabulary for persuasion strategies, we make a superset of persuasion strategies presented in the prior NLP works, which introduced text and domain-specific persuasion tactics, thus making large-scale understanding of persuasion across multiple contexts comparable and replicable.

Constructing a large-scale dataset containing persuasion strategies labels is time-consuming and expensive. We leverage active learning to mitigate the cost of labeling fine-grained persuasion strategies in advertisements. We first introduce an attention-fusion model trained in a multi-task fashion over modalities such as text, image, and symbolism. We use the action-reason task from the Pitts Ads dataset (Hussain *et al.*, 2017) to train the model and then annotate the raw ad images from the same dataset for persuasion strategies based on an entropy based active learning technique.

To sum up, our contributions include:

1. We construct the largest set of generic persuasion strategies based on theoretical and empirical studies in marketing, social psychology, and machine learning literature.
2. We introduce the first dataset for studying persuasion strategies in advertisements. This enables initial progress on the challenging task of automatically

understanding the messaging strategies conveyed through visual advertisements. We also construct a prototypical dataset containing image segmentation masks annotating persuasion strategies in different segments of an image.

3. We formulate the task of predicting persuasion strategies with a multi-task attention fusion model.

4. We conduct extensive experiments on the released corpus, showing the effect of different modalities on identifying persuasion strategies, correlation between strategies and topics and objects with different strategies.

2.1.1 Related Work

How do messages change people's beliefs and actions? The systematic study of persuasion has captured researchers' interest since the advent of mass influence mechanisms such as radio, television, and advertising. Work in persuasion spans across multiple fields, including psychology, marketing, and machine learning.

Persuasion in Marketing and Social Psychology: Sociology and communication science has studied persuasion for centuries now starting from the seminal work of Aristotle on rhetoric. Researchers have tried to construct and validate models of persuasion. Due to space constraints, while we cannot cover a complete list of literature, in Section 2.1.2, we list the primary studies which originally identified the presence and effect of various persuasion tactics on persuadees. We build on almost a century of this research and crystallize them into the persuasion strategies we use for annotation and modeling. Any instance of (successful) persuasion is composed of two events: (a) an attempt by the persuader, which we term as the persuasion strategies, and (b) subsequent uptake and response by the persuadee (Anand *et al.*, 2011; Vakratsas and Ambler, 1999). In this work, we study (a) only while leaving (b) for future work. Throughout the rest of the paper, when we say persuasion strategy, we mean the former without considering whether the persuasion was successful or not.



Figure 2.4: Persuasion strategies in advertisements. Marketers use both text and vision modalities to create ads containing different messaging strategies. Different persuasion strategies are constituted by using various rhetorical devices such as slogans, symbolism, colors, emotions, allusion.

Persuasion in Machine Learning: Despite extensive work in social psychology and marketing on persuasion, most of the work is qualitative, where researchers have looked at a small set of messages with various persuasion strategies to determine their effect on participants. Computational modeling of persuasion is still largely lacking. In the limited work in computational modeling of persuasion, almost all of it is concentrated in the NLP literature, with only very few works in computer vision. Research on persuasion in NLP under the umbrella of argumentation mining is broadly carried out from three perspectives: extracting persuasion tactics, studying the effect of constituent factors on persuasion, and measurement of persuasiveness nature of content. A few examples of research studies that annotate persuasive strategies in various forms of persuader-persuadee interactions like discussion forums, social media, blogs, academic essays, and debates are (Anand *et al.*, 2011; Tan *et al.*, 2016; Chen and Yang, 2021*b*). We use these and other studies listed in Section 2.1.2 to construct our vocabulary of persuasion strategies in advertisements.

Other studies focus on factors such as argument ordering (Shaikh *et al.*, 2020; Li *et al.*, 2020*a*), target audience (Lukin *et al.*, 2017), and prior beliefs (El Baff *et al.*, 2020) for their effect in bringing about persuasion. Studies such as (Althoff *et al.*, 2014; Wei *et al.*, 2016) also try to measure persuasiveness and generate persuasive content. The generation of persuasive (textual) messages has been studied (Donadello *et al.*, 2020) and, in particular, a novel ML method for learning user model tailored persuasion strategy has also been proposed (Hadoux *et al.*, 2021; Donadello *et al.*, 2022).

As one of the first works in the limited work in the computer vision domain, Joo *et al.* (Joo *et al.*, 2014) introduced syntactical and intent features such as facial displays, gestures, emotion, and personality, which result in persuasive images. Their analysis was done on human images, particularly politicians, during their campaigns. Their work on political campaigners is more restrictive than general product and public-service advertisements. Moreover, they deal with low-level features such as gestures and personality traits depicted through the face, which are important for detecting persuasion strategies but are not persuasion strategies themselves. Recently, Bai *et al.* (Bai *et al.*, 2021) studied persuasion in debate videos where they proposed two tasks: debate outcome prediction and intensity

of persuasion prediction. Through these tasks, they predict the persuasiveness of a debate speech, which is orthogonal to the task of predicting the strategy used by the debater. Other similar works which discuss persuasiveness of images and videos are (Joo *et al.*, 2015; Siddiquie *et al.*, 2015).

2.1.2 Generic Taxonomy of Persuasion Strategies

This section introduces the generic taxonomy of persuasive strategies, their definitions, examples, and connections with prior work. Representative literature from a) SPM: Social Psychology and Marketing, b) ML: Machine Learning

1. Authority and Credibility: SPM:(Aronson *et al.*, 1963; Milgram and Gudehus, 1978; Cialdini and Cialdini, 2007; Milgram, 1963; McGinnies and Ward, 1980; Giffin, 1967; Petty and Cacioppo, 1986) ML:(Anand *et al.*, 2011; Iyer and Sycara, 2019; Wachsmuth *et al.*, 2017; Chen and Yang, 2021*a*; Durmus and Cardie, 2018)
   (a) **Guarantees**: Guarantees reduce risk and people try out such products more often.
   (b) **Authority**: Authority indicated through expertise, source of power, third-party approval, credentials, and awards
   (c) **Trustworthiness**: Trustworthiness indicated honesty and integrity of the source through tropes like years of experience, "trusted brand", numbers and statistics

2. Social Identity and Proof: SPM:(Deutsch and Gerard, 1955; Petty *et al.*, 1997; Wood, 2000; Cialdini and Goldstein, 2004; Levesque and Pons, 2020) ML:(Anand *et al.*, 2011; Iyer and Sycara, 2019; Rosenthal and Mckeown, 2017; Yang *et al.*, 2019; Zhang *et al.*, 2016; Stab and Gurevych, 2017; Althoff *et al.*, 2014; Hidey *et al.*, 2017; Durmus and Cardie, 2018)
   (a) **Social Identity**: *Normative* influence, which involves conformity with the positive expectations of "another", who could be "another person, a group, or one's self" (includes self-persuasion, fleeting attraction, altercasting, and exclusivity)
   (b) **Social Proof**: *Informational influence* by accepting information obtained from others as evidence about reality, *e.g.*, customer reviews and ratings

3. Reciprocity: SPM:(Regan, 1971; Cialdini and Cialdini, 2007; Clark, 1984; Clark and Mills, 1979; Clark *et al.*, 1986) ML:(Anand *et al.*, 2011; Iyer and Sycara, 2019; Althoff *et al.*, 2014; Chen and Yang, 2021*a*; Shaikh *et al.*, 2020)
   (a) **Reciprocity**: By *obligating* the recipient of an act to repayment in the future, the rule for reciprocation begets a sense of future obligation, often unequal in nature

4. Foot in the door: SPM: (Freedman and Fraser, 1966; Burger, 1999; Cialdini and Cialdini, 2007) ML:(Chen and Yang, 2021*b*; Wang *et al.*, 2019; Vargheese *et al.*, 2020)
   (a) **Foot in the door**: Starting with small requests followed by larger requests to facilitate compliance while maintaining *cognitive coherence*.

5. Overcoming Resistance: SPM:(McGuire and Papageorgis, 1961; Knowles and Linn, 2004; McGuire, 1964) ML:{None}
   (a) **Overcoming Resistance**: Overcoming resistance (reactance) by postponing consequences to the future, by focusing resistance on realistic concerns, by forewarning that a message will be coming, by acknowledging resistance, by raising self-esteem and a sense of efficacy.

6. Value and Impact Formulation: SPM:(Lee *et al.*, 2010; Furnham and Boo, 2011; Wegener *et al.*, 2001; Tversky and Kahneman, 1974; Strack and Mussweiler, 1997; Bhattacharya and Sen, 2003) ML:(Zhang *et al.*, 2017; Longpre *et al.*, 2019)
   (a) **Concreteness**: Using concrete facts, evidence, and statistics to appeal to the logic of consumers
   (b) **Anchoring and Comparison**: A product's value is strongly influenced by what it is compared to.
   (c) **Social Impact**: Emphasizes the importance or bigger (societal) impact of a product

7. Scarcity: SPM: (Brehm, 1966; Lynn, 1991; Rothman *et al.*, 1999; Tversky and Kahneman, 1985) ML:(Yang *et al.*, 2019; Chen and Yang, 2021*a*; Shaikh *et al.*, 2020)
   (a) **Scarcity**: People assign more value to opportunities when they are less available. This happens due to psychological reactance of losing freedom of choice when things are less available or they use availability as a cognitive shortcut for gauging quality.

8. Anthropomorphism: SPM:(Fournier, 1998; Levesque and Pons, 2020; Epley *et al.*, 2007) ML:{None}
   (a) **Anthropomorphism**: When a brand or product is seen as human-like, people will like it more and feel closer to it.

9. Emotion: Aesthetics, feeling and other non-cognitively demanding features used for persuading consumers SPM:(Hibbert *et al.*, 2007; Petty and Cacioppo, 1986; Petty *et al.*, 1983)
   ML:(Yang *et al.*, 2019; Tan *et al.*, 2016; Hidey *et al.*, 2017; He *et al.*, 2018; Durmus and Cardie, 2018; Zhang *et al.*, 2017; Wachsmuth *et al.*, 2017)
   (a) **Amazed**
   (b) **Fashionable**
   (c) **Active, Eager**
   (d) **Feminine**
   (e) **Creative**

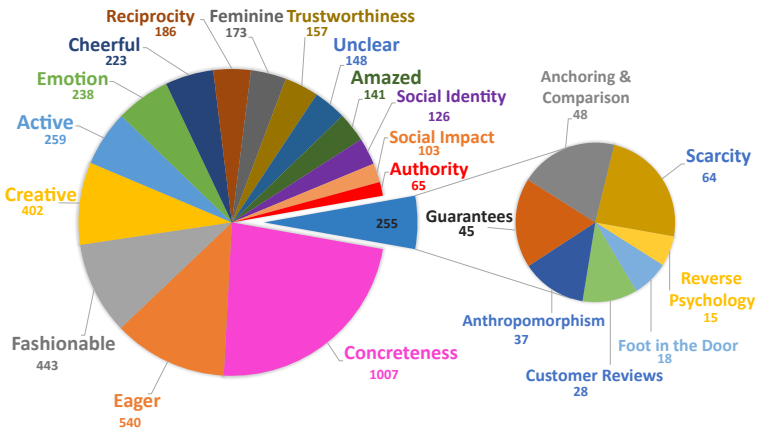


Figure 2.5: Distribution of Persuasion Strategies in the image persuasion strategy dataset. The top-3 strategies are Concreteness, Eager, and Fashionable.

(f) **Cheerful**

(g) **Further Minor**

10. **Unclear**: If the ad strategy is unclear

2.1.3 Persuasion Strategy Corpus Creation

2.1.3.1 Persuasion Strategy Dataset For Image Advertisements

To annotate persuasion strategies on image advertisements, we leverage raw images from the Pitts Ads dataset. It contains 64,832 image ads with labels of topics, sentiments, symbolic references (*e.g.* dove symbolizing peace), and reasoning the ad provides to its viewers (see Fig 2.10 for a few examples). The dataset has ads spanning multiple industries, products, services, and also contains public service announcements, making it representative of the diverse advertising landscape encountered in real-world applications such as social media platforms, websites, and mobile applications. This diversity is crucial for developing models that can generalize to practical scenarios including automated content analysis and marketing intelligence systems. Through this, they presented an initial work for the task of understanding visual rhetoric in ads. Since the dataset already had a few types of labels associated with the ad images, we used active learning on a model trained in a multi-task learning fashion over the reasoning task introduced in their paper. We explain the model and then the annotation strategy followed in §2.1.4.

To commence training, we initially annotated a batch of 250 randomly selected ad images with persuasion strategies defined in Section 2.1.2. We recruited four research assistants to label persuasion strategies for each advertisement. Definitions and examples of different persuasion strategies were provided, together with a training session where we asked annotators to annotate a number of example images and walked them through any disagreed annotations. To assess the reliability of the annotated labels, we then asked them to annotate the same 500 images and computed Cohen's Kappa statistic to measure inter-rater reliability. We obtained an average score of 0.55. The theoretical maximum of Kappa, given the unequal distribution, is 0.76. In such cases, Cohen (Cohen, 1960) suggested that one should divide kappa by its maximum value $\mathbf{k}/\mathbf{k_{max}}$, which comes out to be 0.72. This is a *substantial* agreement. Further, to maintain labeling consistency, each image was double annotated, with all discrepancies resolved by an intervention of the third annotator using a majority vote.

To address annotator diversity challenges and potential biases within our resource constraints, we implemented several strategies to handle annotation challenges: (1) regular calibration sessions were conducted where annotators discussed difficult cases and refined their understanding of strategy definitions, (2) we provided extensive examples covering edge cases and ambiguous scenarios to reduce subjective interpretation, and (3) we tracked individual annotator performance patterns to identify and correct systematic biases. Additionally, whenever cultural differences or ambiguities about strategy interpretation arose during the annotation process, these cases were brought up and discussed in group sessions to establish consensus and ensure consistent labeling across all annotators.

Despite these measures, we acknowledge that potential biases in annotation remain a concern for generalizability. Cultural biases may affect how annotators perceive certain persuasion strategies, particularly those related to social identity, authority, and emotional appeals, which can vary significantly across different cultural contexts. Demographic biases related to age, gender, and socioeconomic background may also influence the interpretation of strategies like social proof, reciprocity, and scarcity. These annotation biases could potentially limit the model's ability to generalize across diverse populations and cultural contexts, as the training data may reflect the specific cultural and demographic perspectives

of our annotator pool rather than universal patterns of persuasion strategy recognition.

The assistants were asked to label each image with no more than 3 strategies. If an image had more than 3 strategies, they were asked to list the top 3 strategies according to the area covered by the pixels depicting that strategy. In total, we label 3000 ad-images with their persuasion strategies; and the number of samples in train, val, and test split are 2500, 250, and 250, respectively[†]. Fig. 2.5 presents the distribution of persuasion strategies in the dataset. It is observed that concreteness is the most used strategy in the dataset, followed by eagerness and fashion. The average number of strategies in an ad is 1.49, and the standard deviation is 0.592. We find that scarcity (92.2%), guarantees (91.1%), reciprocity (84.4%), social identity (83.3%), and cheerful (83%), are the top 5 strategies, which occur in groups of 2 or 3. We observe that the co-occurrence of these strategies is due to the fact that many of them cover only a single modality (*i.e.*, text or visual), leaving the other modality free for a different strategy. For example, concreteness is often indicated by illustrating points in text, while the visual modality is free for depicting, say, emotion. See Fig. 2.6 for an example, where the image depicting *Authority* also has concreteness strategy in it. Similarly, feminine emotion is also depicted in Fig. 2.1, along with concreteness.

Next, we calculate the Dice correlation coefficient[‡] statistics for pairs of co-occurring persuasion strategies. The top-5 pairs are eager-concreteness (0.27), scarcity-reciprocity (0.25), eager-cheerful (0.19), amazed-concreteness (0.17), and eager-reciprocity (0.17). We find that these correlation values are not particularly high since marketers seldom use *common pairings* of messaging strategies to market their products. The visual part mostly shows eager strategy in ads; therefore, we find that the text modality becomes free to show other strategies. That is why primarily text-based concreteness, cheerfulness, and reciprocity strategies are present with the visual-based eager strategy in the text modality. Also, primarily vision-based amazement, eagerness, and scarcity (short-text) strategies co-occur with text-based reciprocity and concreteness (*e.g.*, Fig. 2.1).

[†]Table 2.4 shows the detailed distribution of the number of strategies in ads

[‡]The Dice Coefficient is defined as: $2 * |X \cap Y| \,/\, (|X| + |Y|)$, where X and Y are two sets; a set with vertical bars on either side refers to the cardinality of the set, i.e. the number of elements in that set; and $\cap$ refers to the intersection of two sets.

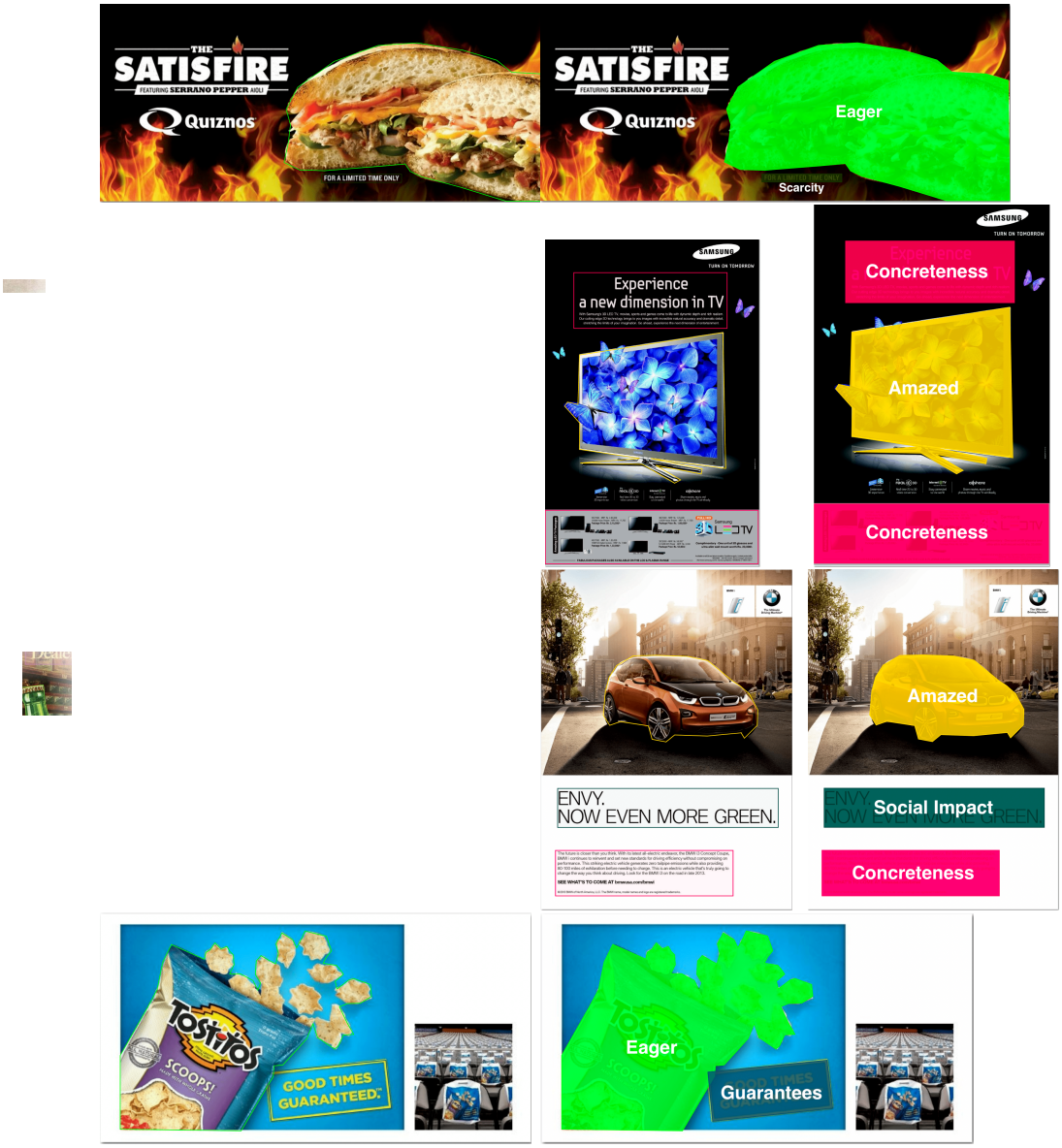


Figure 2.6: Image with a segmentation mask depicting the strategies *Emotion:Cheerful*, *Emotion:Eager* and *Trustworthiness*.

Next, we calculate the correlation between image topics and objects present with persuasion strategies. We see that the emotion:feminine and emotion:fashionable strategies are most often associated with beauty products and cosmetics (corr=0.4256, 0.2891). This is understandable since most beauty products are aimed at women. We see that the fast-food and restaurant industries often use eagerness as their messaging strategy (corr = 0.5877,0.3470). We find that the presence of humans in ads is correlated with the concreteness strategy (see Fig 2.7 for a few examples) (corr=0.3831). On the other hand, vehicle ads use emotion:amazed and concreteness (corr=0.5211,0.2412) (see Fig:2.8 for detailed correlations).

Similar to a low correlation in co-occurring strategies, we find that product



(a) (b) (c)

Figure 2.7: Advertisements containing humans and concreteness

segments and their strategies are not highly correlated. This is because marketers use different strategies to market their products even within a product segment. Fig. 2.1 shows an example in which the footwear industry (which is a subsegment of the apparel industry) uses different strategies to market its products. Further, for a batch of 250 images, we also label segmented image regions corresponding to the strategies present in the image. These segment masks were also double-annotated. Fig. 2.6 presents an example of masks depicting parts of the image masked with different persuasion strategies in a drink advertisement.

2.1.3.2 Persuasion Strategy Dataset For Video Advertisements

For this task, we collected 2203 video advertisements from popular brands publicly available on the web, reflecting the growing importance of video marketing on platforms like YouTube, Instagram, and TikTok where temporal dynamics and multimodal storytelling create complex persuasive narratives. We use the following 12 strategies as our target persuasion strategy set: *Social Identity, Concreteness, Anchoring and Comparison, Overcoming Reactance, Reciprocity, Foot-in-the-Door, Authority, Social Impact, Anthropomorphism, Scarcity, Social Proof,* and *Unclear*. This video-based approach complements our image dataset by capturing persuasion strategies that unfold over time, enabling applications in automated marketing content analysis and planning. We use non-experts human annotators to label this dataset (as compared to expert humans for the image ads dataset). In order

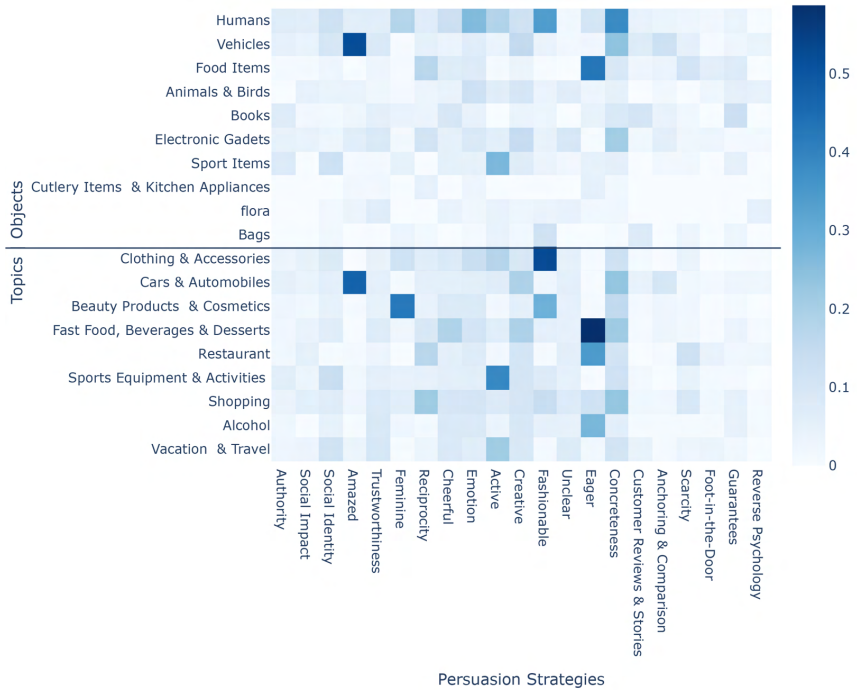


Figure 2.8: Dice correlation between topics and strategies. Topics are taken from the Pitts Ad dataset and further similar topics are combined to get these values.

to make the class labels easier to understand for non-expert human annotators, we make a list of 15 yes/no type-questions containing questions like "*Was there any expert (person or company) (not celebrity) encouraging to use the product/brand? Was the company showcasing any awards (e.g., industrial or government)? Did the video show any customer reviews or testimonials?*" (complete list in Table 2.1).

Each human annotator watches 15 videos such that each video gets viewed by at least two annotators and answers these questions for each video. Based on all the responses for a video, we assign labels to that video. We remove videos with an inter-annotator score of less than 60%. After removing those, we get a dataset with 1002 videos, with an average length of 33 secs and a distribution as shown in Fig. 2.9.

Similar to the image dataset (Kumar *et al.*, 2023*a*), potential annotation biases pose challenges for model generalizability in the video domain. The use of non-expert annotators, while providing broader accessibility to the annotation task, may introduce inconsistencies in strategy interpretation compared to expert annotations. Cultural and demographic biases may be particularly pronounced in video content, where temporal dynamics, audio cues, and visual narratives can

be interpreted differently across cultural contexts. For instance, strategies involving social impact, authority, or emotional appeals may be perceived differently by annotators from varying backgrounds, potentially affecting the model's ability to generalize to global audiences with diverse cultural perspectives on persuasive content. This dataset is then used for the persuasion strategy identification task.

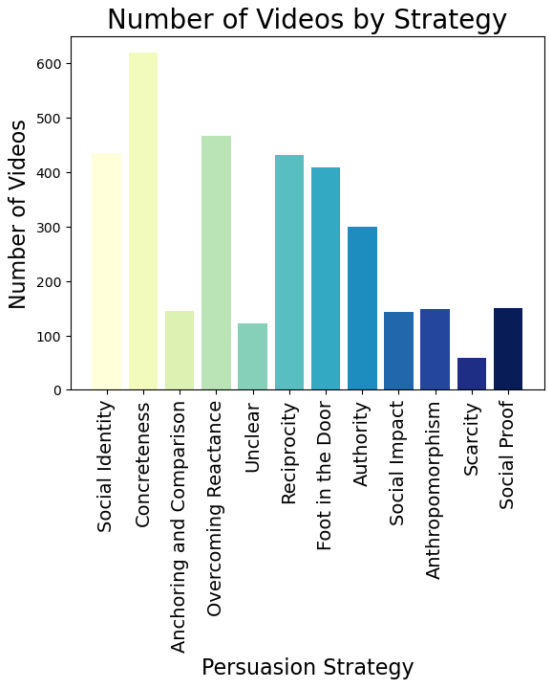


Figure 2.9: Distribution of persuasion strategies in our video persuasion strategy dataset

2.1.4 Modeling: Persuasion Strategy Prediction

2.1.4.1 Modelling Persuasion Strategy For Image Advertisements

The proposed Ads dataset $\mathcal{D}$ annotated with the persuasion strategies comprises of samples where each sample advertisement a_i is annotated with a set of annotation strategies S_i such that $1 \leq |S_i| \leq 3$. The unique set of the proposed persuasion strategies $\mathcal{P}$ is defined in Table 2.1.2. Given a_i, the task of the modeling is to predict the persuasion strategies present in the input ad. As we observe from Fig. 2.4, advertisements use various rhetoric devices to form their messaging strategy. The strategies thus are in the form of multi-modalities, including images, text and symbolism. To jointly model the modalities, we design an attention fusion multi-modal framework, which fuses multimodal features extracted

Question	Strategy	Question	Strategy
Was there any expert (person or company) (not celebrity) encouraging to use the product/brand?	Authority	Did the video show any normal customers (non-expert, non-celebrity) using the product?	Social Identity
Did the video showcase any awards or long usage history of the product/brand?	Authority	Did the video show any customer reviews or testimonials?	Social Proof
Was the product/brand comparing itself with other competitors or existing solutions?	Anchoring and Comparison	Were any number/statistics mentioned?	Concreteness
Did the video talk about any specific features or provide information about the product/brand?	Concreteness	Were there any mention of any offers on the brand/product?	Reciprocity
Were the offers limited or available for a short period of time?	Scarcity	Was the product/brand told to be free or available on a discount?	Foot in the Door, Reciprocity
Was the brand/product described as simple, easy to use, or can start using with minimal resistance?	Overcoming Reactance, Foot in the Door	Was the brand/product talking about bigger societal impact?	Social Impact
Did the brand provide any guarantees that might help reduce the risk of people trying out the product?	Overcoming Reactance	Did the video provide any resources, tips, guides, or tools related to the product?	Reciprocity
Is the brand or product portrayed as human-like?	Anthropomorphism		

Table 2.1: The questions we asked to the non-expert annotators to help them identify persuasion strategy contained in the video advertisement.

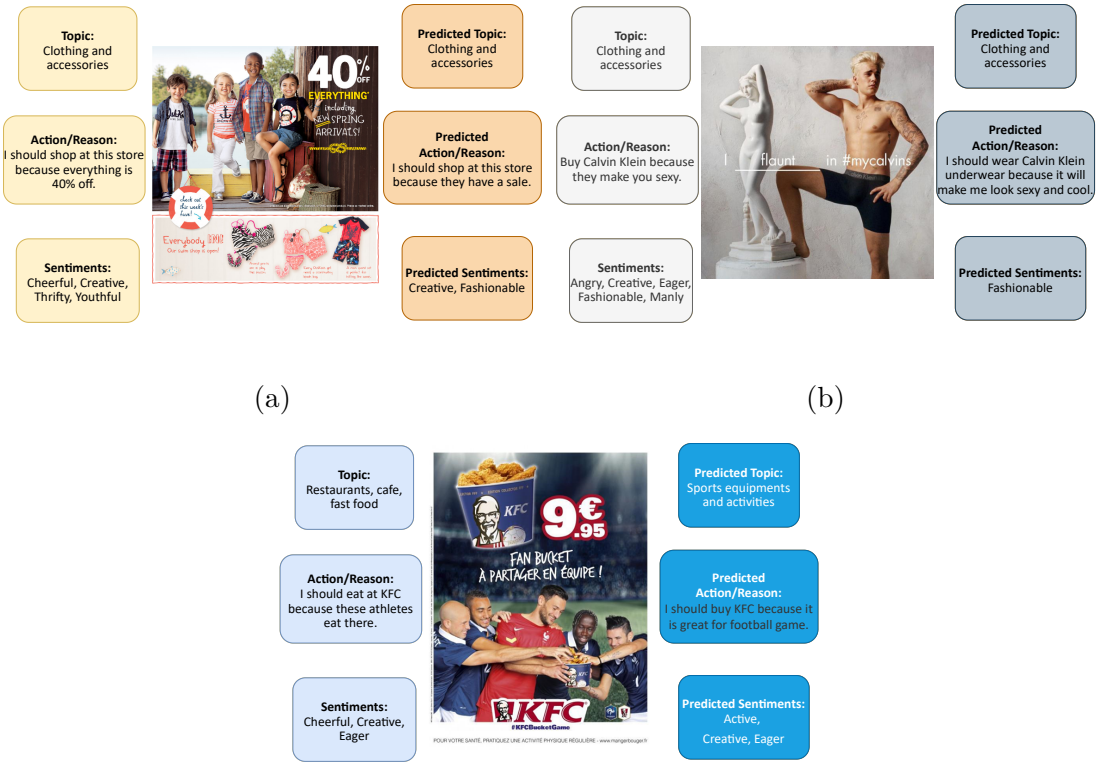


Figure 2.10: Some samples from the Pitts Ads dataset along with the ground truth and predicted action-reason statement, topic and sentiments.

from the ad, *e.g.*, the ad image, text present in the ad extracted through the OCR (Optical Character Recognition), regions of interest (ROIs) extracted using an object detector, and embeddings of captions obtained through an image captioning model (see Fig. 2.11). The information obtained through these modalities are firstly embedded independently through their modality specific encoders followed by a transformer-based cross-attention module to fuse the extracted features from different modalities. The fused embeddings from the attention module are then used as input for a classifier that predicts a probability score for each strategy $p \in \mathcal{P}$. The overall architecture of the proposed model is illustrated in Fig.2.11. In the following, we describe each step in the prediction pipeline in detail.

2.1.4.1.1 Feature Extractors In order to capture different rhetoric devices, we extract features from the image, text, and symbolism modalities.

Image Feature: We use the Vision Transformer (Dosovitskiy *et al.*, 2020) (ViT) model for extracting image features from the entire input image. The

model resizes the input image to size 224×224 and divides it into patches of size 16×16. The model used has been pre-trained on the ImageNet 21k dataset. We only use the first output embedding, which is the CLS token embedding, a 768 dimension tensor, as we only need a representation of the entire image. Then, a fully connected layer is used to reduce the size of the embedding, resulting in a tensor of dimension 256.

Regions of Interest (RoIs) from Detected Objects and Captions: Ad images contain elements that the creator deliberately chooses to create *intentional impact* and deliver some *message* in addition to the ones that occur *naturally* in the environment. Therefore, it is important to identify the composing elements of an advertisement to understand the creator's intention and the ad's message to the viewer. We detect and extract objects as regions of interest (RoIs) from the advertisement images. We get the RoIs by training the single-shot object detector model (Liu *et al.*, 2016) on the COCO dataset (Lin *et al.*, 2014). We compare it with the recent YOLOv5 model (Redmon *et al.*, 2016). We also extract caption embeddings to detect the most important activity from the image using a caption generation mode. We compare DenseCap (Yang *et al.*, 2017) and the more recent BLIP (Li *et al.*, 2022) for caption generation.

OCR Text: The text present in an ad presents valuable information about the brand, such as product details, statistics, reasons to buy the product, and creative information in the form of slogans and jingles that the company wants its customers to remember and thus making it helpful in decoding various persuasion strategies. Therefore, we extract the text from the ads and use it as a feature in our model. We use the Google Cloud Vision API for this purpose. All the extracted text is concatenated, and the size is restricted to 100 words. We pass the text through a BERT model and concatenate the embeddings for those 100 words. Similar to image embeddings, an FC layer is used to convert embeddings to 256 dimensions. The final embedding of the OCR is a tensor of dimension 100×256.

Symbolism: While the names of the detected objects convey the names or literal meaning of the objects, creative images often also use objects for their symbolic and figurative meanings. For example, an upward-going arrow repre-

sents growth or the north direction or movement towards the upward direction depending on the context; similarly, a person with both hands pointing upward could mean danger (*e.g.*, when a gun is pointed) or joy (*e.g.*, during dancing). In Fig. 2.4, in the creative Microsoft ad, a symbol of a balloon is created by grouping multiple mice together. Therefore, we generate symbol embeddings to capture the symbolism behind the most prominent visual objects present in an ad. We use the symbol classifier by Hussain *et al.* (Hussain *et al.*, 2017) on ad images to find the distribution of the symbolic elements present and then convert this to a 256 dimension tensor.

2.1.4.1.2 Cross-Modal Attention To capture the inter-dependency of multiple modalities for richer embeddings, we apply a cross-modal attention (CMA) layer (Frank *et al.*, 2021) to the features extracted in the previous steps. Cross-modal attention is a fusion mechanism where the attention masks from one modality (*e.g.* text) are used to highlight the extracted features in another modality (*e.g.* symbolism). It helps to link and extract common features in two or more modalities since common elements exist across multiple modalities, which complete and reinforce the message conveyed in the ad. For example, the pictures of the silver cup, stadium, and ball, words like "Australian", "Pakistani", and "World Cup" present in the chips ad shown in Fig. 2.11 link the idea of buying *Lays* with supporting one's country's team in the World Cup. Cross attention can also generate effective representations in the case of missing or noisy data or annotations in one or more modalities (Frank *et al.*, 2021). This is helpful in our case since marketing data often uses implicit associations and relations to convey meaning.

The input to the cross-modal attention layer is constructed by concatenating the image, RoI, OCR, caption, and symbol embeddings. This results in a 114×256 dimension input to our attention layer. The cross-modal attention consists of two layers of transformer encoders with a hidden dimension size of 256. The output of the attention layer gives us the final combined embedding of our input ad. Given image embeddings E_i, RoI embeddings E_r, OCR embeddings E_o, caption embeddings E_c and symbol embeddings E_s, the output of the cross-attention layer

E_{att} is formulated as:

$$\text{Enc}(X) = \text{CMA}([E_i(X), E_r(X), E_o(X), E_c(X), E_s(X)]) \,,$$

where $[\ldots, \ldots]$ is the concatenation operation. For the advertisement in Fig. 2.11, we observed that the caption "grey cup on the field" attends to OCR text (containing words like "win") and ViT features of the RoI (of "cup" and "field").

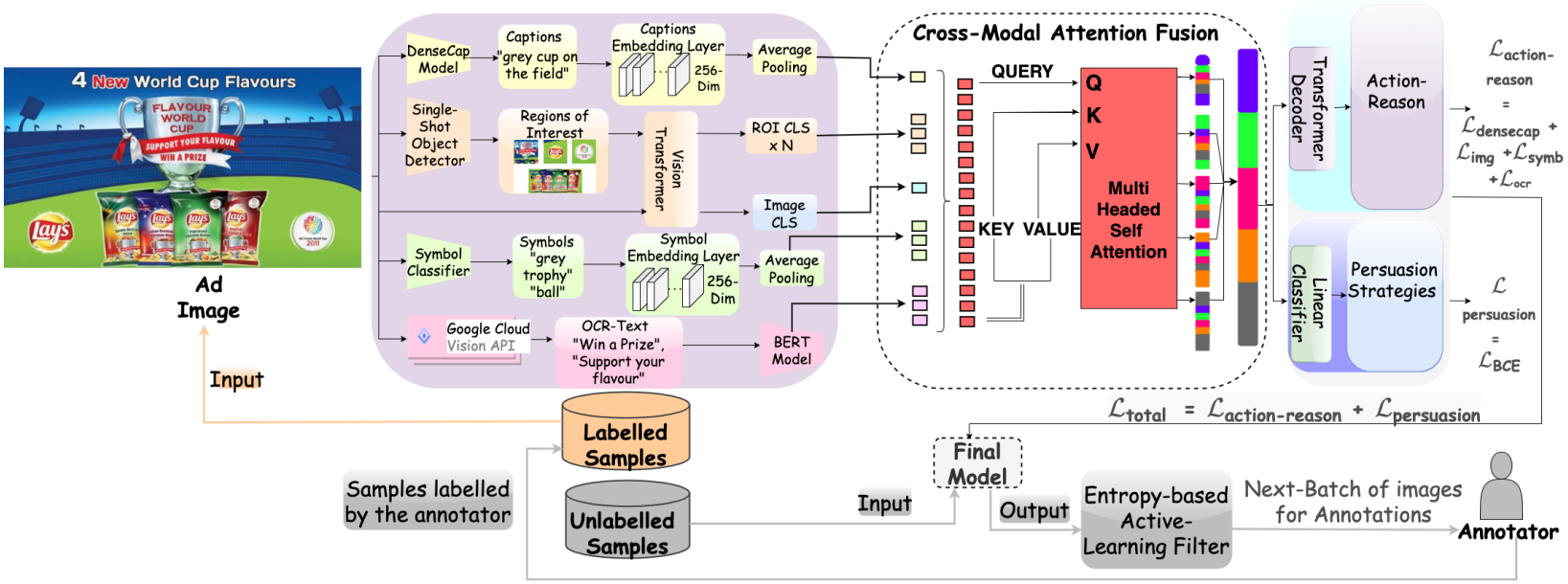


Figure 2.11: Architecture of the Persuasion Strategy Prediction model. To capture the different rhetoric devices, we extract features for the image, text, and symbolism modalities and then apply cross-modal attention fusion to leverage the interdependence of the different devices. Further, the model trains over two tasks: persuasion strategies and the reasoning task of action-reason prediction.

2.1.4.1.3 Persuasion Strategy Predictor This module is a persuasion strategy predictor, which processes the set of feature embedding Enc(X) obtained through cross-modality fusion. Specifically, Enc(X) is passed through a self-attention layer as:

$$o_1 = softmax(\text{Enc}(X) \otimes W_{self-attn})^\intercal \otimes \text{Enc}(X) \tag{2.1}$$

where Enc(X) is of the dimension 114×256, $W_{self-attn} \in \mathcal{R}^{256\times 1}$, $\otimes$ denote tensor multiplication and o_1 denotes the output of self attention layer, which is further processed through a linear layer to obtain $o_{|\mathcal{P}|}$ to represent the logits for each persuasion strategy. We apply sigmoid over each output logit such that the i^{th} index of the vector after applying sigmoid denotes p_i - the probability with which i^{th} persuasion strategy is present in the ad image. Our choice of using sigmoid over softmax is motivated by the fact that multiple persuasion strategies can be present simultaneously in an ad image. Consequently, the entire model is trained in an end-to-end manner using binary cross-entropy loss $\mathcal{L}_s$ over logit for each strategy:

$$\mathcal{L}_s = [-y_i \log(p_i) - (1 - y_i) \log(1 - p_i)] \tag{2.2}$$

where, y_i is 1 if i^{th} persuasion strategy is present in the ad and 0 otherwise. It can be observed in Table 2.2 that our model achieves an accuracy of 59.2%, where a correct match is considered if the strategy predicted by the model is present in the set of annotated strategies for a given ad. Further, we perform several ablations where we exclude each modality while retaining all the other modalities. We note that for each modality, excluding the modality results in a noticeable decrease in accuracy, with significant decreases observed when excluding DenseCap ($\sim$ 3.6%) and OCR ($\sim$ 4.4%). Further, we observe that using DenseCap for obtaining caption embeddings, and SSD for object detection works better than BLIP and YOLOv5, respectively (see Table 2.3). We also explore using focal loss (Lin *et al.*, 2017) in place of cross-entropy loss to handle class imbalance but observed that it led to degradation instead of improvements (top-1 acc.[§] of 56.4% *vs* 59.2%

[§] *Top-1 Accuracy*: It is defined as the fraction of images, where the highest predicted strategy is present in the ground-truth strategies. *Top-3 Accuracy* : It is defined as the fraction of images, where any of the top-3 highest predicted strategies is present in the ground-truth strategies.

Models	Top-1 Acc.	Top-3 Acc.
Our Model	**59.2**	**84.8**
w/o DenseCap	55.6	80.8
w/o Symbol	58.8	81.6
w/o DenseCap & Symbol	55.2	80.8
w/o OCR	54.8	82
w/o Symbol, OCR & DenseCap	58	78.8
w/o Action-Reason Task	56.4	80.4
Random Guess	6.25	18.75

Table 2.2: Effect of different Modalities and Tasks on the accuracy and performance of the strategy prediction task.

using cross-entropy). We also train the model of Hussain *et al.* (Hussain *et al.*, 2017) for strategy prediction through a similar configuration as ours (along with action-reason generation using an LSTM branch). We find that their top-1 and top-3 accuracy is 52.4% (vs. 59.2% ours) and 75.7% (vs. 84.8% ours), which is lesser compared to our model.

2.1.4.1.4 Multi Task Learning One of the key opportunities for our persuasion strategies data labeling and modeling task was the presence of additional labels already given in the base Pitts Ads dataset. In that, authors had given labels about the reasoning task. For the reasoning task, the annotators were asked to provide answers in the form "I should [Action] because [Reason]." for each ad. In other words, they asked the annotators to describe *what the viewer should do and why*, according to the ad. Similar to the reasoning task, persuasion strategies provide various cognitive, behavioral, and affective reasons to try to elicit the motivation of the ad viewers towards their products or services. Therefore, we hypothesize that these natural language descriptions of *why the viewers should follow* the ad will be informative in inferring the ad's persuasion strategy.

We formulate obtaining action-reason statement as a sequence generation task where the model learns to generate a sentence $Y^g = (y_1^g, \ldots, y_T^g)$ of length T conditioned on advertisement X by generating the sequence of tokens present in the action-reason statement. To achieve this, we use a transformer decoder module that attends on the features Enc(X) as shown in Fig. 2.11. The annotated action-reason statement is used to train the transformer decoder as an auxiliary task to

Model Used	Top-1 Accuracy	Top-3 Accuracy	Recall
Model with DenseCap & SSD	59.2	84.8	74.59
Model with BLIP & YOLOv5	58.4	83.8	71.58

Table 2.3: Comparison of caption and object detection models. We noticed that BLIP while being more recent and trained on a larger dataset, generates more informatory captions for background objects which DenseCap successfully ignores.

	#ads with 1 strategy	#ads with 2 strategies	#ads with 3 strategies	Avg. strategies	Std. Dev.
Train-Set	1440	905	155	1.486	0.612
Val-Set	132	98	20	1.552	0.639
Test-Set	147	93	10	1.452	0.574
Total	1719	1096	185	1.49	0.592

Table 2.4: Distribution of test, train, validation, and the total dataset

strategy prediction through the standard teacher forcing technique used in Seq2Seq framework. Please refer to the Supplementary for more architectural details about the action-reason generation branch. As shown in Table 2.2, generating action-reason as an auxiliary task improves the strategy prediction accuracy by 2.8%. We evaluate the performance on action-reason generation on following metrics: BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE, CIDER, SPICE and observed a score of 53.6, 42.0, 33.1, 25.7, 26.3, 48.4, 42.8, 8.9 respectively.

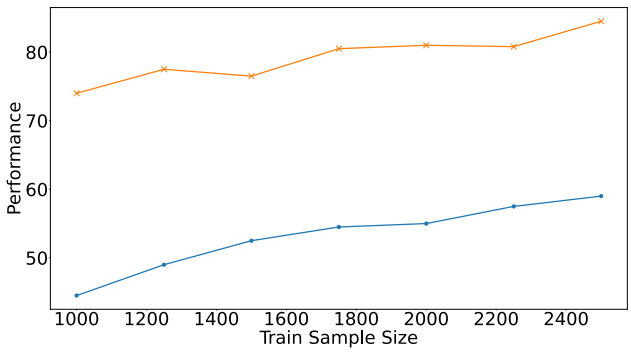


Figure 2.12: Incremental effect of introducing new data through active learning; Results for prediction of persuasion strategies on the test set

2.1.4.1.5 Active Learning We use an active learning method to ease the large-scale label dependence when constructing the dataset. As in every active

learning setting, our goal is to develop a learner that selects samples from unlabeled sets to be annotated by an oracle. Similar to traditional active learners (Gilad-Bachrach *et al.*, 2005; Lewis and Catlett, 1994), we use uncertainty sampling to perform the sample selection. In doing so, such function learns to score the unlabeled samples based on the expected performance gain they are likely to produce and used to update the current version of the localization model being trained. To evaluate each learner, we measure the performance improvements, assessed on a labeled test set at different training dataset sizes.

At every learning step t, a set of labeled samples L_t is first used to train a model f_t. Then, from an unlabeled pool $U_t = D - L_t$, an image instance a is chosen by a selection function g. Afterwards, an oracle provides temporal ground-truth for the selected instance, and the labeled set L_t is augmented with this new annotation. This process repeats until the desired performance is reached or the set U_t is empty.

In our implementation, we instantiate the active learning selection function as the entropy of the probability distribution predicted by the model over the set of persuasion strategies for a given ad image instance a. Formally, $g = -\sum_{i=1}^{|\mathcal{P}|} p_i^n * \log(p_i^n)$, where p_i^n denotes the normalized probability with which i^{th} persuasion strategy is present in a as per the model prediction. The normalized probability p_i^n is estimated as $p_i / \sum_{j=1}^{|\mathcal{P}|} p_j$. Intuitively, ad samples with high entropy selection values indicate that the model trained on limited data has a higher degree of confusion while predicting the persuasion strategy since it is not decisively confident about predicting few strategies. Hence, we rank the unlabeled ad images in the decreasing order of difficulty according to the corresponding values of the entropy selection function and select the top-k ads in the subsequent batch for annotation followed by training. As shown in Fig. 2.12, we set k to be 250 and analyze the effect of incrementally introducing new samples selected through active learning. It can be seen that both top-1 and top-3 accuracy increases with the addition of new training data. We stop at the point when 2500 training samples are used since the model performs reasonably well with a top-1 and top-3 strategy prediction accuracy of 59.2% and 84.8% (see Fig. 2.12).

For sample selection, we rank the ads using the model's entropy as the ranking

function. We choose samples with the highest entropy, indicating the model's confusing in predicting strategies. The ad samples with top-k entropy (where k=250) are selected for annotation in each batch (with the entropy thresholds for each batch being determined by the entropy of the kth sample). The active learning process stops when the model achieves reasonable performance (thus determining the convergence criterion), with a top-1 accuracy of 59.2% and top-3 accuracy of 84.8% after training on 2,500 samples. Incremental improvements in accuracy are observed as new data is introduced through active learning.

2.1.4.2 Modelling Persuasion Strategy For Video Advertisements

Large Language Models (LLMs) have been demonstrated to perform well for downstream classification tasks in the text domain. This powerful ability has been widely verified on natural language tasks, including text classification, semantic parsing, mathematical reasoning, *etc.* Inspired by these advances of LLMs, we aim to explore whether they could tackle reasoning tasks on multimodal data (*i.e.* videos). Therefore, we propose a storytelling framework, which leverages the power of LLMs to verbalize videos in terms of a text-based story and then performs downstream video understanding tasks on the generated story instead of the original video. Our pipeline can be used to verbalize videos and understand videos to perform complex downstream tasks such as emotion, topic, and persuasion strategy detection.

We show the performance of our framework on fifteen distinct tasks across five datasets. Firstly, we employ a video story dataset to evaluate the story generation task. Secondly, we utilize a video advertisements dataset to assess topic and emotion classification, as well as action and reason generation. Then, the persuasion strategy dataset to evaluate the task of understanding persuasion strategies within stories, and finally, HVU and LVU for concept, user engagement, and attribute prediction. These diverse datasets allow us to evaluate the performance and capabilities of our framework thoroughly.

1. The Video story dataset (Li *et al.*, 2020*b*) contains 105 videos, from four types of common and complex events (*i.e.* birthday, camping, Christmas, and wedding) and corresponding stories written by annotators. It has longer

videos (average length 12.4 mins) and longer descriptions (162.6 words on average). Moreover, the sentences in the dataset are more sparsely distributed across the video (55.77 sec per sentence). *Metrics*: Following (Li *et al.*, 2020*b*), we use several NLP metrics, *viz.*, BLEU-N, ROUGE-L, METEOR and CIDEr to measure the similarity between the story generated by the model and ground truth.

2. The Image and Video Advertisements (Hussain *et al.*, 2017) contains 3,477 video advertisements and the corresponding annotations for emotion and topic tags and action-reason statements for each video. There are a total of 38 topics and 30 unique emotion tags per video. Further, we have 5 action-reason statements for each video for the action-reason generation task. For our experiment, we use 1785 videos, due to other videos being unavailable/privated from Youtube.

Metrics: Following (Hussain *et al.*, 2017), for the topic and emotion classification task, we evaluate our pipeline using top-1 accuracy as the evaluation metric. Further, since (Hussain *et al.*, 2017) did not use any fixed set of vocabulary for annotations, rather they relied on annotator-provided labels, the labels are often very close (like cheerful, excited, and happy). Therefore, based on nearness in Plutchik's (PLUTCHIK, 1980) wheel of emotions, we club nearby emotions and use these seven main categories: joy, trust, fear, anger, disgust, anticipation, and unclear. For the action-reason task, following (Hussain *et al.*, 2017), we evaluate our accuracy on the action and reason retrieval tasks where 29 random options along with 1 ground truth are provided to the model to find which one is the ground truth. Further, we also generate action and reason statements and evaluate the generation's faithfulness with the ground truth using metrics like ROUGE, BLEU, CIDEr, and METEOR.

3. Persuasion strategy dataset: This is the dataset we contribute for understanding persuasion strategies.

Metrics: We evaluate the performance using top-1 accuracy metric. Videos have a varied number of strategies, therefore, we consider a response to be correct if the predicted strategy is present among the list of ground-truth strategies.

4. Long-Form Video Understanding (LVU): We *et al.* (Wu and Kra-

henbuhl, 2021) released a benchmark comprising of 9 diverse tasks for long video understanding and consisting of over 1000 hours of video. The various tasks consist of content understanding ('relationship', 'speaking style', 'scene/place'), user engagement prediction ('YouTube like ratio', 'YouTube popularity'), and movie metadata prediction ('director', 'genre', 'writer', 'movie release year'). We *et al.* (Wu and Krahenbuhl, 2021) use top-1 classification accuracy for content understanding and metadata prediction tasks and MSE for user engagement prediction tasks.

5. Holistic Video Understanding (HVU): HVU (Diba *et al.*, 2020) is the largest long video understanding dataset consisting of 476k, 31k, and 65k samples in train, val, and test sets, respectively. A comprehensive spectrum includes the identification of various semantic elements within videos, consisting of classifications of scenes, objects, actions, events, attributes, and concepts. To measure performance on HVU tasks, similar to the original paper, we use the mean average precision (mAP) metric on the validation set.

Next, we explain our pipeline to solve these tasks.

2.1.4.2.1 Video Verbalization To obtain a verbal representation of a video, we employ a series of modules that extract unimodal information from the multimodal video. This information is then used to prompt a generative language model (such as GPT-3.5 (Brown *et al.*, 2020) and Flan-t5 (Chung *et al.*, 2022)) to generate a coherent narrative from the video. The overall pipeline is depicted in Fig. 2.14. In the following, we delve into each component of the framework in details.

1. Video Metadata: Understanding the context of a story is crucial, and we achieve this by gathering information about the communicator (brand). We leverage the publicly available video title and channel name from the web. Additionally, we utilize Wikidata (Vrandečić and Krötzsch, 2014), a collaborative knowledge base that provides comprehensive data for Wikipedia, to obtain further details such as the company name, product line, and description. This information helps us comprehend the story elements and establish connections with the brand's business context. For non-advertisement videos, we skip this step and retrieve only

the video title.

2. Text Representation of Video Frames: We extract two types of textual information from video frames. Firstly, we capture the literal text present on the frames. Secondly, we analyze the scene depicted in each frame to gain a deeper understanding. In the upcoming sections, we will elaborate on both of these aspects.

Method	Frame Extraction	METEOR	CIDEr	Rougle-l	BLEU-1	BLEU-2	BLEU-3	BLEU-4
GPT-3.5	Uniform Sampling	24.8	102.4	24.3	63.8	56.4	47.2	38.6
GPT-3.5	Pyscenedetect	24.17	67.8	21.17	54.59	49.05	41.54	33.88

Table 2.5: Comparison of Pyscenedetect (Breakthrough, 2023) with uniform sampling of choosing video frames. Based on downstream performance, we can see that uniform sampling works better than Pyscenedetect

a. Visual and Scenic Elements in Frames: For videos with a duration shorter than 120 seconds, we employ an optical flow-based heuristic using the GMFlow model (Xu *et al.*, 2022*a*) to extract keyframes. In shorter advertisement videos, scene changes often indicate transitions in the story, resulting in keyframes with higher optical flow values. The GMFlow model effectively captures these story transitions. We select frames with an optical flow greater than 50 and prioritize frames with maximum pixel velocity. However, for longer videos, this approach yields a large number of frames that are difficult to accommodate within a limited context. To address this, we sample frames at a uniform rate based on the native frames-per-second (fps) of the video (see Table 2.5 for a comparison between uniform sampling and Pyscenedetect). Additionally, we discard frames that are completely dark or white, as they may have high optical flow but lack informative content.

Using either of these methods, we obtain a set of frames that represent the events in the video. These frames are then processed by a pretrained BLIP-2 model (Li *et al.*, 2023*a*). The BLIP model facilitates scene understanding and verbalizes the scene by capturing its most salient aspects. We utilize two different prompts to extract salient information from the frames. The first prompt, "*Caption this image*", is used to generate a caption that describes what is happening in the image, providing an understanding of the scene. The second prompt, "*Can you tell the objects that are present in the image?*", helps identify and gather information



Transcript : well hey Zuma that's the wrong pinata …. how old are you …. my birthday cake is its zoo …. paw patrol you …. got as a present … Thank You Marshall…. are you ready to blow out the candle…. Marshall let's push chase and the pool … Wow all the surprises we got…. Go first okay put on the blindfold ready …… for more fun click the red subscribe button

Generated Story : A family had the perfect idea for a Paw Patrol-themed birthday party. The kids had a blast blowing bubbles in the backyard and playing in the pool with a mascot in a Paw Patrol costume. The birthday boy got to play with a pinata in the backyard and a hot air balloon flew over the house. The kids played with a toy bike and a Paw Patrol playhouse, and the birthday girl hugged a mascot. They also set up a LEGO train set and a DC Comics mascot in the pool. The birthday party ended with a game of catch and a pin the badge on the mascot game. Everyone had a great time, and the birthday cake was a Paw Patrol-themed delight. This vlog shows that family fun can be had at any birthday party with a little imagination.

Topic : Celebrations	**Emotion** : Cheerful	**Persuasion Strategy**: Social Identity, Social Proof
Action : Consider using your imagination to create a fun and memorable birthday party for your family or friends.	**Reason** : The reason is to create a uniquely fun-filled and memorable birthday party for your family or friends	

Figure 2.13: An example of a story generated by the proposed pipeline along with the predicted outputs of the video-understanding tasks on the generated story. The generated story captures information across scenes, characters, event sequences, dialogues, emotions, and the environment. This helps the downstream models to get adequate information about the video to reason about it correctly. The original video can be watched at `https://youtu.be/_amwPjAcoC8`.

about the objects depicted in each frame.

b. Textual elements in frames: We also extract the textual information present in the frames, as text often reinforces the message present in a scene and can also inform viewers on what to expect next (Wang *et al.*, 2021*a*). For the OCR module, we sample every 10th frame extracted at the native frames-per-second of the video, and these frames are sent to PP-OCR (Vrandečić and Krötzsch, 2014). We filter the OCR text and use only the unique words for further processing.

3. Text Representation of Audio: The next modality we utilize from the video is the audio content extracted from it. We employ an Automatic Speech Recognition (ASR) module to extract transcripts from the audio. Since the datasets we worked with involved YouTube videos, we utilized the YouTube API to extract the closed caption transcripts associated with those videos.

4. Prompting: We employ the aforementioned modules to extract textual representations of various modalities present in a video. This ensures that we capture the audio, visual, text, and outside knowledge aspects of the video. Once the raw

text is collected and processed, we utilize it to prompt a generative language model in order to generate a coherent story that represents the video. To optimize the prompting process and enable the generation of more detailed stories, we remove similar frame captions and optical character recognition (OCR) outputs, thereby reducing the overall prompt size.

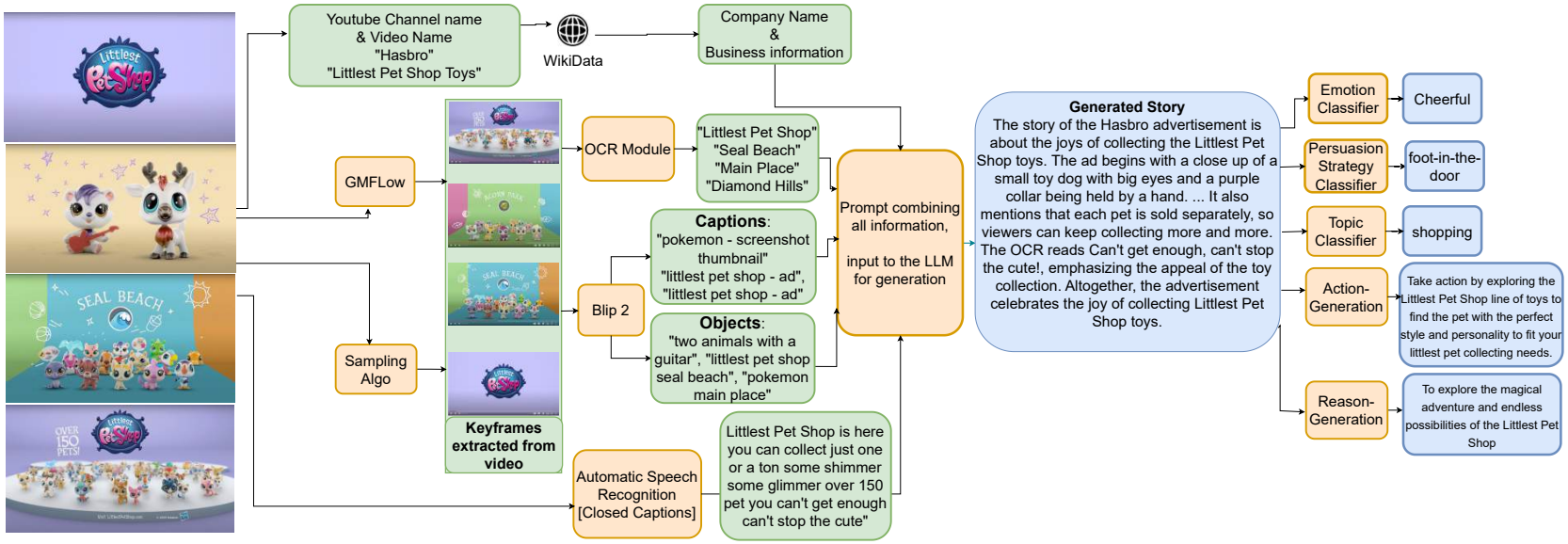


Figure 2.14: The overview of our framework to generate a story from a video and perform downstream video-understanding tasks. First, we sample keyframes from the video which are verbalized using BLIP-2. We also extract OCR from all the frames. Next, using the channel name and ID, we query Wikidata to get company and product information. Next, we obtain automatically generated captions from Youtube videos using the Youtube API. All of these are concatenated as a single prompt and given as input to an LLM and ask it to generate the story of the advertisement. Using the generated story, we then perform the downstream tasks of emotion and topic classification and persuasion strategy identification. This video can be watched at https://youtu.be/ZBLkTALi1CI.

The prompt template is given in Section 2.1.4.2.2. Through experimentation, we discovered that using concise, succinct instructions and appending the text input signals (such as frame captions, OCR, and automatic speech recognition) at the end significantly enhances the quality of video story generation. For shorter videos (up to 120 seconds), we utilize all available information to prompt the LLM for story generation. However, for longer videos, we limit the prompts to closed captions and sampled frame captions. The entire prompting pipeline is zero-shot and relies on pre-trained LLMs. In our story generation experiments, we employ GPT-3.5 (Brown *et al.*, 2020), Flan-t5 (Chung *et al.*, 2022), and Vicuna (Chiang *et al.*, 2023). A temperature of 0.75 is used for LLM generation. The average length of the generated stories is 231.67 words. Subsequently, these generated stories are utilized for performing video understanding tasks.

2.1.4.2.2 Prompt format For verbalization, a template prompt format has been used, including all the data components as objects, captions, asr, ocr, metadata.

"Please write a coherent story based on the following video advertisement. Use only the information provided and make sure the story feels like a continuous narrative and at the end include one sentence about what product the advertisement was about. Do not include any details not mentioned in the prompt.Use the elements given below to create a coherent narrative,but don't use them as it is.The advertisement for the company {company_name} The video is titled {title}, with captions that include {caption}, voice-over : {transcripts}, and object recognition descriptions : {ocr}. The following objects are present in the advertisement and should be used to help create the story: {objects} Please exclude any empty or stop words from the final text."

For downstream tasks, a template prompt format with an instruction about the specific task,the previous generated verbalization and vocabulary for the downstream task is prompted to the LLM. Here is the example for the topic detection task, for other tasks context and vocab were changed accordingly.

"Given {topics} identify the most relevant topic from the dictionary keys from topic_vocab related to the story of the video advertisement given below.Consider

the definitions given with topics in the topic_vocab dictionary, to identify which topic is most relevant, don't add any extra topics that are not given in dictionary keys and answer with just the most relevant topic. Story : {verbalization}"

Training	Model	Topic	Emotion		Persuasion	Action	Reason
			All labels	Clubbed			
Random	Random	2.63	3.37	14.3	8.37	3.34	3.34
Finetuned	VideoMAE (Tong *et al.*, 2022)	24.72	29.72	85.55	11.17	-	-
	Hussain *et al.* (Hussain *et al.*, 2017)	35.1	32.8	-	-	-	48.45
	Intern-Video (Wang *et al.*, 2022)	57.47	36.08	86.59	5.47	6.8	7.1
Zero-shot	VideoChat (Li *et al.*, 2023*c*)	9.07	3.09	5.1	10.28	-	-
Our Framework	GPT-3.5 Generated Story + GPT-3.5 Classifier	51.6	11.68	79.69	35.02	66.27	59.59
Zero-shot	GPT-3.5 Generated Story + Flan-t5-xxl Classifier	60.5	10.8	79.10	33.41	79.22	81.72
	GPT-3.5 Generated Story + Vicuna Classifier	22.92	10.8	67.35	29.6	21.39	20.89
	Vicuna Generated Story + GPT-3.5 Classifier	46.7	5.9	80.33	27.54	61.88	55.44
	Vicuna Generated Story + Flan-t5-xxl Classifier	57.38	9.8	76.60	30.11	77.38	80.66
	Vicuna Generated Story + Vicuna Classifier	11.75	10.5	68.13	26.59	20.72	21.00
Finetuned	Generated Story + Roberta Classifier	71.3	33.02	84.20	64.67	42.96	39.09

Table 2.6: Comparison of all the models across topic, emotion, and persuasion strategy detection tasks. We see that our framework, despite being zero-shot, outperforms finetuned video-based models on the topic classification, persuasion strategy detection and action and reason classification tasks and comes close on the emotion classification task. Further, the Roberta classifier trained on generated stories outperforms both finetuned and zero-shot models on most tasks. Best models are denoted in green and runner-ups in blue.

	Method	Model Type	METEOR	CIDEr	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Random	Random	Retrieval	13.1	30.2	21.4	43.1	23.1	10.0	4.8
Finetuned	Narrator (Li *et al.*, 2020*b*)	Retrieval	19.6	98.4	29.5	69.1	43.0	25.3	15.0
	EMB (Li *et al.*, 2020*b*)	Retrieval	19.1	88.8	28.9	64.5	39.3	22.7	13.4
	BRNN (Li *et al.*, 2020*b*)	Retrieval	18.1	81.0	28.3	61.4	36.6	20.3	11.3
	ResBRNN (Li *et al.*, 2020*b*)	Retrieval	19.6	94.3	29.7	66.0	41.7	24.3	14.7
	Pseudo-GT+ ResBRNN-kNN (Li *et al.*, 2020*b*)	Retrieval	20.1	103.6	29.9	69.1	43.5	26.1	15.6
	GVMF (Lu and Wu, 2022)	Retrieval	20.7	107.7	30.8	70.5	44.3	26.9	15.9
Zero-shot	VideoChat (Li *et al.*, 2023*c*)	Generative	15.49	42.9	17.88	50.00	43.30	34.76	27.21
Zero-shot Our Framework	GPT-3.5	Generative	24.8	102.4	24.3	63.8	56.4	47.2	38.6
	Vicuna	Generative	17.4	73.9	20.9	70.49	60.0	48.25	38.20
	Flant-t5-xxl	Generative	4.8	34.6	10.58	7.9	6.8	5.4	4.3
	Uniformly Sampled BLIP-2 Captions	Generative	21.7	108.9	24.04	55.19	48.5	40.7	33.76

Table 2.7: Comparison on story generation task on the video-story dataset. We see that our framework despite being zero-shot outperforms all the fine-tuned generative prior art on all metrics. Further, it also outperforms fine-tuned retrieval models, which choose from a fixed set of frame descriptions on most metrics. Best models are denoted in green and runner-ups in blue.

2.1.4.2.3 Results Video Storytelling: The performance comparison between our pipeline and existing methods is presented in Table 2.7. We evalu-

Task	Model	METEOR	CIDEr	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Action	GPT-3.5	20.46	41.7	9.5	18.7	14.8	11.8	9.4
Action	Flan-t5-xxl	15.75	61.5	13.6	50.0	34.8	26.9	21.8
Action	Vicuna	21.20	42.6	7.6	16.8	13.08	10.08	7.7
Reason	GPT-3.5	13.34	16.7	7.8	27.1	20.8	14.7	10.4
Reason	Flan-t5-xxl	8.35	24.9	5.9	39.4	24.7	16.7	12.0
Reason	Vicuna	15.82	27.9	7.75	24.6	19.3	14.1	10.3
Reason given action	GPT-3.5	13.77	29.4	8.7	33.5	24.9	17.9	13.2
Reason given action	Flan-t5-xxl	4.29	19.0	7.6	23.2	15.0	10.2	7.5
Reason given action	Vicuna	13.62	24.4	7.61	22.6	17.7	12.8	9.2

Table 2.8: Comparison of the different zero-shot models on the action and reason generation tasks. Note that there are no fine-tuned generative models in the literature for this task and the number of annotated videos is too small to train a generative model. Best models are denoted in green.

Training	Model	relationship	way_speaking	scene	like_ratio	view_count	director	genre	writer	year
Trained	R101-slowfast+NL (Wu and Krahenbuhl, 2021)	52.4	35.8	54.7	0.386	3.77	44.9	53.0	36.3	52.5
Trained	VideoBert (Sun *et al.*, 2019*a*)	52.8	37.9	54.9	0.320	4.46	47.3	51.9	38.5	36.1
Trained	Xiao *et al.* (Xiao *et al.*, 2022)	50.95	34.07	44.19	0.353	4.886	40.19	48.11	31.43	29.65
Trained	Qian *et al.* (Qian *et al.*, 2021)	50.95	32.86	32.56	0.444	4.600	37.76	48.17	27.26	25.31
Trained	Object Transformers (Wu and Krahenbuhl, 2021)	53.1	39.4	56.9	0.230	3.55	51.2	54.6	34.5	39.1
Zero-shot (Ours)	GPT-3.5 generated story + Flan-t5-xxl	64.1	39.07	60.2	0.061	12.84	69.9	58.1	52.4	75.6
Zero-shot (Ours)	GPT-3.5 generated story + GPT-3.5 classifier	68.42	32.95	54.54	0.031	12.69	75.26	50.84	32.16	75.96
Trained (Ours)	GPT-3.5 generated story + Roberta	62.16	38.41	68.65	0.054	11.84	45.34	39.27	35.93	7.826

Table 2.9: Comparison of various models on the LVU benchmark. We see that our framework, despite being zero-shot, outperforms fine-tuned video-based models on 8/9 tasks. Best models are denoted in green and runner-ups in blue.

Training	Model	Scene	Object	Action	Event	Attribute	Concept	Overall
Trained	3D-Resnet	50.6	28.6	48.2	35.9	29	22.5	35.8
Trained	3D-STCNet	51.9	30.1	50.3	35.8	29.9	22.7	36.7
Trained	HATNet	55.8	34.2	51.8	38.5	33.6	26.1	40
Trained	3D-Resnet (Multitask)	51.7	29.6	48.9	36.6	31.1	24.1	37
Trained	HATNet (Multitask)	57.2	35.1	53.5	39.8	34.9	27.3	41.3
Zero-shot (Ours)	GPT-3.5 generated story + Flan-t5-xxl classifier	59.66	98.89	98.96	38.42	67.76	86.99	75.12
Zero-shot (Ours)	GPT-3.5 generated story + GPT-3.5 classifier	60.2	99.16	98.72	40.79	67.17	88.6	75.77

Table 2.10: Comparison of various models on the HVU benchmark (Diba *et al.*, 2020). The models scores are as reported in (Diba *et al.*, 2020). We see that our framework, despite being zero-shot, outperforms fine-tuned video-based models on all the tasks. Best models are denoted in green and runner-ups in blue.

ate multiple generative and retrieval-based approaches and find that our pipeline achieves state-of-the-art results. It is important to note that as our method is entirely generative, the ROUGE-L score is lower compared to retrieval-based methods due to less overlap with ground truth reference video stories. However, overall metrics indicate that our generated stories exhibit a higher level of similarity to the reference stories and effectively capture the meaning of the source video.

Video Understanding: The performance comparison between our pipeline and other existing methods across six tasks (topic, emotion, and persuasion strategy classification, as well as action and reason retrieval and generation) is presented in Tables 2.6 and 2.8. Notably, our zero-shot model outperforms finetuned video-based baselines in all tasks except emotion classification. Further, our text-based finetuned model outperforms all other baselines on most of the tasks.

Unlike the story generation task, there are limited baselines available for video understanding tasks. Moreover, insufficient samples hinder training models from scratch. To address this, we utilize state-of-the-art video understanding models, VideoMAE and InternVideo. InternVideo shows strong performance on many downstream tasks. Analyzing the results, we observe that while GPT-3.5 and Vicuna perform similarly for story generation (Table 2.7), GPT-3.5 and Flan-t5 excel in downstream tasks (Table 2.6). Interestingly, although GPT-3.5 and Vicuna-generated stories yield comparable results, GPT-3.5 exhibits higher performance across most tasks. Vicuna-generated stories closely follow GPT-3.5 in terms of downstream task performance.

Next, we compare the best models (as in Table 2.6) on the LVU and HVU benchmarks with respect to the state-of-the-art models reported in the literature. Tables 2.9 and 2.10 report the results for the comparisons. As can be noted, the zero-shot models outperform most other baselines. For LVU, the zero-shot models work better than the trained Roberta-based classifier model. For HVU, we convert the classification task to a retrieval task, where in a zero-shot way, we input the verbalization of a video along with 30 randomly chosen tags containing an equal number of tags for each category (scene, object, action, event, attribute, and concept). The model is then prompted to pick the top 5 tags that seem most relevant to the video. These tags are mapped back to the main category tags, which are treated as the predicted labels.

Furthermore, as a comparative and ablation study of our approach, we evaluate the performance using only the BLIP-2 captions and audio transcriptions (Table 2.11). Our findings highlight that generated stories leveraging both audio and visual signals outperform those using vision or audio inputs alone. This emphasizes the significance of verbalizing a video in enhancing video understanding.

Model	Topic	Emotion		Persuasion	Action	Reason
		All labels	Clubbed			
BLIP-2 Captions + Flant-t5-xxl	32.2	7.4	43.11	32.1	52.98	76.26
BLIP-2 Captions + GPT-3.5	32.7	7.9	76.69	30.1	49.91	58.71
Audio Transcription + Flant-t5-xxl	49.37	10.1	63.56	21.9	66.17	79.68
Audio Transcription + GPT-3.5	32.88	6.4	75.97	32.25	64.98	61.78

Table 2.11: Ablation study of using only visual (caption) or audio (transcripts) and LLMs for downstream tasks. It can be noted that the overall model does not perform as well (compared to Table 2.6) when using only audio or scene description without generating story.

2.1.4.2.4 Ablation Among the different components of information input present in the prompt, the LLM utilizes them differently while constructing the verbalization for the videos.For this experiement we use a subset of (Hussain *et al.*, 2017) dataset, considering videos that have spoken audio present.

We use ROUGE-l to get the longest common subsequence (LCS) between the generated verbalization and the individual components,which captures the overlapping content, providing an indication of their semantic similarity.

As generated verbalizations are abstractive as compared to extractive, we also use cosine similarity between the Roberta embeddings of the generated verbalization and the individual components.

We find that despite the order of the components in the prompt, the LLMs tend to utilize the audio components in the videos, in an extractive way.

Model	Top-5 Accuracy	mAP
VideoMAE	25.57	24.79
InternVideo	7.477	15.62
GPT-3.5 Generated Story + GPT-3.5	34.2	27.53
Vicuna Generated Story + GPT-3.5	31.54	27.24
GPT-3.5 Generated Story + Flant5	37	27.96
Vicuna Generated Story + Flant5	31.13	27.32

Table 2.12: Top-5 accuracy, and mAP for persuasion strategy detection task

2.1.4.2.5 A few examples of the stories generated using our method 1. "Cathy 30 Master" by Grand Canyon University (`https://www.youtube.com/watch?v=lPdD8NvVfw0`): Kathy Ames had always wanted to pursue a doctoral degree but was unsure about the time commitment. When she discovered Grand

Canyon University, she knew she had found the perfect fit. Grand Canyon University offered a flexible schedule that would allow her to balance her personal and family life with her studies. She - along with other students - gathered in the classroom, excitedly listening to their coach, Scott Saunders, explain the program. Afterward, Kathy made her way to the library and settled into a chair with her laptop.

She studied diligently, surrounded by her peers and classmates. In the evenings, she met with her peers around the table to discuss the topics of the day. Everyone was always eager to help and support each other. After a long day, Kathy made her way back to her living room where she relaxed on the couch with a glass of water and a lamp providing a soothing light.

Kathy was grateful for the opportunity to pursue her dream at Grand Canyon University. She was able to learn from experienced faculty and gain real-world experience that would prepare her for success after graduation.

The advertisement for Grand Canyon University was about offering a private, Christian education at an affordable price.

2. "Today is the Day for Doing | The Home Depot" by Home Depot (`https://www.youtube.com/watch?v=f_6QQ6IVa6E`): The woman holding the book stepped onto the patio and looked up to the sky. She was ready to take on the day. Taking out her phone, she opened the furniture catalog app, scrolling through the various designs. She quickly decided on the perfect pieces to brighten up her home. Next, she headed to The Home Depot for the supplies she needed. As she entered the store, the woman was delighted to find all the tools and materials she needed, from the Ryobi Cordless Vacuum to the Leaf Blower. She was even more excited when she spotted the Splatter an object recognition tool that allowed her to easily find the perfect paint color for her project. With her shopping done, the woman made her way to the checkout line with a cup of coffee in hand. She couldn't wait to get to work and make her home more beautiful. She knew that with the help of The Home Depot, Today was the Day for Doing. This advertisement was for The Home Depot - the one-stop-shop for all your home improvement needs.

3. “October 2018 Color of the Month: Shell White - Sherwin-Williams” by Sherwin-Williams (https://www.youtube.com/watch?v=PJlHiQJBDMw): The advertisement for the company Sherwin-Williams opens on a kitchen table strewn with shells and wicker baskets, with two glasses of iced water beside them. A vase with a blue pattern sits in the foreground, and a person holds up a phone with the Sherwin-Williams logo on the screen. A girl appears from behind a white sheet, peeking out of a white tent as if to signify the timelessness and neutrality of this color. The voice-over begins, as the camera pans to a living room with a staircase, and then to a dining room with a white table, chairs, and a white vase. The words “Color of the Month: Shell White, Sherwin-Williams” appear on the screen, as the camera zooms in on the vase. The words are followed by Öur app makes it a snap,ŕeferring to Color Snap, the company’s new way of painting a home. The advertisement ends with the Sherwin-Williams logo, emphasizing the company’s commitment to excellence in home painting. This advertisement was promoting the company’s color of the month, Shell White.

4. “Side-by-side comparison of SolidWorks w/NVIDIA Quadro FX 1700 vs. Quadro 2000” by NVIDIA (https://www.youtube.com/watch?v=CDjBIt70fp4): The story began with a green light glowing in the dark, symbolizing the presence of a powerful technology that can change the way we work. This technology was an advanced graphics card, the NVIDIA Quadro FX 1700. It was compared side-by-side with its successor, the Quadro 2000, and it was clear that the Quadro 2000 was far more powerful. The Quadro FX 1700 had a qt - x700 vs qt - x700 capacity, while the Quadro 2000 had a green light that shone brighter and further than before. As the comparison was being made, the results were clear: the Quadro 2000 was the superior product. This advertisement for the company NVIDIA showcased the power of the Quadro 2000, and the improved performance it could bring to an organization. The advertisement concluded with a green light, signaling that NVIDIA had the answer to improving workflows. The product the advertisement was about was the NVIDIA Quadro 2000.”

5. “Training sessions in the nature - Mary Kom | Priyanka Chopra | In Cinemas NOW” by Viacom 18 Studios (https://www.youtube.com/watch?v=SXTSJC5N

rLs): Mary Kom was training hard to make her mark as a boxer, and she had nothing but her focus and determination to help her succeed. She was a force to be reckoned with in the boxing ring and her hard work was soon paying off.But she knew that she had to take her training to the next level if she wanted to win the boxing championship. That's when she decided to take her own advice and step out of her comfort zone. She headed out to the woods, determined to make the most of her time in nature.The woods were a perfect place for Mary to challenge herself physically. She was soon standing on the rocks of a river, pushing her body to its limits with every movement. She was standing tall, her black top glinting in the sunlight, and her energy was palpable. The sun was setting, and it was the perfect time to relax and unwind.Finally, she ended her day with a yoga routine, her body twisting and turning in a beautiful display of strength and grace. It was a day of hard work and fun, and it left Mary feeling inspired and ready to take on anything. The video was an advertisement for the movie Mary Kom, released by Viacom18 Studios."

2.1.4.2.6 Hallucinations Present In the Automatically Generated Stories The current large language models , do hallucinate or produce imaginary outputs. However, they are able to capture the gist of the videos, helping improve on downstream tasks. Some examples are shown below:

1. "The Fundamentals Of Caring | Clip: "Perverts" | Netflix" by Netflix (https://youtu.be/_Usyxm9S_Jg)
   (a) Story: The woman in the advertisement is standing in front of a red background, smiling brightly. She and a man sitting at a table look content, as if they've just come from a conversation. The Netflix logo is visible, indicating that the advertisement is for the company and its products. The voice-over for the advertisement introduces Selena Gomez as the star of the trailer for the popular Twilight Saga. Her popularity has made her a perfect choice to promote Netflix's products. The voice-over then informs viewers that the company's streaming service is not available. The man and woman in the advertisement seem to have a strong connection, and we can infer that they are friends. Perhaps they share an understanding of each other's struggles, since they both seem to be comfortable and happy with each other. The advertisement ends with the Netflix logo, leaving viewers with the knowledge that they can find solace and understanding in the company's products, no matter what they are going through. The advertisement was for Netflix."

(b) Hallucinations: The story falsely attributes “Selena Gomez” to “Twilight Saga”. The LLMs often confuse named entity linkages.

2. “Back-to-School 2016” by Academy Sports + Outdoors (https://youtu.be/SZwWT5lnHf8)
   (a) Story: A young girl is walking through her bedroom when she notices a man running up the stairs in his sneakers. She is intrigued and follows him, curious to see where he is going. Arriving at the top of the stairs, she finds a young man standing with his backpack. Across the room, she sees a girl in a dress standing in front of a mirror. The girl smiles and looks content. At that moment, the voice-over of the Academy Sports + Outdoors ad can be heard: ”Academy Sports and Outdoors is the place to get all the best brands in thousands of combinations for back to school. Right stuff, low price, every day.” Suddenly, a black cat can be seen flying in the sky and the girl in the tank top is now standing in front of the mirror. The man is wearing blue pants and sneakers, and the shoes, rug, and sneakers of the apartment are visible. The girl smiles and looks content. The scene shifts to the bedroom where the bed, dresser, and mirror are present, along with a sign for Academy Sportswear. This advertisement for Academy Sports + Outdoors was about the company’s back-to-school collection of products. It showcased how Academy Sports + Outdoors provides the right stuff, at a low price, every day, for students to get ready for school.
   (b) The scene with a black cat is hallucinated though not mentioned in the prompt

3. “Many Worlds, One Home l Anime 2019 l NX” By Netflix (https://youtu.be/B63ubKGNqos)
   (a) Story: The advertisement by Netflix, titled ”Many Worlds, One Home”, is a stunning vision of many different universes. We see a man standing on a red moon, a desert landscape, and a live performance by Adam Sandler. Purple backgrounds with the word ”worlds” convey the idea of many possibilities, while a boy playing with a ball in front of a statue shows a more peaceful and homely side to this. As the night falls, a man is standing on a car, and the Ultraman. There are also glimpses of Tse Seeds, Devilman Crybaby, Saint Seiya: Knights of the Zodiac, Eden Vs. Eden, Carole and Tuesday, Aggretsuko, Seis Manos, Godzilla, Castlevania, B: The Beginning, Baki, Cannon Busters, Rilakkuma and Kaoru, Beastars, and Ultramarine Magmell. We see a teddy bear, knights of the zodiac, a character walking in the water, a man in a snowy area, a pair of shoes, a man running away from the fire, the moon, a man wearing a suit with a green light, eyes being the only thing that can be seen, birds flying in the sky, a red and black background with a red light, bats and castles, the eyes, a man holding a gun, a man flying in the air, a face, a girl with long hair, and two hands. All of this culminates to signify the many different worlds that one can find ”home”, and the advertisement concludes with the Netflix logo. This advertisement for Netflix captures the idea of many possibilities and of being at home in a single place.

(b) Hallucination: Blip2 hallucinates entities such as "Adam Sandler" who is not associated with the advertisement. But it causes the generated story to have the same association.

4. "African Cats: Herding Cats - Clip" by Walt Disney Studios (https://youtu.be/WCIEmrN8ot4)
   (a) Story: The sun shone brightly in the savanna as the cheetah cub frolicked in the grass. Suddenly, a cheetah emerged from the shadows. The cheetah and it's cub plan to chase the gazzeles. Despite the presence of death, life in the wild continued on, just as it had for centuries. This story is featured in the advertisement for Walt Disney Studios, which captures the beauty and unpredictable nature of life in the African savanna. This advertisement ultimately serves to promote the film African Cats, which tells the story of a cheetah family as they strive to protect their respective prides against a variety of obstacles in the wild.
   (b) Hallucination: The generated caption "a fox cub is playing in the grass cheetah eats in the wild" was insufficient to capture the overall information of the video, however, due to the presence of ASR and world knowledge of the LLM. The LLM is able to generate a coherent story.

5. "Bright & Dreamy: Pastel Tint| Ulta Beauty" by ultabeauty (https://youtu.be/Lh5n4HcLvCU)
   (a) Story: The woman in the Ulta Beauty advertisement skillfully applied makeup to her face, creating a beautiful pastel tint. With each brush stroke, her look became brighter and more dreamy. She beamed with confidence and joy, her face glowing with beauty. As the advertisement continued, she highlighted her eyes and cheeks with subtle pastel shades, adding to the stunning effect. With each addition, her face lit up with a vibrant luminosity. When she was finished, her face was a work of art, the perfect combination of bright and dreamy. The Ulta Beauty advertisement showcased the power of a perfect pastel tint, and how it can enhance any look. This advertisement was for Ulta Beauty.
   (b) Hallucination: The video contains only music. However, the generated verbalization can pick the advertisement's gist.

6. "Round Brilliant Diamond Hoop Earrings 14kt White Gold" by Costco https://youtu.be/VdhRmVnbqQo
   (a) Story: Jackie had been wanting diamond hoop earrings for what seemed like forever. She had been dreaming of the perfect pair that she could wear on special occasions and that would bring out the best in her. One day Jackie happened to come across an advertisement for Costco, which featured a pair of Round Brilliant Diamond Hoop Earrings in 14kt White Gold. She immediately knew this was the perfect pair for her. The advertisement had a voice-over that said, "These earrings sparkle with glamour and sophistication." Jackie was mesmerized by the sparkle of the diamonds, and she was sure anyone who saw them would also be taken in by their beauty. Jackie quickly decided to purchase the earrings. She was sure they would make the perfect accessory for any

special occasion. From date nights to family gatherings, she knew these earrings would make her look simply stunning. The advertisement for the Costco Round Brilliant Diamond Hoop Earrings in 14kt White Gold proved to be the perfect choice for Jackie. They not only added a touch of glamour and sophistication to her wardrobe, but made her feel extra special when she wore them. The advertisement for the company Costco was for a pair of Round Brilliant Diamond Hoop Earrings in 14kt White Gold.

(b) Hallucination: The video has no audio transcription, and minimal visual information. However the LLM hallucinates a story, maintaining the essence of the advertisement. It also attributes a false voice-over which is not present in the video.

2.2 MINIMAL: Mining models for universal adversarial triggers

2.2.1 Abstract

It is well known that natural language models are vulnerable to adversarial attacks, which are mostly input-specific in nature. Recently, it has been shown that there also exist input-agnostic attacks in NLP models, special text sequences called universal adversarial triggers. However, existing methods to craft universal triggers are data intensive. They require large amounts of data samples to generate adversarial triggers, which are typically inaccessible by attackers. For instance, previous works take 3000 data samples per class for the SNLI dataset to generate adversarial triggers. In this paper, we present a novel data-free approach, *MINIMAL*, to mine input-agnostic adversarial triggers from models. Using the triggers produced with our data-free algorithm, we reduce the accuracy of Stanford Sentiment Treebank's positive class from 93.6% to 9.6%. Similarly, for the Stanford Natural Language Inference (SNLI), our single-word trigger reduces the accuracy of the entailment class from 90.95% to less than 0.6%. Despite being completely data-free, we get equivalent accuracy drops as data-dependent methods.

2.2.2 Introduction

In the past two decades, deep learning models have shown impressive performance over many natural language tasks, including sentiment analysis (Zhang *et al.*, 2018*b*), natural language inference (Parikh *et al.*, 2016), automatic essay scoring (Kumar *et al.*, 2019), question-answering (Xiong *et al.*, 2017), keyphrase extraction (Meng *et al.*, 2017), *etc.* At the same time, it has also been shown that these models are highly vulnerable to adversarial perturbations (Behjati *et al.*, 2019). The adversaries change the inputs to cause the models to make errors. Adversarial examples pose a significant challenge to the rising deployment of deep learning based systems[¶].

Commonly, adversarial examples are found on a per-sample basis, *i.e.*, a separate optimization needs to be performed for each sample to generate an adversarially perturbed sample. Since the optimization needs to be performed for each sample, it is computationally expensive and requires deep learning expertise for generation and testing. Lately, several research studies have shown the existence of input-agnostic universal adversarial trigger (UATs) (Moosavi-Dezfooli *et al.*, 2017; Wallace *et al.*, 2019). These are a sequence of tokens, which, when added to any example, cause a targeted change in the prediction of a neural network. The existence of such word sequences poses a considerable security challenge since the word sequences can be easily distributed and can cause a model to predict incorrectly for all of its inputs. Moreover, unlike input-dependent adversarial examples, no model access is required at the run time for generating UATs. At the same time, the analysis of universal adversaries is interesting from the point of view of model, dataset analysis and interpretability (§2.2.6). They tell us about the global model behaviour and the general input-output patterns learnt by a model (Wallace *et al.*, 2019).

Existing approaches to generate UATs assume that an attacker can obtain the training data on which a targeted model is trained (Wallace *et al.*, 2019; Behjati *et al.*, 2019). While generating an adversarial trigger, an attacker firstly *trains* a proxy model on the training data and then generates adversarial examples by

[¶]The code and reproducibility steps are given in `https://github.com/midas-research/data-free-uats`

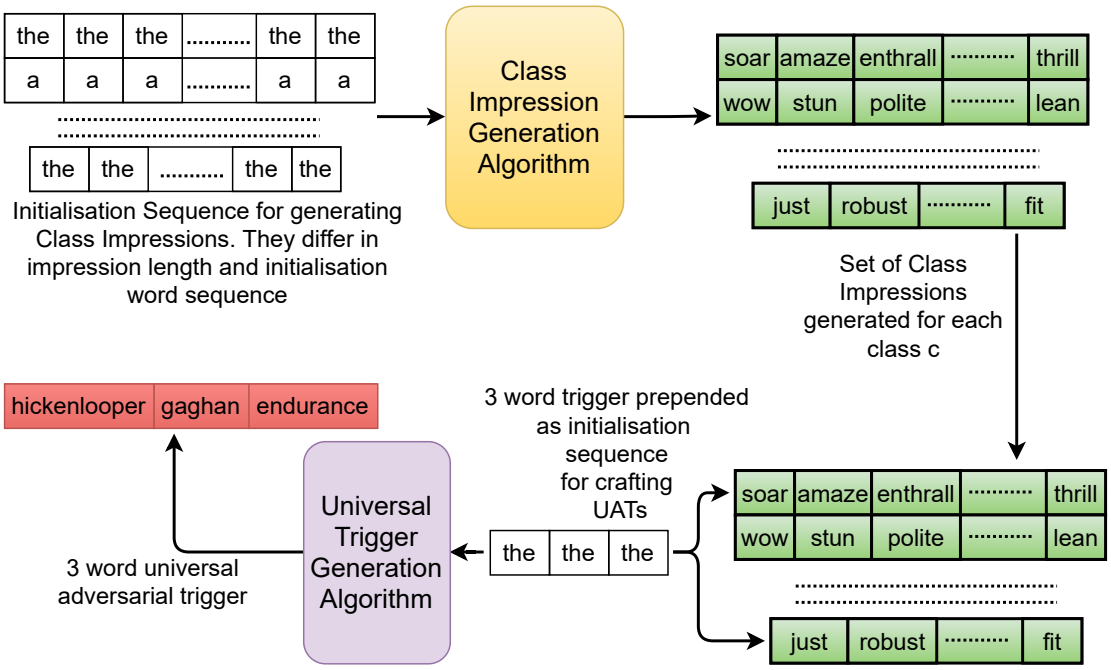


Figure 2.15: Two step process to generate universal adversarial triggers. First, we generate multiple class impressions for each class c. For this, we take multiple initialization sequences differing in starting word and length. After generating class impressions, we use them as our dataset for generating universal adversarial triggers.

using gradient information. Table 2.13 presents the data requirements during training for the current approaches. For instance, to find universal adversaries on the natural language inference task, one needs 9000 training examples. Also, the adversarial ability of a perturbation has been shown to depend on the amount of data available (Mopuri *et al.*, 2017, 2018*a*). However, in practice, an attacker rarely has access to the training data. Training data are usually private and hidden inside a company's data storage facility, while only the trained model is publicly accessible. For instance, Google Cloud Natural Language (GCNL) API only outputs the scores for the sentiment classes (Google, 2021) while the data on which the GCNL model was trained is kept private. In this real-world setting, most of the adversarial attacks fail.

In this paper, we present a novel data-free approach for crafting universal adversarial triggers to address the above issues. Our method is to mine a trained *model* (but not data) for perturbations that can fool the target model without any knowledge about the data distribution (*e.g.*, type of data, length and vocabulary of samples, *etc.*). We only need access to the embedding layer and model outputs. Our method achieves this by solving first-order Taylor approximation of two tasks:

first, we generate "class-impressions" (§2.2.4.1), which are reconstructed text sentences from a model's memory representing the learned parameters for a certain data class; and second, we mine universal adversarial triggers over these generated class impressions (§2.2.4.2). Class-impression can be considered as the general representation of samples belonging to a particular class (Fig 2.19) and are used to emulate samples belonging to that class in our method. The concept of data leaving its impression on a trained model has also been observed in prior work in model inversion attacks in computer vision (Micaelli and Storkey, 2019; Nayak *et al.*, 2019). We build on that concept to mine universal adversarial triggers. We propose a combination of general model inversion attacks methodology with trigger generation to mine data-free adversarial triggers and show our results for several NLP models (Fredrikson *et al.*, 2015; Tramèr *et al.*, 2016).

The major contributions of our work are summarized as:

- For the first time in the literature, we propose a novel data-free approach, MINIMAL (MINIng Models for AdversariaL triggers), to craft universal adversarial triggers for natural language processing models and achieve state-of-the-art success (adversarial) rates (§2.2.5). We show the efficacy of the triggers generated using our method on three well-known datasets and tasks, *viz.*, sentiment analysis (§2.2.5.1) on Stanford Sentiment Treebank (SST) (Socher *et al.*, 2013), natural language inference (§2.2.5.2) on the SNLI dataset (Bowman *et al.*, 2015), and paraphrase detection (§2.2.5.3) on the MRPC dataset (Dolan and Brockett, 2005).

- We use both class impressions and universal adversarial triggers generated by our models to try to understand the models' global behaviour (§2.2.6). We observe that the words with the lowest entropy (*i.e.*, the most informative features) appear in the class impressions (Fig. 2.18). We find that these low entropy word-level features can also act as universal adversarial triggers (Table 2.24). The class-impression words are good representations of a class since they form distinct clusters in the manifold representations of each class.

2.2.3 Related Work

Universal Adversarial Attacks: Moosavi-Dezfooli *et al.* (2017) showed the existence of universal adversarial perturbations. They showed that a *single perturbation* could fool DNNs most of the times when added to all images. Since then, many universal adversarial attacks have been designed for vision systems (Khrulkov and Oseledets, 2018; Li *et al.*, 2019; Zhang *et al.*, 2021). To the best of our knowledge, there are only three recent papers for NLP based universal adversarial attacks, and all of them require data for generating universal adversarial triggers (Wallace *et al.*, 2019; Song *et al.*, 2021; Behjati *et al.*, 2019). In simultaneous works, (Wallace *et al.*, 2019; Behjati *et al.*, 2019) show universal adversarial triggers for NLP. Song *et al.* (2021) extend it to generate natural (data-distribution like) triggers. We compare our work with (Wallace *et al.*, 2019) since they show improved adversarial success rates over (Behjati *et al.*, 2019). We leave mining natural triggers from models as a future study. Our results demonstrate comparable performance as (Wallace *et al.*, 2019) but without using any data. Table 2.13 mentions the data requirement of (Wallace *et al.*, 2019).

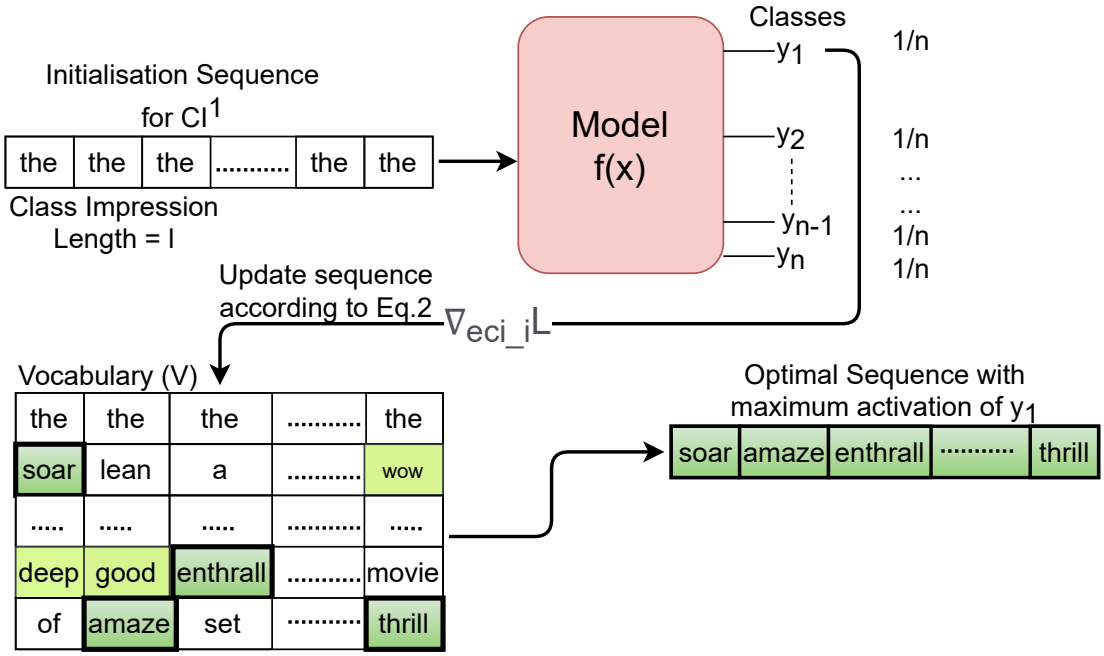


Figure 2.16: Class Impression Generation (CIG) Algorithm. We start with an initial sequence of *"the the ... the"* and continuously update it based on its gradient with respect to output probabilities (Eq. 2.4). The final sequence we get represents the class impression CI^c for the class c.

While there are many proposed classifications of adversarial attacks, from the point of view of our work, they can be seen in two ways: (a) data-based attacks; (b) data-free attacks. Data-based computer vision attacks depend on training and validation dataset to craft adversaries, while data-free attacks rely on other signals. There are some data-free approaches in computer vision, for example, by maximizing activations at each layer (Mopuri *et al.*, 2017, 2018*a*), class activations (Mopuri *et al.*, 2018*b*), and pretrained models and proxy dataset (Huan *et al.*, 2020). However, there has been no work in NLP systems for data-free attacks.

2.2.4 The Proposed Approach

In summary, our algorithm of crafting data-free universal adversarial triggers is divided into two steps, as shown in Fig 2.15. First, we generate a set of class-impressions (§2.2.4.1) (Fig 2.16) for each class. These natural language examples represent the entire class of samples and are generated solely from the weights learnt by the model. Second, we use the set of class impressions generated in the first step to craft universal adversarial triggers corresponding to those impressions (§2.2.4.2) (Fig 2.17).

2.2.4.1 Class-Impressions Generation (CIG) Algorithm

To generate the class impression CI^c for a class c, we propose to maximize the confidence of the model $f(x)$ for an input text sequence $\mathbf{t}_c$. Formally, we maximize:

$$CI^c = \arg\max_{\mathbf{t}_c} \mathbb{E}_{\mathbf{t}_c \sim \mathcal{V}} \left[\mathcal{L}(c, f(\mathbf{t}_c)) \right], \tag{2.3}$$

where $\mathbf{t}_c$ is sampled from a vocabulary $\mathcal{V}$. The input $\mathbf{t}_c$ in NLP is not continuous, but is made up of discrete tokens. Therefore, we use the first-order Taylor approximation of Eq. 2.3 (Michel *et al.*, 2019; Ebrahimi *et al.*, 2018; Wallace *et al.*, 2019). Formally, for every token $\mathbf{e}_{ci_i}$ in a class impression CI^c, we solve the following equation:

$$\mathbf{e}_{ci_i} = \arg\min_{\mathbf{e}'_i \in \mathcal{V}} [\mathbf{e}'_i - \mathbf{e}_{ci_i}]^\intercal \nabla_{\mathbf{e}_{ci_i}} \mathcal{L}, \tag{2.4}$$

Dataset	Validation Size (Real samples)	Impressions Size (Generated samples)
SST	900	300
SNLI	9000	400
MRPC	800	300

Table 2.13: Number of Samples required to generate Universal Adversarial Triggers for each Dataset. In a data-based approach like (Wallace *et al.*, 2019), validation set (column 2) is used to generate the UATs. The third column lists the number of queries we make to generate artificial samples. These artificial samples are then used to craft UATs. Note that no real samples are required for our method.

where $\mathcal{V}$ represents the set of all words in the vocabulary, and $\nabla_{\mathbf{e}_{ci_i}} \mathcal{L}$ is the gradient of the task loss. We model the Eq. 2.4 as an iterative procedure by starting out with an initialisation value of $\mathbf{e}_{ci_i}$ as *'the'*. We then continually optimize it until convergence. For computing the optimal $\mathbf{e}'_i$, we take $|\mathcal{V}|$ d-dimensional dot products where d is the dimensionality of the token embedding. We use beam-search for finding the optimal sequence of tokens $\mathbf{e}'_i$ to get the minimum loss in Eq. 2.4. We score each beam using the loss on the batch in each iteration of the optimization schedule.

Finally, we convert the optimal $\mathbf{e}_{ci_i}$ back to their associated word tokens. Fig. 2.16 presents an overview of the process. It shows the case where we initialized $\mathbf{e}_{ci_i}$ with a sequence of *"the the .... the"* and then follow the optimization procedure for finding the optimal CI^c for the class c[‖].

To generate class impressions for the models that use contextualized embeddings like BERT (Devlin *et al.*, 2019), we perform the above optimization over character and sub-word level. We also replace the context-independent embeddings in Eq. 2.4 with contextual embeddings as obtained from BERT after passing the complete sentence to it.

We generate multiple class impressions for each class for all models by varying the number of tokens and the starting sequence. This gives us a number of class impressions for the next step where we generate triggers over these class impressions.

[‖]We vary the initialization sequence and sequence length to generate multiple class impressions for the same class

2.2.4.2 The Universal Trigger Generation (UTG) Algorithm

Class	Class Impression
Positive	energizes energizes captivated energizes enthrall eye-catching captivating aptitude artistry passion
Positive	captures soul-stirring captivates mesmerizing soar amaze excite amaze enthrall thrill captivating impress artistry accomplishments
Negative	spiritless ill-constructed ill-conceived ill-fitting aborted fearing bottom-rung woe-is-me uncharismatically pileup
Negative	laziest third-rate insignificance stultifyingly untalented hat-in-hand rot leanest blame direct-to-video wounds urinates

Table 2.14: Class Impressions for BiLSTM-Word2Vec Sentiment Analysis Model. Note that the words in the class impression examples highly correspond to the respective sentiment classes.

After generating class impressions in the previous step, we generate adversarial triggers as follows. From the last algorithm, we get a batch of class impressions CI^c for the class c. The task of crafting universal adversarial triggers is defined as minimizing the following loss function:

$$\arg\min_{\mathbf{t}_{adv}} \mathbb{E}_{\mathbf{t}\sim\mathcal{C}I^c}\left[\mathcal{L}(\tilde{c}, f(\mathbf{t}_{adv};\mathbf{t}))\right], \tag{2.5}$$

where $\tilde{c}$ denotes target class (distinct from the class c), $f(\mathbf{t}_{adv};\mathbf{t})$ denotes the evaluation of $f(x)$ on the input containing concatenation of adversarial trigger tokens at the start of the text t. The text t is sampled from the set of all class impressions CI^c. Again, we use the Taylor approximation of the above equation. Therefore, we get:

$$\mathbf{e}_{adv_i} = \arg\min_{\mathbf{e}'_i\in\mathcal{V}} [\mathbf{e}'_i - \mathbf{e}_{adv_i}]^\intercal \nabla_{\mathbf{e}_{adv_i}}\mathcal{L}, \tag{2.6}$$

where $\mathcal{V}$ represents the set of all words in the vocabulary, and $\nabla_{\mathbf{e}_{adv_i}}\mathcal{L}$ is the average gradient of the task loss over a batch. We model Eq. 2.6 as an iterative procedure where we initialize $\mathbf{e}_{adv_i}$ with an initialisation value of 'the'. For computing the optimal $\mathbf{e}'_i$, similar to the previous step, we take $|\mathcal{V}|$ d-dimensional dot products where d is the dimensionality of the token embedding. We use beam-search for finding the optimal sequence of tokens $\mathbf{e}'_i$ to get the minimum loss in Eq. 2.6. We score each beam using the loss on the batch in each iteration of the optimization schedule. Additionally, to generate impressions of varying difficulty, we randomly select the token from a N-sized beam of possible minimal candidates, instead of

the least scoring candidate.

Type	Direction	Trigger	Acc. Before	Acc. After
Data-based	P → N	worthless endurance useless	93.6	9.6
Data-free	P → N	useless endurance useless	93.6	9.6
Data-based	N → P	kid-empowerment hickenlooper enjoyable	80.3	7.9
Data-free	N → P	compassionately hickenlooper gaghan	80.3	8.1

Table 2.15: The table reports the accuracy drop for the BiLSTM-Word2Vec sentiment analysis model after prepending 3-word adversarial triggers generated using MINIMAL and data-based methods.

Type	Direction	Trigger	Acc. Before	Acc. After
Data-free	P → N	useless endurance useless	86.2	32
Data-free	N → P	compassionately hickenlooper gaghan	86.9	35

Table 2.16: Accuracy drop for transfer attack with data-free UAT generated by our method. We prepend 3-word adversarial triggers to the SST BiLSTM-ELMo model.

Finally, we convert the optimal $\mathbf{e}_{adv_i}$ back to their associated word tokens. Fig. 2.17 presents an overview of the process. Similar to Sec. 2.2.4.1, we initialize the iterative algorithm with a sequence ($\mathbf{e}_{adv}$) of *"the the .... the"*[**] and then follow the optimization procedure to find the optimal $\mathbf{e}_{adv}$. We handle contextual embeddings in a similar manner as in Sec. 2.2.4.1. Next, we show the application of the algorithms developed on several downstream tasks.

2.2.5 Experiments

We present our experimental setup and the effectiveness of the proposed method in terms of the success rates achieved by the crafted UATs. We test our method on several tasks including sentiment analysis, natural language inference, and paraphrase detection.

[**]We vary the initialisation sequence and sequence length to generate multiple adversarial triggers

Class	Class Impressions
Contradiction	**Hypothesis**: lynched cardinals giraffes lynched lynched a mown extremist natgeo illustration **Premise**: zucchini restrooms swimming golds weekday rock 4 seven named dart
Entailment	**Hypothesis**: civilization va physical supersonic prohibits biathlon body land muffler mobility **Premise**: gecko robed abroad teetotalers blonds pluggling sprinter speeds corks dogtrack
Neutral	**Hypothesis**: porters festivals fluent a playgrounds ratatouille buttercups horseback popularity waist **Premise**: bowler teaspoons group tourism tourism spiritual physical physical person

Table 2.17: Class Impressions for ESIM model trained for the Natural Language Inference Task

2.2.5.1 Sentiment Analysis

We use the Stanford Sentiment Treebank (SST) dataset (Socher *et al.*, 2013). Previous studies have extensively used this dataset for studying sentiment analysis (Devlin *et al.*, 2019; Cambria *et al.*, 2013). We use two models on this dataset: Bi-LSTM model (Graves and Schmidhuber, 2005) with word2vec emeddings (Mikolov *et al.*, 2018), Bi-LSTM model with ELMo embeddings (Peters *et al.*, 2018). The same models have been used in previous work (Wallace *et al.*, 2019) for generating data-dependent universal adversarial triggers. The models achieve an accuracy of 84.4% and 86.6% over the dataset, respectively. We compare our algorithm with (Wallace *et al.*, 2019) since it is demonstrated to work better than other works (Behjati *et al.*, 2019).

Class Impressions: First, we generate class impressions for the model. Table 2.14 presents 2 class impressions per class. As can be seen from the table, the words selected by the CIG algorithm highly correspond to the class sentiment. For instance, the algorithm selects positive words such as *energizes, enthrall* for the positive class, and negative words such as *spiritless, ill-conceived, laziest* for the negative class. We posit that the class impressions generated through our algorithm can be used to interpret what a model has learnt.

UAT: Next, using the class impressions generated for the models, we generate universal adversarial triggers with the UTG algorithm (Sec 2.2.4.2). In order to

Class Type: Entailment→Neutral Original Accuracy: (ESIM: 91%, DA: 90.3%)				
Data-Inputs	**Data-Type**	**Trigger**	**ESIM**	**DA**
Hypothesis and Premise	Data-Based	nobody	0.06	0.18
		whatsoever	0.6	43
		cats	0.69	0.7
	Data-Free	nobody	0.06	0.18
		no	0.1	2
		mars	0.1	0.3
Hypothesis-Only	Data-Free	monkeys	0.7	0.54
		zebras	0.5	0.39
		cats	0.69	0.7
Class Type: Neutral→Contradiction Original Accuracy: (ESIM: 88%, DA: 80%)				
Data-Inputs	**Data-Type**	**Trigger**	**ESIM**	**DA**
Hypothesis and Premise	Data-Based	shark	18	28
		moon	17	13
		spacecraft	12	8.4
	Data-Free	skydiving	14	20
		orangutan	12	75
		spacecraft	12	8.4
Hypothesis-Only	Data-Free	sleep	11	19
		drowning	15	29
		spacecraft	12	8.4
Class Type: Contradiction→Entailment Original Accuracy: (ESIM: 79%, DA: 85%)				
Data-Inputs	**Data-Type**	**Trigger**	**ESIM**	**DA**
Hypothesis and Premise	Data-Based	expert	64	73
		siblings	66	68
		championship	65	74
	Data-Free	inanimate	67	82
		final	66	68
		championships	68	85
Hypothesis-Only	Data-Free	humans	70	79
		semifinals	68	74
		championship	65	74

Table 2.18: We prepend a single word (Trigger) to SNLI hypotheses. We display the top 3 triggers created using both Validation set and Class Impressions for ESIM and show their performance on the DA. The original accuracies are mentioned in brackets.

Class	Class Impressions
Paraphase Detected	**Sentence 1**:nintendo daredevil bamba bamba the the lakers dodgers weekend rhapsody seahawks **Sentence 2**: nintendo multiplayer shawnee dodgers anthem netball the olympics soundtrack overture martial
Paraphase Detected	**Sentence 1**: mon submitted icus submit arboretum templar desires them requirements kum **Sentence 2**: lection rahul organizers postgraduate qualifying your exercises signifies its them
No Paraphase Detected	**Sentence 1**: b 617 matrices dhabi ein wm spelt rox a proportional alamo swap **Sentence 2**: drilled traced 03 02 said mattered million 0% 50% corporations a a
No Paraphase Detected	**Sentence 1**: cw an hung kanda singapore tribu chun mid 199798 nies bula latvia **Sentence 2**: came tempered paced times than an saying say shone say s copp

Table 2.19: Class Impressions for ALBERT model trained for the Microsoft Research Paraphrase Corpus

avoid selecting construct-relevant words, we remove such words[††] from our vocabulary for this task. Table 2.15 shows the results for the performance of adversarial triggers generated using our method and those by the data-based approach of (Wallace *et al.*, 2019). Despite being completely independent of data, we achieve comparable accuracy drops as (Wallace *et al.*, 2019). We are able to reduce the sentiment prediction accuracy by more than 70% for both the classes.

Transfer of Mined UATs: We check whether the triggers mined from one model also work on other models. For this, we test the triggers mined from BiLSTM-Word2Vec model on the BiLSTM-ELMo model. Table 2.16 notes the results for the same. The triggers reduce the accuracy for both the classes by more than 50%. This is significant since they are completely mined from the model without any information of the underlying distribution. We also compare the attack success rate as a function of trigger length (Fig. 2.20).

[††] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

2.2.5.2 Natural Language Inference

For natural language inference, we use the well-known Stanford Natural Language Inference (SNLI) Corpus (Bowman *et al.*, 2015). We use two models for our analysis on this task: Enhanced Sequential Inference Model (ESIM) (Chen *et al.*, 2017) and Decomposable Attention (DA) (Parikh *et al.*, 2016) with GloVe embeddings (Pennington *et al.*, 2014). The accuracies reported by ESIM is 86.2%, and DA is 85%.

Class Impressions: Modelling natural language inference involves taking in two inputs: premise and hypothesis and deciding the relation between them. The relation can be one amongst entailment, contradiction, and neutral. Following the algorithm in Sec. 2.2.4.1, we find both premise and hypothesis together after starting out from a common initial word sequence. Through this, we get a *typical* premise and its corresponding hypothesis for the three output classes (entailment, contradiction, and neutral).

One example per class for the ESIM model is given in Table 2.17. Unlike sentiment analysis, class impressions for SNLI are not readily interpretable. This is because that while a sentence from the SST corpus can be considered a combination of latent sentiments, the same cannot be assumed of a hypothesis sentence from the SNLI corpus. A statement by itself is not a characteristic hypothesis (or premise). For instance, the SST sentence "You'll probably love it." is a characteristic positive polarity sentence and can be understood to be so by the word 'love'. The same cannot be said for the SNLI premise sentence "An older and younger man smiling." SNLI class impressions give us a glance into a model's learnt deep manifold representation of premise-hypothesis pair. They are generally far away from the training data. Strong priors about the natural training distribution might be needed to make them closer to the training data, . We leave this task for future investigation.

UAT: After obtaining a batch of class impressions from the previous step, we craft the universal adversarial triggers. A comparison of the results for UATs generated using our method, and those of (Wallace *et al.*, 2019) are given in Table 2.18. As can be seen, we achieve comparable results as (Wallace *et al.*,

2019). A single word trigger is able to reduce the accuracy of the entailment class from 90.3% to 0.06%.

Hypothesis Only UATs: Several recent research studies have indicated that the annotation protocol for SNLI leaves artefacts in the dataset such that by giving just hypothesis, one can obtain 67% accuracy (Gururangan *et al.*, 2018; Poliak *et al.*, 2018). Following that line of study, we generate only the hypothesis class impressions using the CIG algorithm. Then, we generate triggers over the hypothesis-only generated class impressions. Table 2.18 notes the results for the hypothesis-only attacks. We find that hypothesis-only triggers perform equivalently to hypothesis and premise attacks. This provides further proof that there are many biases in the SNLI dataset and more importantly, the models are using those biases as class representations and adversarial triggers actively exploit these (§2.2.6).

Transfer of Mined UATs to Other Models: To determine how the triggers mined from one model transfer to another, we test both data-based and our data-free triggers generated using the ESIM model on the DA model. Table 2.18 shows the results. We check the transfer attack performance in two cases: where both hypothesis and premise are given and where only the hypothesis is given. It can be seen that even though both the models are architecturally very different, the triggers transfer remarkably well for both cases. For instance, for the entailment class, the original and transfer attack accuracy drops are comparable. It is also noteworthy that our results are equivalent to (Wallace *et al.*, 2019) even for transfer attacks.

2.2.5.3 Paraphrase Identification

For paraphrase identification, we use the Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005). Paraphrase identification is the task of identifying whether two sentences are semantically equivalent. We use the ALBERT model (Lan *et al.*, 2020) for the task. It reports an accuracy of 89.9% over this.

Class Impressions: Similar to natural language inference, here, the models require two input sentences. The task of the model is to identify whether the two

sentences are semantically the same. The class impressions generated on the ALBERT model are given in Table 2.19. We find that similar to the SNLI corpus, the MRPC class impressions are not readily interpretable. For specific examples like the first example in the table, we find that sometimes words related to one topic occur as class impressions. Words like 'nintendo' and 'daredevil' in sentence one and 'multiplayer' and 'anthem' often occur in the context of multiplayer digital games. We should have got similar class impressions in an ideal scenario for sentences 1 and 2 for actual paraphrases. However, we find that the model considers even those sentence pairs (example 2) as paraphrases that have zero vocabulary or topic overlap. This indicates that the model is performing a similarity match in the high dimensional data manifold. We do some analysis for this in Sec. 2.2.6. We leave the further investigation of this for future work.

UAT: Table 2.20 notes the performance of 3 word data-free adversarial triggers generated using MINIMAL. As can be seen, the mined artefacts reduce the accuracy for both classes by more than 70%.

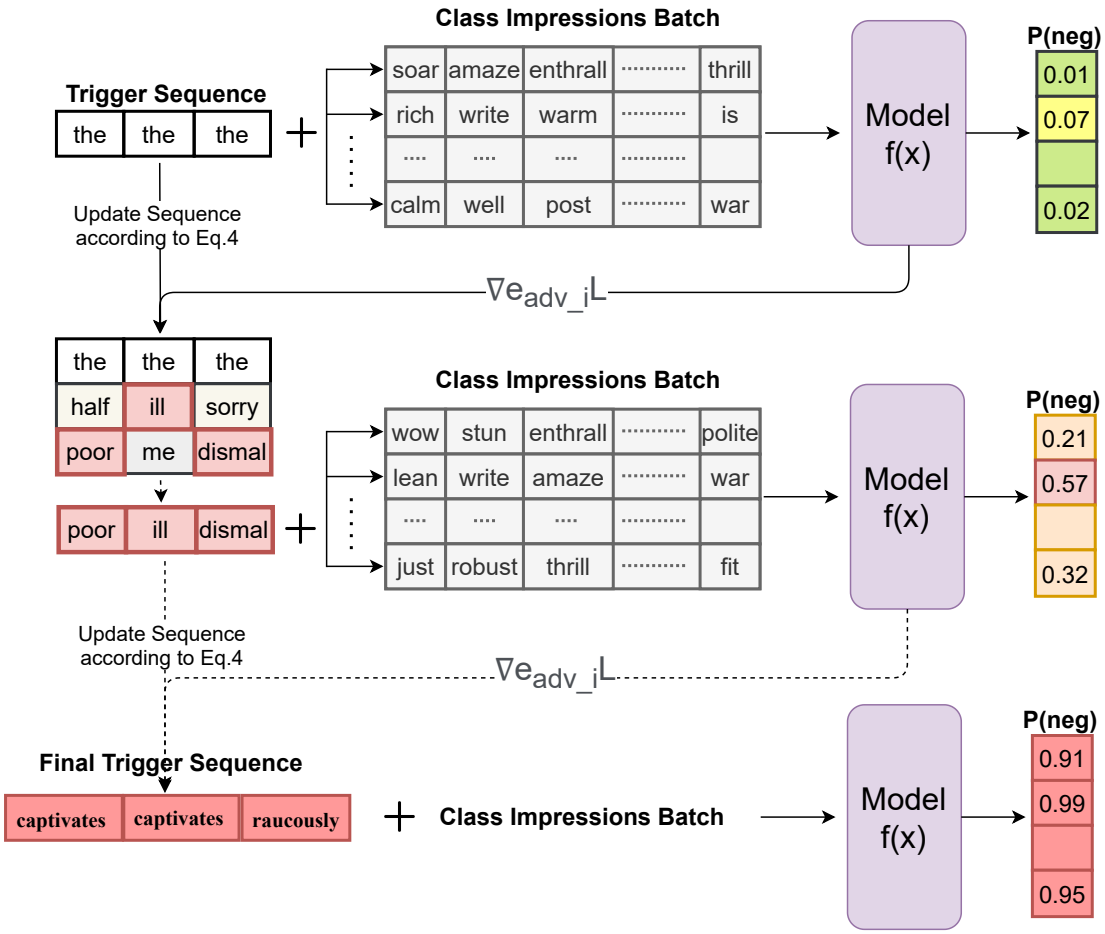


Figure 2.17: Iterative Universal Trigger Generation (UTG) Algorithm.

2.2.6 Analyzing the Class Impressions

We futher analyze class impressions and their relationship with universal adversarial triggers. Specifically, we try to answer these questions: which words get selected as class impressions, why are we able to find universal adversarial triggers from a batch of class impressions and no train data distribution is required? We also try to relate it to the observation made by (Gururangan *et al.*, 2018; Poliak *et al.*, 2018), which ranked the dataset artefact words by calculating their pointwise-mutual information (PMI) values for each class. We further show that the trigger words align very well with dataset artefacts.

Type	**Direction**	**Trigger**	**Acc. Before**	**Acc. After**
Data-free	P → N	insisting sacrificing either	95	45
Data-free	N → P	waistband interests stomped	80.9	61.6

Table 2.20: Accuracy drop for the ALBERT paraphrase identification model after prepending 3-word adversarial triggers generated using MINIMAL.

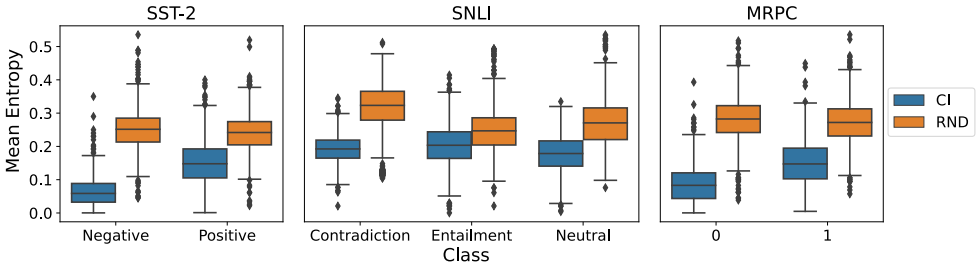


Figure 2.18: Mean Entropy of class impression words and 350 words randomly selected from the SST, SNLI, and MRPC dataset vocabularies.

Class Impression Words: For analyzing why certain words are selected as representatives of a particular class, we find the discriminative power of each word by calculating its entropy. Concretely, we calculate entropy of the random

Stanford Sentiment Treebank			
Positive	**%**	**Negative**	**%**
beautifully	99.97	dull	99.99
wonderful	99.95	worst	99.99
enjoyable	99.94	suffers	99.98
engrossing	99.94	stupid	99.98
charming	99.89	unfunny	99.97
Impression Average	73.89	Impression Average	77.97

Table 2.21: PMI percentiles for sample class impression words and their average

Microsoft Research Paraphrase Corpus			
Paraphrase	**%**	**Non-Paraphrase**	**%**
experts	99.89	biological	99.91
such	99.84	important	99.39
only	99.67	drug	99.92
due	99.65	case	98.91
said	99.57	among	98.73
Impression Average	77.23	Impression Average	81.89

Table 2.22: PMI percentiles for sample class impression words and their average

Stanford Natural language Inference					
Contradiction	**%**	**Entailment**	**%**	**Neutral**	**%**
naked	99.99	human	99.91	about	99.73
sleeping	99.97	athletic	99.73	treasure	99.06
tv	99.96	martial	99.71	headed	99.05
asleep	99.96	clothes	99.53	school	98.87
eats	99.93	aquatic	99.38	league	98.83
Average	67.89	Average	70.89	Average	68.97

Table 2.23: PMI percentiles for sample class impression words and their average

Ground Truth→Attacked Target	**Trigger**	**ESIM**
Entailment→Neutral Accuracy:88%	beatboxing	77
	insects	68
	reclining	83
Entailment→Contradiction Accuracy:79%	qualities	70
	coexist	71
	stressful	70
Neutral→Contradiction Accuracy: 79%	disoriented	69
	arousing	67
	championship	65
Neutral→Entailment Accuracy: 91%	championship	0.1
	semifinals	0.9
	aunts	0.5
Contradiction→Entailment Accuracy: 91%	ballet	5
	nap	2
	olives	9
Contradiction→Neutral Accuracy: 88%	nap	14
	hubble	21
	snakes	9

Table 2.24: We prepend a single word trigger to SNLI hypotheses. We take the first word from all ground truth class impressions and evaluate them on class impressions of the target class.

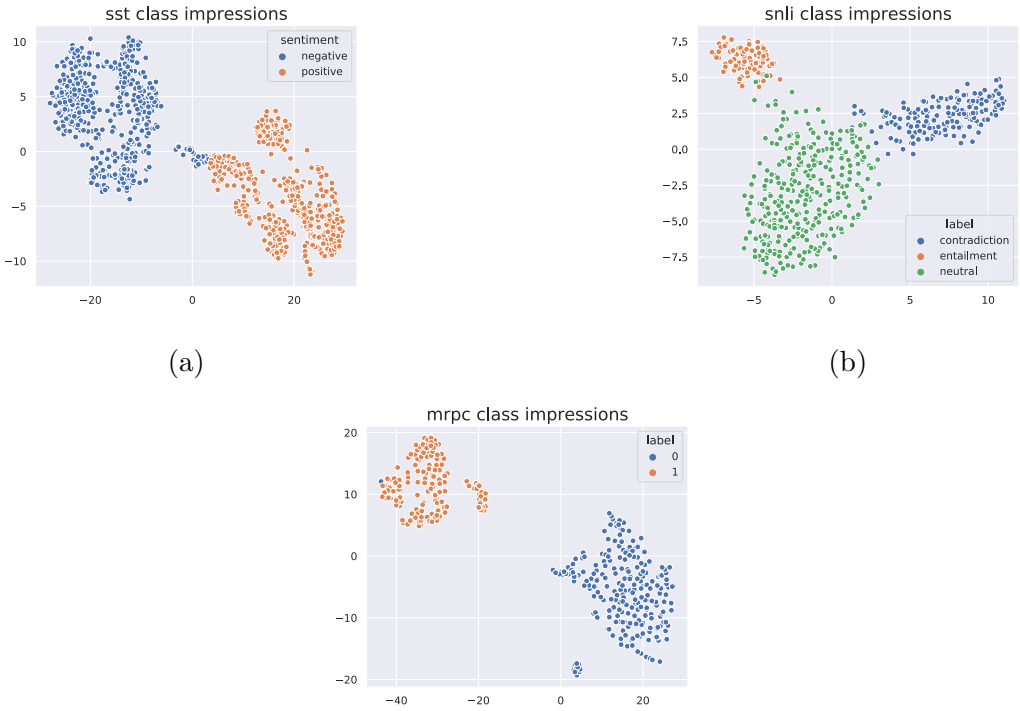


Figure 2.19: t-SNE plots for SST, SNLI, and MRPC class impression words. The words from the different class impressions form distinct clusters based on their classes for all three datasets. The clusters are shown in different colors based on their classes.

variable $Y|X$ where Y denotes a model class and X denotes the word level feature. Formally, we compute:

$$\mathbb{H}(Y|X) = -\sum_{k=1}^{K} p(Y=k|X)\log_2 p(Y=k|X) \tag{2.7}$$

for the class impression words and we compare them with randomly chosen words from the model vocabulary. Fig. 2.18 shows the results for SST, SNLI, and MRPC datasets. Interestingly, we find that the words which form class impressions are low entropy features. These words are much more discriminative than other randomly sampled words for all three datasets. This is further reinforced by Fig. 2.19 where we show t-SNE plots for all the datasets . They show that words from different class impressions form distinct clusters.

Fig. 2.18 shows that CIG algorithm selects low entropy features as representatives of different classes. However, it does not show the class-preference of these low entropy word-features. We hypothesize that those words become representa-

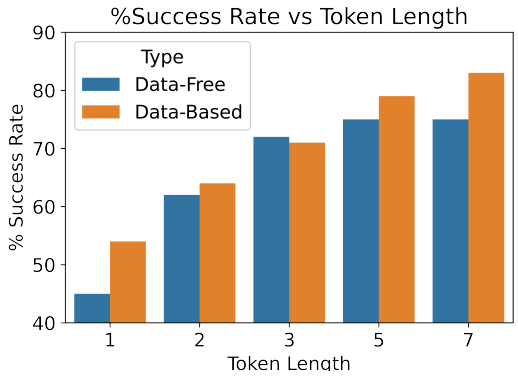


Figure 2.20: Attack success rate as a function of trigger length

tives of a particular class with a higher PMIs with respect to that class. In order to show this, we calculate PMI values of class representatives for each class and note that class representatives have a higher PMI for their own class than other classes. Formally, we compute:

$$PMI(word, class) = log \frac{p(word, class)}{p(word, .)p(., class)} \tag{2.8}$$

We use add-10 smoothing for calculating this. We then group each class impression word based on its target class and report their PMI percentile. We show the results in Tables 2.21-2.23. It can be seen that class representatives have very high PMI percentiles. Previous studies have characterized high PMI words as *dataset artefacts* (Gururangan *et al.*, 2018; Poliak *et al.*, 2018). Wallace *et al.* (2019) have also shown that universal adversarial triggers have a high overlap with these dataset artefacts and consequently have high PMI values. Since we observe that class representatives too have high PMI values, we hypothesize that they could act as good adversarial triggers.

Following this, we postulate that adding class impression words of one class to a real example of another class should change the prediction of that example. For validating this, we conduct an experiment where we take words from class impressions of class c_i and prepend them to real examples of class c_j. Table 2.24 shows the results of the experiment over SNLI dataset. As can be seen, the results are very promising.

We observe that the class which was more adversarially unsecure (*Entailment*$>_{adv-unsecure}$*Contradiction*) has better class impression words. These words,

when added to examples of other classes, produce more successful perturbations. For e.g., when entailment words are added to contradiction examples, they reduce the accuracy from 91% to less than 10%. On the other hand, contradiction was adversarially more secure, and hence there is no appreciable reduction in the accuracy of any other class upon adding the contradiction class impression words[‡‡]. This result can potentially help dataset designers design more secure datasets on which the model-makers can train adversarially robust models.

The above analysis shows that we can get class-impressions and adversarial triggers from dataset itself by computing entropy and PMI values. Moreover, our experiments in Sec. 2.2.5 show that one can equivalently mine models to get class impressions and adversarial triggers. Therefore, we conclude that we can craft both class impressions and adversarial triggers given either dataset or a well-trained model (*i.e.*, the one which can model training data distribution well). Further, the models represent their classes with dataset artefacts. These artefacts are also responsible for making them adversarially unsecure. The lesser the dataset artefacts in a class, the lesser is a trained model's representative capacity for that class, and the more is the model's adversarial robustness for that class. We would like to further develop on these initial results to better dataset design protocols in future work.

2.2.7 Conclusion and Future Work

This paper presents a novel data-free approach, MINIMAL to mine natural language processing models for input-agnostic (universal) adversarial triggers. Our setting is more natural, which assumes an attacker does not have access to training data but only the trained model. Therefore, existing data-dependent adversarial trigger generation techniques are unrealistic in practice. On the other hand, our method is data-free and achieves comparable performance to data-based adversarial trigger generation methods. We also show that the triggers generated by our algorithm transfer remarkably well to different models and word embeddings. We achieve this by developing a combination of model inversion and adversarial

[‡‡]We find similar results on the MRPC dataset. We did not do these experiments for the SST dataset since SST class impression words are construct-relevant words and hence are bound to change sentiment scores while the same is not true for the other two datasets.

trigger generation attacks. Finally, we show that low entropy word-level features occur as adversarial triggers and hence one can equivalently mine either a model or a dataset for these triggers.

We conduct our analysis on word-level triggers and class impressions based model inversion. While this analysis leads to crucial insights into dataset design and adversarial trigger crafting techniques, it can be extended to multi-word contextual analysis. This will also potentially lead to better dataset design protocols. We are actively engaged in this line of research. Further, another research focus can be to generate natural-looking class impressions and, consequently adversarial triggers.

2.3 Chapter Conclusion

What does an advertisement say that makes people change their beliefs and actions? With limited works, the computational study of rhetoric of this all-pervasive form of marketing communication is still in its infancy. In this chapter, based on the well-developed social psychology and marketing literature, we develop and release the largest vocabulary of persuasion strategies and labeled dataset. We develop several models for predicting persuasion strategies for video and image based ads. Further, we show the performance of these models on several other advertisement-understanding related tasks, including topic, emotion, and question-answering.

Chapter 3

Modeling Behavior: A Case For Large Content and Behavior Models

In the previous chapter, we dealt with the first culture of social science, explanation, and how to enable it at scale by using machine learning techniques of computer vision and NLP. Marketers make many decisions on a regular basis: what marketing campaign to launch, who to target, what the message should be, which channel it should be sent on, when it should be sent, and how frequently. Extraction of information about advertisements (for example, emotion, persuasion strategy, topic, and question answering) and correlating them with key performance indicators (KPIs) helps decision-makers (in this case, human marketers) to understand and execute campaigns better. Now, in this chapter, we turn to the question of how to encode the complete communication pipeline to enable better and possibly completely automated decision-making.

Thanks to the digitization of various aspects of life, humanity has been collecting a lot of data over the last two decades. For example, let's take the case of email marketing, one of the first marketing tools leveraging Internet technology. Say a Walmart marketer sends a Black Friday offer about a price drop on Apple devices to John, a 27-year-old male grad student living in Buffalo. The email was received at 09:57 AM and opened at 02:00 PM. Upon opening the email, email content consisting of a carousel of four images and three lines is dynamically fetched from the backend. John takes 5 seconds to scan the email quickly, scrolling halfway through, before deciding to click on a photo. During this single macro-transaction, a series of micro-transactions are recorded and a host of machine learning and software systems are required to function together to make a sequence of decisions.

Amongst all the recorded transactions and algorithms, let's discuss the most prominent ones that are important for our use case. Much before sending the email, depending on business needs, the marketer decides to launch a particular

campaign. The business need, for example, in this case, could be precipitated by an upcoming event or festival (Black Friday) or a rising inventory of Apple products. The next step is the creative process, where the marketer designs the email pods consisting of text and images by herself or with a team of creatives. The marketer has to decide the target segments (of which John will be a part). Next, an algorithm has to decide when to send the email and the subject line. Post this, a series of software technologies helps to send the email to the right people on time. When John decides to open the email, an event gets recorded in the backend recording `(customer ID, transaction ID, email ID, time of opening the email, device, email client, [other metadata])`. A personalization system then dynamically selects the email content and sends it to John's device. Those get recorded with the transaction ID. Scrolling on the email also generates transactions recording which images and text were sent to John's device. Further, when John decides to click on one link, another transaction gets recorded of the type `(transaction ID, customer ID, link, time of click, email client, device, [other metadata])`. On an abstract level, all of these transactions can be represented by the seven factors of communication: `(communicator, message, time of message, channel, receiver, time of effect, effect)`.

If this email were sent to a million subscribers, one email message would result in several hundred thousand transactions getting recorded (assuming single digit click through rate, which is typical of most campaigns). These transactions capture behavior data of the subscribers in response to a single email sent by the communicator, Walmart. This example illustrates the size and nature of behavioral data that gets captured. Notice that for a message, it is always the case that there is one sender and multiple receivers (an invariance noticed as early as 1950s (Meier, 1959)). Therefore, the scale of behavioral transactions generated is several orders higher than the number of unique pieces of content.

Given the magnitude of behavioral data collected, the natural question is can all that data be used to answer questions related to human behavior prediction, explanation, and optimization. Therefore, the research questions that we investigate in this chapter follow this natural line of inquiry:

1. How can behavior data help? Can behavior data help us to achieve the following goals:

   (a) Behavior Prediction
   (b) Behavior Explanation
   (c) Behavior Optimization?

2. How should we encode behavior data?

3. What kind of behavior can help?
   (a) How can implicit (like eye movements) and explicit (like clicks, likes, and views) behaviors help?
   (b) Can synthetically generated behavior data help?

To solve the behavior problems listed before, we can take inspiration from how the problem of learning natural language is being solved in the domain of large language models (LLMs). Raffel *et al.* (Raffel *et al.*, 2020), in their seminal work on T5, mention that the basic idea underlying large language models is to treat every text processing problem as a "text-to-text" problem, *i.e.*, taking the text as input and producing new text as output. This framework allows for a direct application of the same model, objective, training procedure, and decoding process to every task we consider. Further, this allows us to pre-train a model on a data-rich task like the next-word prediction, which can then be transferred to downstream tasks. Notably, thanks to the Internet, the next-word prediction task has huge amounts of available data. Consider the Common Crawl project (https://commoncrawl.org), one common source of data included in most language models. It produces more than 20TB of text per month sampled from random web pages across the internet.

T5 and other language models like GPT-3, Pythia (Biderman *et al.*, 2023), and LLama (Touvron *et al.*, 2023) can solve a wide variety of tasks, including the ones for which they were not explicitly trained. For instance, language models trained on the next word prediction task showed generalization capabilities across a wide variety of tasks like question-answering, summarization, natural language inference, and translation (Brown *et al.*, 2020). Recently, a series of papers have shown that this generalized "world understanding" captured in LLMs can be leveraged to enable them to "see" (Liu *et al.*, 2023*a*; Li *et al.*, 2023*c*,*a*; Zhu *et al.*, 2023; Ge *et al.*, 2023; Zhang *et al.*, 2023*a*; Bhattacharyya *et al.*, 2023). This is a significant capability enhancement since a model trained in language only settings can be made to reason about images and videos. These papers follow the same

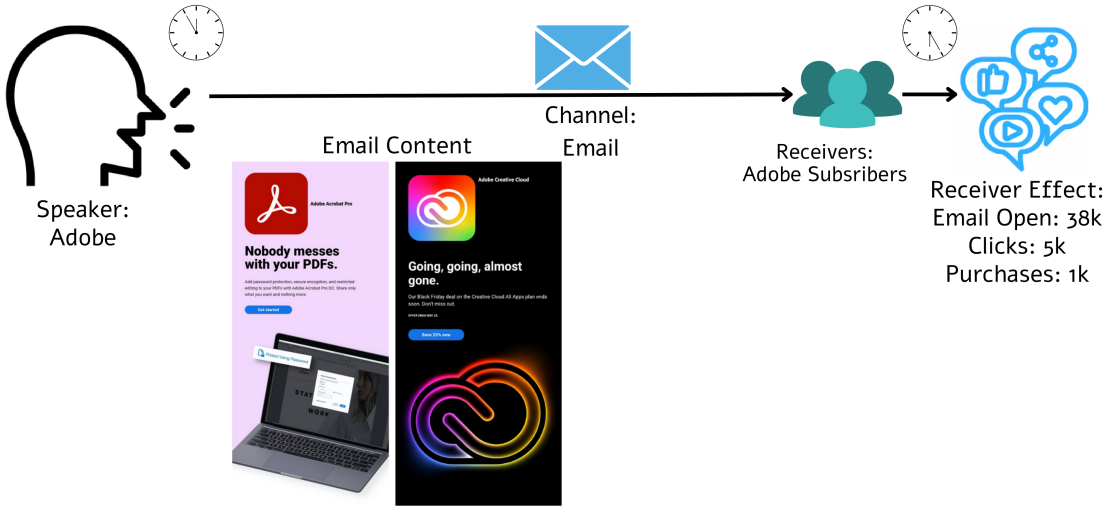


Figure 3.1: Communication process can be defined by seven factors: Communicator, Message, Time of message, Channel, Receiver, Time of effect, and Effect. Any message is created to serve an end goal. For marketers, the end goal is to bring in the desired receiver effect (behavior) (like clicks, purchases, likes, and customer retention). The figure presents the key elements in the communication pipeline - the marketer, message, channel, receivers, and finally, the receiver effect.

transfer learning approach advocated by T5, where they convert visual information to language space to leverage the "text-to-text" framework. They show that it is possible to teach a large language model, the new modality of vision, without needing to pre-train the model from scratch. Rather, using only a few million tokens, it is possible to scale LLMs' abilities to vision as well. Following this chain of thought, it could be possible to solve the effectiveness problem by posing it as a "text-to-text" problem. This is one of the paradigms we explore in this work. We show behavior generalization using several different types of behaviors.

Another possible way to integrate behavior with text is an encoder approach, which we will detail next. While behavior is a downstream effect of content, behavior contains signals about the content sent to the receiver and can help improve content-understanding and natural language processing. For instance, integration of human gaze data into neural network architectures has been explored for a range of computer vision tasks such as image captioning, visual question answering, and tagging (Karessli *et al.*, 2017; Yu *et al.*, 2017; He *et al.*, 2019; Boyd *et al.*, 2022). In language processing, tracking a reader's eye movements provides information about the cognitive processes of text comprehension (Rayner *et al.*, 2006; Just and Carpenter, 1980). Hence, recent research has utilized features

gleaned from readers' eye movement to improve the performance of complex NLP tasks such as sentiment analysis (Long *et al.*, 2017; Mishra *et al.*, 2016*c*), sarcasm detection (Mishra *et al.*, 2016*b*), part-of-speech tagging (Barrett *et al.*, 2016*b*), NER (Hollenstein and Zhang, 2019*a*), and text difficulty (Reich *et al.*, 2022). While these studies show promise that behavior can be used to extract information about content, these are done in relatively small-scale lab settings needing real-time behavior to infer about content. Given these limitations, these approaches are not possible to scale. Scale helped LLMs to learn language. We therefore explore the paradigm of synthetic behavior generated over content and then scale it over to fine-tune a large language model to understand the possibilities of this paradigm better. We cover both the approaches next.

A notable advantage of both "text-to-text" and encoder paradigms is their inherent ability to handle missing modalities. Despite LLMs being trained on trillions of tokens, they exclude image pixels yet can reason about visual content through text conversion (Bhattacharyya *et al.*, 2023) or via image encoders that map visual information to LLM-compatible representations. As demonstrated by Li *et al.* (2023*b*); Liu *et al.* (2023*a*), teaching vision capabilities requires only a few million image-text pairs. Similarly for behavior data, we demonstrate two effective approaches: (1) converting behavioral signals directly to text (Sec 3.2.1), or (2) employing a behavior encoder that transforms behavioral data into dense representations compatible with LLMs (Sec 4.1.1). Importantly, these methods function robustly with incomplete datasets through implicit data imputation—neither approach requires simultaneous presence of all modalities (content, behavior, sender, receiver, time, channel) during training. This partial supervision framework allows models to learn cross-modal relationships even when faced with missing or sparse behavioral signals.

3.1 Introduction

Shannon and Weaver (1949), in their seminal paper on communication, includes all of the procedures by which one mind may affect another. This includes all forms of expression, such as words, gestures, speech, pictures, and musical sounds.

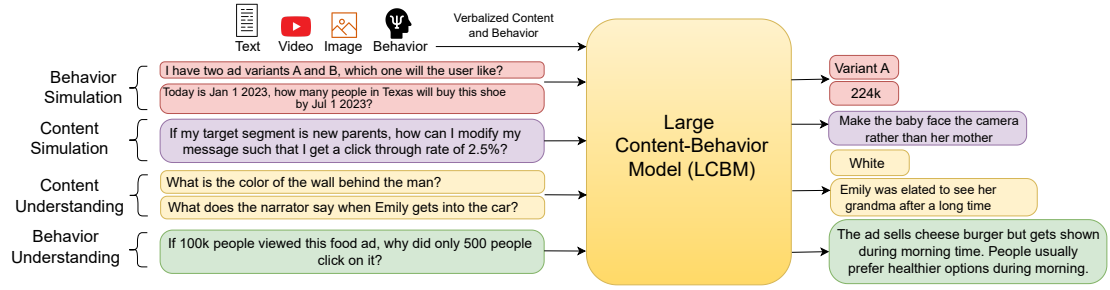
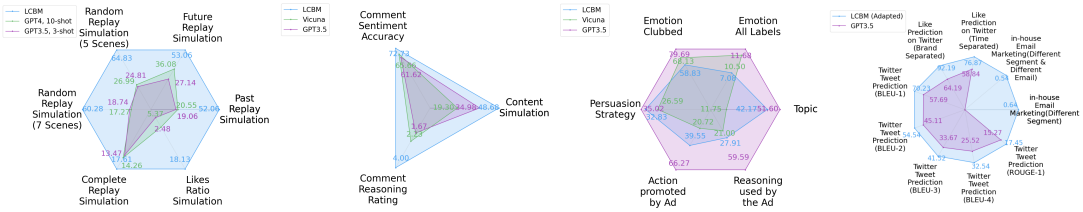


Figure 3.2: Encoding and predicting content (images, videos, and text) and behavior in the language space. Large Content Behavior Models (LCBMs), once trained, can enable a host of different applications, including behavior simulation, content understanding, content-behavior optimization, and content-behavior understanding.



(a) (b) (c) (d)

Figure 3.3: Comparison of GPT-3.5, GPT-4, Vicuna-13B, and LCBM-13B on: (a) Behavior Simulation accuracy on two types of behaviors: replay value prediction and likes/views prediction. The task is, given the video content and channel information, to predict replay values corresponding to each scene and the ratio of likes to views. (b) Content simulation and behavior understanding tasks. The task for content simulation is, given the channel information and scene-level behavior, to predict the scene content. Given information on the video platform and the video content, the task of behavior understanding is to predict and explain the sentiments of the viewers and the commenters. Six evaluators scored the models' explanations between 0-5 to get the predicted sentiment and explanation scores by comparing the ratings and reasons with the user comments. The annotators did not know which model gave the reasoning. (c) Content understanding tasks. We evaluate four tasks: emotion, topic, and persuasion strategy prediction, and action-and-reason understanding. (d) Behavior Simulation on in-house Email Marketing Data (R^2 score) and Twitter likes (accuracy), and Content Simulation on Twitter tweet prediction (BLEU/ROUGE scores). It can be noted that on the behavior simulation, content simulation, and behavior understanding tasks, LCBM performs better than 3-shot GPT-3.5 and 10-shot GPT-4 (covering a larger area. On the content understanding tasks, while LCBM outperforms similar-sized Vicuna models, GPT-3.5 performs better. However, we also note that GPT-3.5 and GPT-4 are at least 12 times larger than LCBM-13B. Further, we show the behavior domain adaptation results in Table 3.9, 3.3, 3.4.

They mentioned that the broad problem of communication can be studied at three levels: technical, semantic, and effectiveness.

Level A: Technical. How accurately can the symbols of communication be transmitted?

Level B: Semantic. How precisely do the transmitted symbols convey the desired meaning?

Level C: Effectiveness. How well does the received meaning induce the desired conduct in the receiver?

These three levels build on top of each other. Thus, solving the problem at Level C necessarily requires solving the corresponding problems at Levels A and B.

Since the publication of this seminal paper, the tremendous growth in the field of telecommunications, particularly the advent of the Internet and mobile devices, has led to affordable, wide-scale solutions for Level A. With the recent advances in large language models (LLMs) such as BERT (Devlin *et al.*, 2019), GPT-3 and 4 (Brown *et al.*, 2020; OpenAI, 2023), T5 (Raffel *et al.*, 2020), and many more, we have witnessed a significant improvement in the performance of various Natural Language Processing (NLP) tasks. LLMs in zero- or few-shot settings can easily handle tasks such as question answering, summarization, translation, and many more. This has helped us progress towards solving Level B to a large extent. However, the Level C problem of effectiveness remains largely unsolved. Effectiveness refers to designing messages that can fulfill the communicators' underlying objectives, such as explaining complex concepts to the receivers and informing the receivers' choices (*e.g.*, when making purchase decisions).

How do we solve the effectiveness problem while retaining the other two levels? To solve the effectiveness problem, we can take inspiration from how the semantic problem is being solved. Raffel *et al.* (2020), in their seminal work on T5, mention that the basic idea underlying large language models is to treat every text processing problem as a "text-to-text" problem, *i.e.*, taking the text as input and producing new text as output. This framework allows for a direct application of the same model, objective, training procedure, and decoding process to every

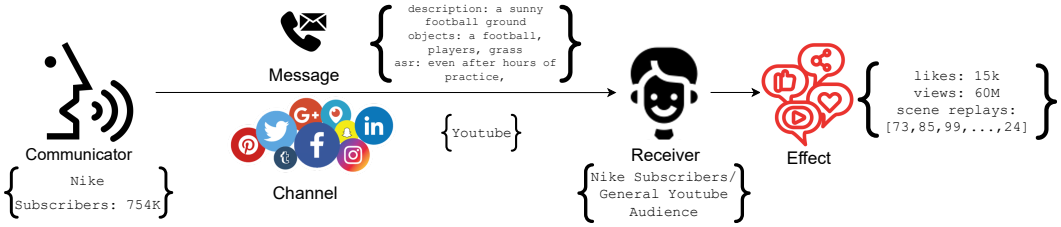


Figure 3.4: Five factors of communication: Communicator, Message, Channel, Receiver, and Effect.

task we consider. Further, this allows us to pre-train a model on a data-rich task like the next-word prediction, which can then be transferred to downstream tasks. Notably, thanks to the Internet, the next-word prediction task has huge amounts of available data. Consider the Common Crawl project (https://commoncrawl.org), one common source of data included in most language models. It produces more than 20TB of text per month sampled from random web pages across the internet.

T5 and other language models like GPT-3, Pythia (Biderman *et al.*, 2023), and Llama (Touvron *et al.*, 2023) can solve a wide variety of tasks, including the ones for which they were not explicitly trained. For instance, language models trained on the next word prediction task showed generalization capabilities across a wide variety of tasks like question-answering, summarization, natural language inference, and translation (Brown *et al.*, 2020). Recently, a series of papers have shown that this generalized "world understanding" captured in LLMs can be leveraged to enable them to "see" (Liu *et al.*, 2023*a*; Li *et al.*, 2023*c*,*a*; Zhu *et al.*, 2023; Ge *et al.*, 2023; Zhang *et al.*, 2023*a*; Bhattacharyya *et al.*, 2023). This is a significant capability enhancement since a model trained in language only settings can be made to reason about images and videos. These papers follow the same transfer learning approach advocated by T5, where they convert visual information to language space to leverage the "text-to-text" framework. They show that it is possible to teach a large language model, the new modality of vision, without needing to pre-train the model from scratch. Rather, using only a few million tokens, it is possible to scale LLMs' abilities to vision as well. Following this chain of thought, it could be possible to solve the effectiveness problem by posing it as a "text-to-text" problem. This is the paradigm we explore in this work.

How can we pose the effectiveness problem as a text-to-text prob-

lem? The problem of effect is to know what the receiver does after receiving the message (Shannon and Weaver, 1949). In general, for a piece of content, other than the content itself, we often have information about *who* consumes the content and what his *action* is on consuming the content. The latter is the effect described in Shannon's three levels of communication. For instance, an email, as a message from the communicator to the receiver, elicits certain actions from the receiver like link-clicks, replies, and read-time. While LLMs are trained on trillions of tokens of content, the training does not include the receiver effect. For instance, Enron Email (Klimt and Yang, 2004) is a popular corpus that is included in the training of LLMs like Pythia (Biderman *et al.*, 2023). It contains 600K email content sourced from the Enron corporation, which LLMs use to learn how to write emails. However, it does not contain data about the receivers' activities, such as whether they opened the email, how long they kept it open (read-time), and what their reply was. Similarly, while major text corpora include a large number of public blogs and user forums to train LLMs like CommonCrawl, they are stripped of receiver behavior on forum messages, such as the number of likes, shares, and replies, before including them in LLM training (for instance, see (Biderman *et al.*, 2022; Penedo *et al.*, 2023)). To pose the effectiveness problem as a text-to-text problem, we can include these *behavior tokens* in the text along with content tokens and train the LLM to model both of those in the same space. This might help the LLM simulate the receiver effect, optimize for it, and reason about it.

In this paper, we show initial experiments to integrate behavior as a new modality to increase the scope of multimodal LLMs from only content to both content and behavior. We call this new type of model a Large Content Behavior Model (LCBM). This class of models shows promise in enabling the LLMs to not only reason about content but also reason about and predict human behavior over that content. Further, LCBMs have the potential for behavior domain adaptation where models trained on one type of behavior can generalize on another behavior type (Fig. 3.3). Behavior simulation can enable many real-world applications, such as content recommendation, customer journey optimization, and A/B testing. To build LCBM, we introduce behavior instruction tuning (§3.2.4), an attempt to extend the instruction tuning paradigm to behavior space, bringing all five communication factors (communicator, message, channel, receiver, and effect)

into the same space (Fig. 3.4). Similar to Brown *et al.* (2020); Raffel *et al.* (2020); Liu *et al.* (2023*a*); Ge *et al.* (2023), we do not design best-in-class predictors for any of the downstream tasks. Rather, we show a model which shows generalization capabilities across a wide variety of content- and behavior-related tasks. To summarize, our paper makes the following two contributions:

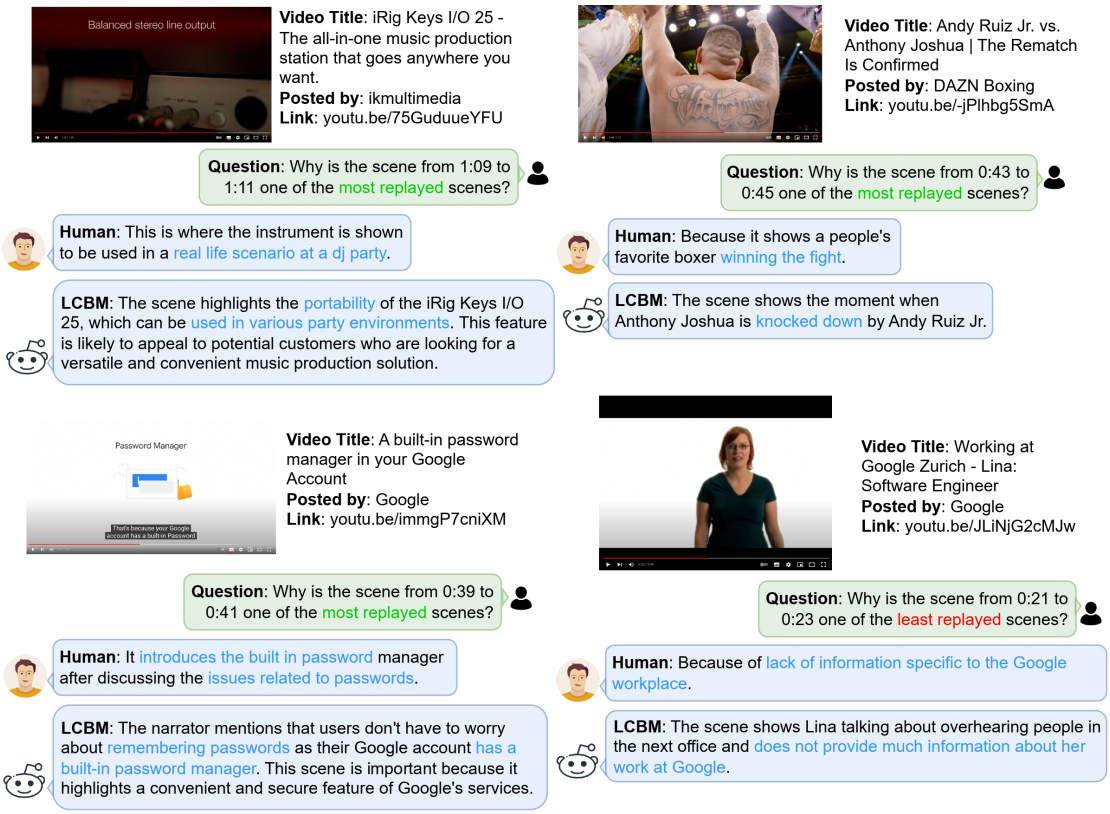


Figure 3.5: A few examples showing LCBM's ability to understand and explain human behavior of scene replayability. We compare it against human-provided explanations of the same.

- **Large Content Behavior Model (LCBM).** We develop a large multimodal model that shows capabilities of behavior simulation (given content), content simulation (given behavior), content understanding, and behavior understanding (Fig. 3.2). Following the text-to-text framework, we connect the Vicuna LLM (Touvron *et al.*, 2023; Chiang *et al.*, 2023) with an open-set visual encoder of EVA-CLIP (Sun *et al.*, 2023) and instruction fine-tune it end-to-end on behavior instruction data. EVA-CLIP and QFormer (Li *et al.*, 2023*a*) help the model to understand visual content in the language space, making it a Vision Language Model (VLM). During behavior instruction tuning, we teach the model to predict behavior given content and content given behavior using various instruction tasks (§3.2.4). This helps us teach behavior modality to the VLM while grounding it in the natural language space. We use three datasets to show the performance of LCBM: a dataset consisting of YouTube videos as the content and the corresponding retention graph, likes, the number of views, and comment sentiment as receiver behavior; a dataset consisting of Twitter posts (text, images, and videos) and corresponding human behavior (like counts) extracted from 168 million tweets across 10135 enterprise Twitter accounts from

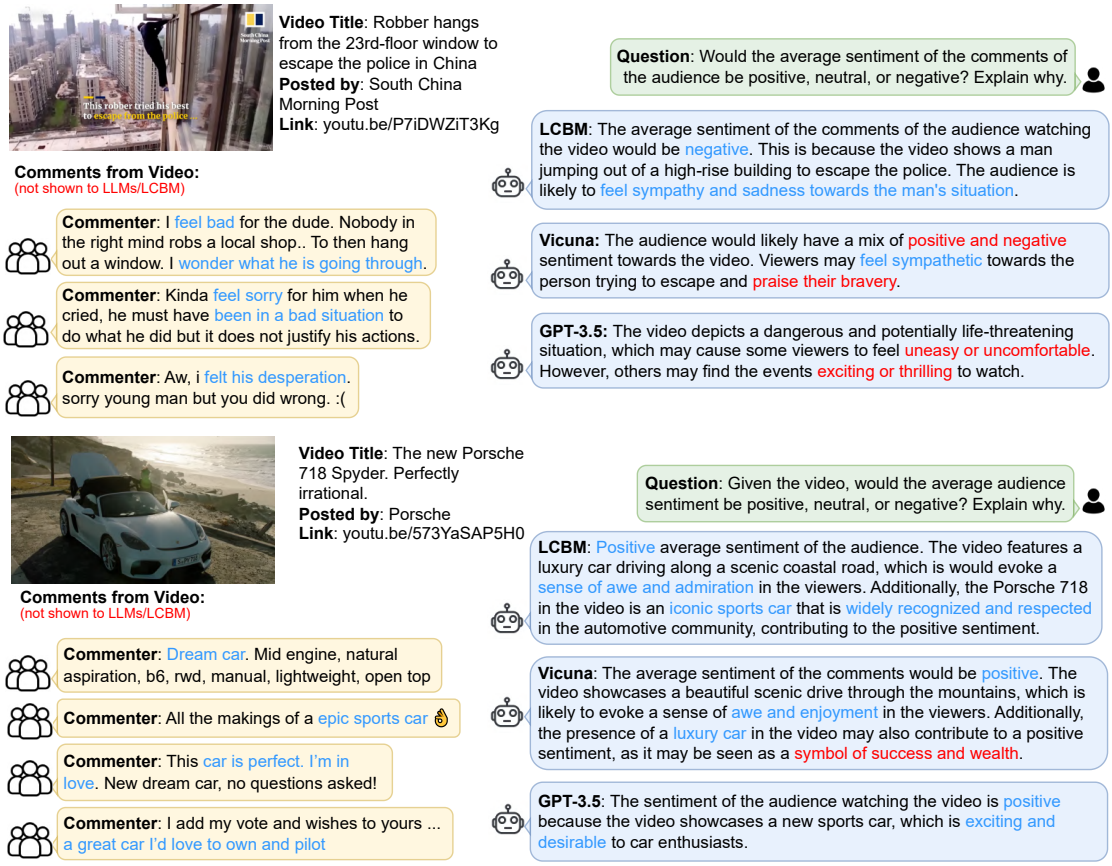


Figure 3.6: A few examples showing LCBM's ability to understand and explain human behavior of audience sentiment. We also compare it against other models like Vicuna and GPT-3.5.

2007 to 2023 (Khurana *et al.*, 2024); and an internal dataset of in-house Marketing Emails[‖] (content) and the click-through rate corresponding to each segment they were sent to (behavior). We observe that teaching the LCBM behavior and content simulation improves its capabilities on them (expected), but the model also shows signs of domain-adaptation in behavior modality (few-shot capability, *unexpected*) (Tables 3.9,3.3,3.4) and improvements in behavior understanding (Figs. 3.6,3.5,§3.3) (zero-shot capability, *unexpected*) (Brown *et al.*, 2020). See Fig. 3.3 for a radar plot of all the capabilities and comparisons of performances across LCBM and state-of-the-art LLMs: GPT-3.5 and GPT-4.

- **Dataset and Test Benchmark.** To spur research on the topic of large content and behavior models, we release our generated behavior instruction fine-tuning data from over 40,000 public-domain YouTube videos and 168 million Twitter posts. The data contains: 1) YouTube video links, automatically extracted key scenes, scene verbalizations, replay graph data, video views, likes, comments, channel name, and subscriber count at the time of collection, and 2) Twitter extracted account names, tweet text, associated media (image and video) verbalizations (including image captions, keywords, colors, and tones), tweet timestamps, and like counts (Khurana *et al.*, 2024). We also release a benchmark to test performance on the joint content behavior space (§3.2.3), introducing two types of tasks in this space: predictive and descriptive. In the predictive benchmark, we test the model's ability to predict behavior given the content and predict content given the behavior. In the descriptive benchmark, we validate its explanation of human behavior by comparing it with ground-truth annotations we obtain from human annotators that try to explain human behavior. See Figs. 3.6,3.5 for a few examples.

3.1.1 Related Work

Models of Human Communication: Communication is the situation in which a source transmits a message to a receiver with conscious intent to affect the latter's behaviors (Osgood *et al.*, 1957; Miller, 1966). Thus, in the most general terms, communication implies a sender, a channel, a message, a receiver, a relationship between sender and receiver, an effect, a context in which communication occurs and a range of things to which 'messages' refer (McQuail and Windahl, 2015; Lasswell, 1948). As per this, all of the content produced by humanity is essentially communication from a sender to a receiver over some channel and with some effect. Despite much research on communication in social sciences since the 1900s, there has been little adoption of it in machine learning modeling. A prime artefact of this is that the biggest models in machine learning (LLMs) are trained only on content (messages) and ignore other factors in communication (the intended

[‖]We obtain in-house Marketing Emails dataset by collaborating with the in-house team.

receiver, channel, and behavior) even when they are available.

Prior Efforts To Model Behavior: While there has been much research in ML to model human behavior, it has been disconnected from language and, sometimes, real-world data. For instance, Agent-based modeling (ABMs), a popular paradigm in Reinforcement Learning, has been employed to model behavior (Bankes, 2002; Romero *et al.*, 2023; Park *et al.*, 2023). Nevertheless, ABMs tend to view humans as rational economic agents who communicate primarily through their actions, neglecting the significance of content in communication. In ABMs, agents strive to maximize their rewards, whereas communication does not always aim to optimize specific, well-defined reward signals. Moreover, the scarcity of large repositories containing extensive records of human actions poses a challenge when training ABMs to learn human behavior. Consequently, existing large models trained on human behavior, such as the ABMs and decision transformer and its variants, often rely on simulated data, such as game environments, rather than real human behavior (Chen *et al.*, 2021). This reliance on artificially generated data introduces biases inherent to the creators of the training data, making it difficult to capture authentic human behavior. However, recent advancements have demonstrated the potential of large models trained on real-world tokens encompassing various modalities, like images, videos, audio, and text, as the basis for diverse tasks (Ge *et al.*, 2023; Li *et al.*, 2023*a*). Notably, LLMs, as exemplars of foundation models, have exhibited impressive performance across a range of tasks, including those they were not explicitly trained for, such as emotion recognition, named entity recognition, and complex tasks like table understanding (Ye *et al.*, 2023; Bhattacharyya *et al.*, 2023).

Further, there has also been much work in modeling behavior using conventional modeling techniques, such as regression, bagging and boosting (Mazloom *et al.*, 2016; Villarroel Ordenes *et al.*, 2019), neural networks (Ding *et al.*, 2019; Wang *et al.*, 2018*b*; Khosla *et al.*, 2014), and transformers (Wu and Krahenbuhl, 2021; Xiao *et al.*, 2022). While these models can certainly model behavior, LLMs show generalization powers which extend to capabilities much beyond just behavior simulation. For instance, once trained on behavior tokens, other than behavior simulation, LLMs can now generate behavior optimized content (Table 3.5), explain behavior (Table 3.6), and domain-adapt to other behaviors (Table 3.9), none

of which are shown by other models. The other concurrent works which model behavior using LLMs (Kang *et al.*, 2023) model just behavior (for example, by CTR prediction) by attaching classification or regression heads to LLMs and thereby lose out on the text-to-text paradigm where LLMs show their best performance and generalization capabilities. In addition, similar to non LLM paradigm, this method loses out on other capabilities like generating behavior optimized content and explaining behavior.

3.2 Setup

In this section, we introduce our approach to model content and behavior together as a text-to-text problem. Since most publicly available corpora strip off receiver behavior from content, we first introduce our dataset, "The Content Behavior Corpus (CBC)", a dataset consisting of content and the corresponding receiver behavior. Next, we introduce our methodology to convert the content and behavior into text and our approach to model it using an LLM. Then, we cover the tasks through which we test various capabilities of LCBM (Fig. 3.2): content-understanding, behavior understanding, content simulation, behavior simulation, and behavior domain adaptation.

3.2.1 The Content Behavior Corpus (CBC)

The availability of large-scale unlabeled text data for unsupervised learning has fueled much of the progress of LLMs. In this paper, we are interested in modeling content and the corresponding receiver behavior in the same space. While available datasets contain trillions of content tokens (text, images, audio, and videos), they unfortunately do not contain the receiver effect. To address this, we utilize YouTube and Twitter, two large publicly available sources of content-behavior data, consisting of (a) account name, account description, and number of subscribers and followers (*communicator data*) , (b) rich content in the form of videos, images, creator-provided captions, titles, and descriptions (*message*), (c) behavior in the form of likes, views, user comments, and replay graph (*receiver effect*). This covers all the five factors of communication (Fig. 3.4), with the

channel being fixed (as YouTube or Twitter) and receivers being average channel followers and viewers of the communicator. Since content data is multimodal in the form of a combination of images, videos, and text, and behavior data is in the form of numbers, to model it using a text-to-text paradigm, we *verbalize* both of them following the methodology we detail next.

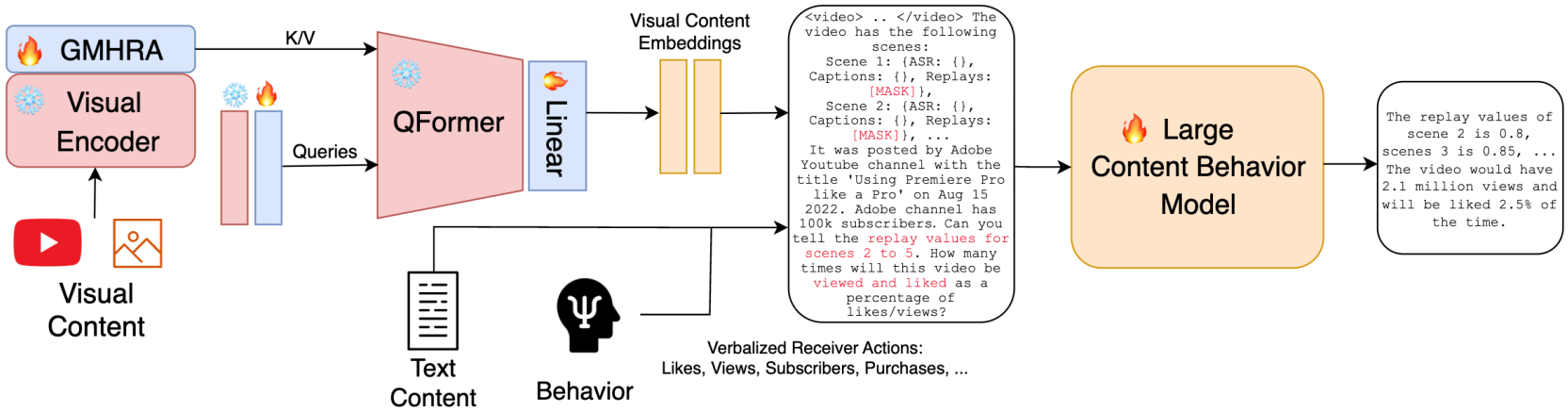


Figure 3.7: Encoding and predicting content (images, videos, and text) and behavior in the language space. Strategy to behavior instruction fine-tune (BFT) LLMs to create LCBMs. We capture visual concepts through the visual encoder (EVA-CLIP), and world knowledge is through an LLM (Llama). To leverage the rich knowledge of LLMs, we use GMHRA and QFormer to convert visual tokens of ViT to language tokens that Llama can understand. Further, we find that verbalizing the visual stimulus helps Llama to gather information more explicitly than what is provided by ViT+QFormer. We fine-tune the combined model end-to-end to predict 1) behavior given content and 2) content given behavior. Snowflake and fire symbols denote the frozen and unfrozen parts of the architecture.

Verbalization: For the video V, YouTube provides us with 100 average viewer retention values r_i for $i \in [0..100)$, corresponding to the entire video. The sampling rate of 100 is constant and independent of video length (T). Replay value r_i corresponds to video frames between the timestamps $(T/100 \times i, T/100 \times (i+1))$, which denotes how often these frames were replayed compared to the most replayed frames. The metric has a value between 0 and 1 that identifies the video's relative retention performance at a given point in the video. To accommodate longer video lengths, we merge replay values until $T/100 \times (i+j) - T/100 \times i >$ 1 second with $j \in \{i+1, 100\}$. We choose the replay value for this merged group of scenes as $max(r_i, ..., r_j)$. Using this logic, we get replay values R_i for $i \in [0..m]$, where $m = \lfloor 100/(\lceil 100/T \rceil) \rfloor$. Next, we sample two frames randomly corresponding to each $i \in [0..m]$. We caption the frames using BLIP (Li *et al.*, 2023*a*). We also obtain the automatic speech recognition for the speech for the video between the timestamps corresponding to replay value R_i using Whisper (Radford *et al.*, 2023). The ASR and BLIP captions are content for scenes, and replay values are the behavior corresponding to them. We include the scene content and behavior in the video verbalization (Listing 3.1) with the sampling for both scene content and behavior as described above.

We also include video content by encoding video frames through EVA-CLIP (Sun *et al.*, 2023) (explained in §3.2.2). Other than video embeddings, we include the video title and description as part of the video content. Corresponding to the overall video content, we verbalize overall video behavior metrics like video views and the ratio of likes and views. Finally, we append it with communicator information on the video channel and the subscriber count. The Listing 3.1 presents the overall verbalization for video and frame level content and behavior. The verbalization for Twitter posts is similar and is given in Listing 3.7.

Listing 3.1: Verbalization pattern for inputting content and behavior in the same space

```
Input: <video> ..[Video Tokens] .. </video>
The video has the following scenes:
Scene 1: {ASR: Welcome to a quick tutorial, OCR: Adobe Premiere Pro, Captions: A desktop interface,
    Replays: 60},
Scene 2: {ASR: on using Premiere Pro to edit, Captions: A computer interface, with an image of a white horse.
    Objects – Horse, Grass, Fence., Replays: 53},
...
It was posted on Adobe's YouTube channel with the title 'Using Premiere Pro like a Pro' on Aug 15 2022.
```

Adobe's YouTube channel has 100k subscribers. This video was viewed by 346 thousand people and liked (as a percentage of likes/views) by 2.3% people.

3.2.2 Model

To understand both visual and textual contents, we follow a similar approach as was taken by recent models like BLIP, Llava, VideoLlama, and others (Liu *et al.*, 2023*a*; Ge *et al.*, 2023; Li *et al.*, 2023*a*; Zhu *et al.*, 2023), we use visual encoders to encode visual knowledge and an LLM to encode text and world knowledge. Fig. 3.7 shows our architecture to encode visual content into the language space. We include video content by encoding video frames through EVA-CLIP (Sun *et al.*, 2023) and Global Multi-Head Relation Aggregator (GMHRA) from Uniformer (Li *et al.*, 2021). GMHRA helps aggregate the information better across the time dimension. The combination of ViT and GMHRA gives us a good representation of the visual content. Next, to effectively leverage the LLM's rich language representations, we use Q-Former from BLIP-2 (Li *et al.*, 2023*a*) with an extra linear layer and additional query tokens to convert from visual tokens to language tokens. Further, similar to Bhattacharyya *et al.* (2023), we find that while encoding visual tokens is powerful, converting visual content to text adds to the downstream performance. Therefore, we include the BLIP caption for each scene along with the scene replay graph.

We use the Llama-based Vicuna-13B LLM (Touvron *et al.*, 2023; Chiang *et al.*, 2023) as our base LLM. Similar to prior works (Liu *et al.*, 2023*a*; Ge *et al.*, 2023; Li *et al.*, 2023*a*; Zhu *et al.*, 2023), we follow a two-stage training paradigm where in the first stage, we utilize the WebVid (Bain *et al.*, 2021), COCO caption (Chen *et al.*, 2015), Visual Genome (Krishna *et al.*, 2017*a*), CC3M (Sharma *et al.*, 2018), and CC12M (Changpinyo *et al.*, 2021) datasets to align the visual encoder embeddings with LLM. In the second stage, we train the model with behavior instructions prepared by following the approach described in §3.2.4. In summary, LCBM takes concatenated inputs of visual tokens, scene ASR, caption, scene behavior of replays, channel information, and video title and behavior metrics of views and a ratio of likes to views. Based on the instruction, we test LCBM's abilities on various tasks we cover in the next paragraphs.

Computational Requirements: We employed a distributed training setup using 32 NVIDIA A100 GPUs (80GB each), providing a total of 2.56TB of GPU memory across the cluster. The distributed training was implemented using PyTorch's DistributedDataParallel (DDP) framework with gradient synchronization across all GPUs to ensure consistent model updates. We used gradient accumulation with micro-batch sizes of 4 samples per GPU, accumulating gradients over 8 steps before performing parameter updates, effectively achieving a global batch size of 1,024 samples. Additionally, we employed gradient checkpointing to trade computation for memory during backpropagation through the vision encoder components. The two-stage training paradigm required approximately 180 hours of total compute time per run. We used adaptive learning rate scheduling with warmup periods and cosine annealing to ensure stable convergence across the large parameter space of the 13B model. Data loading and preprocessing were optimized through asynchronous data pipelines with prefetching, allowing continuous GPU utilization while handling the multimodal nature of our datasets. Video frame extraction and encoding were performed on-the-fly during training to minimize storage requirements, while maintaining training throughput through parallelized preprocessing workers. This infrastructure enabled us to efficiently process the behavioral tokens alongside traditional content tokens, making the behavior instruction fine-tuning paradigm computationally feasible at scale.

Technical Training Pipeline Description: The training pipeline employs a similar optimization setup as LLaVA (Liu *et al.*, 2023*a*) using AdamW optimizer with a 2e-5 learning rate and cosine annealing schedule with 3% warmup ratio. The system utilizes DeepSpeed ZeRO Stage 2 optimization with CPU offloading for memory efficiency, enabling distributed training across multiple GPUs. Training uses micro-batches of 4 samples per device with gradient accumulation disabled (steps=1), while bfloat16 mixed precision and TensorFloat-32 (TF32) acceleration optimize computational efficiency. Weight decay is disabled (0.0), but the pipeline incorporates gradient checkpointing for memory conservation and LoRA (Low-Rank Adaptation) fine-tuning with 5% dropout for parameter-efficient training. The custom LLaVATrainer implements sophisticated parameter grouping, applying different weight decay policies to layer normalization and bias parameters versus other weights, with optional learning rate multipliers for specific parameter

groups. Additional regularization comes through quantization support (4-bit/8-bit with double quantization) and length-grouped sampling that batches samples of similar sequence lengths to improve training stability and efficiency.

Model	#Params	Training	Past		Future		Random Window Size				All Masked	
							5		7			
			RMSE	Accuracy	RMSE	Accuracy	RMSE	Accuracy	RMSE	Accuracy	RMSE	Accuracy
LCBM	13B	3-BFT	8.12	55.10	15.05	42.42	8.55	61.41	9.91	55.10	-	-
LCBM		5-BFT	11.53	52.06	12.02	53.06	8.13	64.83	9.22	60.26	31.34	17.16
LCBM		7-BFT	16.17	35.61	15.14	44.11	9.02	59.22	10.47	53.84	-	-
LCBM		11-BFT	18.25	30.95	15.05	41.44	10.01	55.15	10.49	52.61	-	-
GPT-4	>100B[6]	10-shot-ICL	34.45	20.55	19.51	36.08	22.99	26.99	27.25	17.27	38.52	14.26
GPT-4		2-shot-ICL	35.05	19.34	18.07	39.33	17.42	38.10	21.26	28.05	37.60	13.73
GPT-3.5	175B	3-shot-ICL	34.10	19.06	24.71	27.14	24.52	24.81	26.30	18.74	38.77	13.47
GPT-3.5		2-shot-ICL	33.36	18.02	26.44	25.42	23.35	25.35	24.68	21.24	37.16	13.39
Random	-	-	34.10	10.00	34.10	10.00	34.10	10.00	34.10	10.00	34.10	10.00

Table 3.1: **Behavior Simulation.** Mean RMSE and accuracy scores for scene-by-scene predictions of video replay values. Replay values are the normalized replay scores of each scene as provided by YouTube. The normalized scores are considered to 2 decimal places and multiplied by hundred to convert the score to an integer score in the range 0-100. RMSE is calculated for each video in the test set and the mean is calculated for this score and reported. The model is said to classify correctly if the absolute error between the predicted and ground truth value is less than or equal to 5. The scores are calculated in four regimes: past, future, random, and all-masked. In the past (future) regimes, first (last) 5-20% scenes are masked; in the random setting, 5-20% scenes are masked randomly, and in all masked setting, everything is masked. LCBM was behavior-fine-tuned (BFT) with 3,5,7,11 context window masking strategy, while GPT was compared with an in-context learning (ICL) setting. We note that behavior fine-tuned LCBM, while being at least 10x smaller than other models, performs the best. Best models are denoted in green and runner-ups in blue.

3.2.3 Content Behavior Test Benchmark

We test the capabilities of large content-behavior models on predictive and descriptive abilities on content and behavior, as illustrated in Fig: 3.2. We design the following five tasks to test these capabilities: behavior simulation, content simulation, content understanding, behavior understanding, and behavior domain adaptation. We cover each of these tasks next.

1. **Behavior Simulation.** We test simulation capability on four behaviors across two datasets: YouTube replay values, the ratio of YouTube likes to views, Twitter likes, and the number of views of the YouTube video. The common task amongst all of them is to predict the behavior given the content and content attributes like captions, scene-by-scene descriptions for videos, and sender characteristics like account and subscriber count and date of posting. The behavior to be predicted is masked and asked as a question to the LLM. Listings 3.6 and 3.7 lists the verbalization pattern for this task. For replay value prediction, we test the masked behavior in three settings: *Masked Past* (all replay values of the first 5-20% scenes are masked), *Masked Future* (all replay values of last 5-20% scenes are masked), and *Random Masks* (random masking of replay values for 5-20% scenes). **Metrics:** For replay value prediction (Table 3.1), we use RMSE and accuracy, where a prediction is considered correct if the absolute error between predicted and ground truth is ≤ 5. For Twitter like prediction (Table 3.3), we use classification accuracy in a binary high/low prediction task. For like/view ratio prediction (Table 3.7), we use RMSE, R^2 score, and accuracy (correct if error 10% of ground truth).

2. **Content Simulation.** Here, the task is to predict content given receiver behavior (Listing 3.5, 3.4). For YouTube, given the video content in terms of scene-by-scene descriptions with the content of one group of five consecutive scenes content being masked, behavior values of all scenes, and channel information, the task is to choose the masked scene speech from a list of 25 options, chosen randomly from the entire test set. For YouTube, we chose to model this task as a discriminative task instead of a generative one since videos are generally long, and there could be multiple possible contents for a given behavior, whereas the ground truth is available only for one specific characterization of the content for a given behavior. For Twitter, we model this task as content generation. The Listing 3.8 presents the format for this task. **Metrics:** For Twitter content simulation (Table 3.4), we use BLEU-1/2/3/4 and ROUGE-L to evaluate the quality of generated tweet text. For YouTube content simulation (Table 3.5), due to the open-endedness of videos, we turn content prediction to a classification task, where the task is to select speech segments out of a given number of choices; we use accuracy as the performance metric (percentage of correctly selected speech segments).

3. **Behavior Understanding.** The goal of this task is to check if the model can reason about observed or unobserved receiver behavior. For this task, we could ask the model to explain any behaviors given the content. However, only the

§Note that we cannot compare this model with GPT-3 due to the private nature of data.

YouTube receiver comments have ground truth available with the video. Without ground truth, we found that other behaviors, such as replay values, likes, and views, are difficult to explain by non-experts. Therefore, we ask the model to simulate the sentiment of the receivers' comments and describe its reasoning. To evaluate, we asked six annotators to annotate the reasons provided by the model on a scale of 0-5, with 0 implying the LLMs provided no sentiment or reasoning and 5 implying perfect reasoning. The annotators were free to rate the LLMs as they seemed fit. The annotators were asked to review the video content and the comments to help them evaluate the reasons. We average the ratings of three annotators to get an average rating for every video. Similarly, to review the sentiment correctness, we asked the annotators to judge the predicted sentiment rating with respect to user comments. **Metrics:** We use sentiment accuracy (percentage of correct sentiment predictions compared to ground truth) (Table 3.6) and reasoning score (average human-assigned rating from 0-5).

4. **Content Understanding.** To check if a model trained on both content and behavior tokens does not forget its original content understanding capabilities, we test the content understanding tasks on YouTube videos, following Bhattacharyya *et al.* (2023). They use the following tasks for video-understanding: topic, emotion, persuasion, and action-reason classification. For topic, emotion, and action-reason classification tasks, they use the advertisements dataset by Hussain *et al.* (2017), which contains 3,477 video advertisements and the corresponding annotations for emotion and topic tags and action-reason statements for each video. There are a total of 38 topics and 30 unique emotion tags per video. Further, we have 5 action-reason statements for each video for the action-reason generation task. For our experiment, we use the subset of 1,785 public videos. Following Bhattacharyya *et al.* (2023), for the topic and emotion classification task, we evaluate our pipeline using top-1 accuracy as the evaluation metric. Further, we evaluate emotion classification on clubbed emotion labels as well. For action and reason prediction, we evaluate our accuracy on the action and reason retrieval tasks where 29 random options along with 1 ground truth are provided to the model to find which one is the ground truth. In the persuasion strategy classification, we use the 1002 persuasion strategy videos and corresponding labels released by Bhattacharyya *et al.* (2023). Given the video, the model has to predict which persuasion strategy the video conveys. Persuasion strategy classification could be an important task for evaluating LCBM since the concept of persuasion in psychology views human communication as the means to change the receiver's beliefs and actions (*i.e.*, to persuade) (Kumar *et al.*, 2023*a*), and understanding the different strategies present in communication may help understand human behavior better. **Metrics:** Following previous works (Bhattacharyya *et al.*, 2023; Kumar *et al.*, 2023*a*), we evaluate using top-1 accuracy across all content understanding tasks (topic, emotion, persuasion strategy, action, and reason).

5. **Behavior Domain Adaptation.** In the past work, we have observed strong generalization capabilities from LLMs (OpenAI, 2023; Ouyang *et al.*, 2022; Raffel *et al.*, 2020). While training on next token prediction, LLMs show generalization across tasks, including question answering, natural language inference, and sentiment analysis. Given this, the natural question is, does LCBM, too, show this kind of generalization, where a model trained on one kind of

Date Range	April 1, 2022 to June 12, 2023
Number of Countries	225
Target Products	Top Products used by millions of users
Customers Segmented on the basis of	Type of use, user expertise, frequency of use, and others

Table 3.2: Details of the in-house Marketing Email dataset used to evaluate behavior generalization capabilities of the LCBM

behavior, can show performance on another behavior? To understand this, we test the model on a different dataset and task than what it was originally trained for. We do this over three datasets, LVU (Wu and Krahenbuhl, 2021), in-house Email Marketing[‡], and generalization between Twitter and YouTube likes.

- **LVU Benchmark.** Wu and Krahenbuhl (2021) released a benchmark for long video understanding with over 1000 hours of video. In the benchmark, they have two behavior related tasks: ratio of likes to likes+dislikes and view prediction. YouTube has discontinued the dislike count, therefore, our corpus does not contain the dislike count. We use the LVU test benchmark to check if a model trained on other available behaviors (views, likes, and replay graphs) is able to predict the like ratio. **Metrics:** Following previous works (Bhattacharyya *et al.*, 2023) (Table 3.9), we use MSE (Mean Squared Error) for evaluating performance on the LVU benchmark.
- **in-house Email Marketing.** In this task, we ask the model to predict the click-through rate for a given target segment of an email, given the email content, subject, and verbalized descriptions of the images in the email. We use the emails sent by in-house marketing team to its subscribers. The emails were sent from April 1, 2022 to June 12, 2023 and covered many of the premiere products. The emails were sent to many customer segments (as defined by the marketing team) across 225 countries (Table 3.2). Listing 3.3 lists the verbalization format to verbalize emails to input to the LCBM. **Metrics:** We use RMSE and R^2 score to evaluate click-through rate prediction performance (Table 3.9).

3.2.4 Behavior Instruction Fine-Tuning (BFT)

To teach an LLM the behavior modality over multimodal content, we convert both the visual tokens and behavior modality in the text format and instruction fine-tune the LLM end to end. This follows a two-stage approach: first, we teach the LLM the visual modality (§3.2.2), and next, we teach the LLM the behavior

[‡]Brand Separated means that the train and test set don't have any overlap in terms of brands, Time Separated means that the test set starts after the last tweet in the train set. BFT denotes behavior fine-tuning, and ICL stands for in-context learning. The best results over four runs are reported for all models. Best models are denoted in green and runner-ups in blue.

Model	#Params	Training type	Training	Time Separated	Brand Separated
GPT-3.5	175B	ICL	Few-shot	58.84	64.19
LCBM	13B	BFT	Twitter	74.3	97.69
LCBM	13B	BFT	Twitter and YouTube data	76.87	92.19

Table 3.3: **Behavior Simulation and Behavior Domain Adaptation**[‡]. Two-way classification accuracies for like prediction on Twitter. Given content, channel, and time, predict behavior (High, Low). We note that LCBM trained on Twitter and YouTube performs better than the one trained only on Twitter, showing signs of performance improvement by domain adaptation.

Model	Training	Test	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-l
GPT-3.5	ICL	Brand Separated	53.95	42.36	31.84	24.28	15.24
		Time Separated	57.69	45.11	33.67	25.52	15.27
LCBM	BFT on Twitter	Brand Separated	62.29	46.59	33.98	25.64	14.44
		Time Separated	70	54.4	41.43	32.48	17.38
LCBM	BFT on Twitter + Youtube	Brand Separated	64.28	48.1	35.17	26.63	14.83
		Time Separated	70.23	54.54	41.52	32.54	17.45

Table 3.4: **Content Simulation and Behavior Domain Adaptation**[‡]. Given behavior, channel, time, tweet media caption as prompt, predict content (tweet text). We note that LCBM trained on Twitter and YouTube performs better than the one trained only on Twitter, showing signs of performance improvement by domain adaptation.

modality. We call the latter "Behavior Instruction Fine-Tuning (BFT)" inspired by instruction fine-tuning (IFT) and its variants like visual instruction tuning (Liu *et al.*, 2023*a*).

We prepare the content-behavior instruction datasets as explained next.

Teaching behavior in the forward direction (predict behavior given content): In this instruction tuning task, we teach the model to predict behavior given the message sent by the communicator. Essentially, this teaches the model to predict behavior in the forward direction (as in Fig. 3.4). Concretely, we include the following information as part of verbalization - image and video embedding converted to the text space (using EvaCLiP (Sun *et al.*, 2023)), scene-by-scene verbalization covering automatic speech recognition, scene captions, video/post caption and description, receiver behavior covering replay rates, views, and likes, and communicator information covering account name and follower count. The verbalisation pattern for this task is the same as given in the Listing 3.6.

Teaching behavior in the reverse direction (predict content given behavior): This task teaches the model to learn about behavior in the reverse direction (Fig. 3.4). Here, the model learns to simulate content given behavior. The instruction for this task is given in Listing 3.4.

Using the prepared content and behavior instruction datasets consisting of pairs of content and behavior tokens, we treat the content tokens ($\mathbf{X}_C$) as input and behavior tokens ($\mathbf{X}_B, x_i \in \mathbf{X}_B$) as output of the language model. We then perform instruction-tuning of the LLM on the prediction tokens, using its original auto-regressive training objective. Specifically, for a sequence of length L, we compute the probability of generating target answers ($\mathbf{X}_B$) by:

$$p(\mathbf{X}_B|\mathbf{X}_C) = \prod_{i=1}^{L} p_\theta(x_i|\mathbf{X}_C, \mathbf{X}_{B,<i}) \tag{3.1}$$

For the behavior instruction tuning, we keep the visual encoder weights frozen and continue to update the pre-trained weights of the LLM in LCBM.

Model	#Params	Accuracy
Vicuna	13B	19.30%
LCBM	13B	48.68%
GPT-3.5	175B	34.98%
Random	-	4%

Table 3.5: **Content Simulation.** In this task, the models have to choose the speech segment from a list of 25 options given the video description, non-masked scenes. and replay behavior. We see that despite being similar to masked language modeling (which is a content-only task), LCBM performs better than both Vicuna and GPT-3.5. Best models are denoted in green and runner-ups in blue.

Model	#Params	Sentiment Accuracy	Reasoning Score
Vicuna	13B	65.66%	2.23
LCBM	13B	72.73%	4.00
GPT-3.5	175B	61.62%	1.67

Table 3.6: **Behavior Understanding.** In this task, the models have to simulate the sentiment of comments that a video would get by looking at only the video. Further, they also have to explain the reason for such sentiment. The responses were annotated by humans on a scale of 0-5 for the reason, with 0 being no response provided and 5 being the response matches exactly with the (ground truth) comments received on the video. Best models are denoted in green and runner-ups in blue.

3.3 Results and Discussion

Here, we discuss the results for the five tasks we discuss in Section 3.2.3, namely, behavior simulation, content simulation, behavior understanding, content understanding, and behavior domain adaptation. We compare the behavior fine-tuned model discussed in §3.2.4 with state-of-the-art content-only models like GPT-3.5, GPT-4, and Vicuna-13B. This allows us to compare how much including behavior tokens in the training of an LLM helps in improving the LLM's understanding of behavior and joint content and behavior spaces while retaining its understanding of the content space.

The results for the five tasks are presented in Tables 3.1,3.7,3.5,3.6,3.8, 3.9,3.3, and 3.4. We note a few general trends. LCBM, while being 10x smaller than GPT-3.5 and 4, performs better than them on all behavior-related tasks. Further, we see that there is no significant difference between 10-shot and 2-shot GPT-4 or between GPT-3.5 and GPT-4, indicating that unlike other tasks, it is harder to achieve good performance through in-context-learning on the behavior modality. It can be observed that often GPT-3.5 and 4 achieve performance comparable to

(or worse than) random baselines. Interestingly, the performance of GPTs on the content simulation task is also substantially behind LCBM. The way we formulate the content simulation task (Listing 3.5), it can be seen that a substantial performance could be achieved by strong content knowledge, and behavior brings in little variance. We still see a substantial performance gap between the two models. All of this indicates that large models like GPT-3.5 and 4 are not trained on behavior tokens.

For the content understanding tasks (Table 3.8), predictably GPT-3.5, being the largest model, achieves the best results. However, we see that BFT helps the LLM to learn most content understanding tasks better than the base LLM. LCBM gets better results than both Vicuna and VideoChat. This indicates that behavior modality might carry additional information about the content, which might help an LLM understand content better (Khurana *et al.*, 2023; Klerke *et al.*, 2016*a*; Plank, 2016*a*). Next, we see that LCBM also shows signs of domain adaptation in the behavior modality. We see that on five tasks: comment sentiment prediction, comment sentiment reasoning (Table 3.6), email behavior simulation (Table 3.9), and Twitter behavior (Table 3.3) and content simulation (Table 3.4). We note that if the LCBM is trained on only email behavior simulation samples, it underperforms the model trained on both YouTube data and a few samples to make the model learn email format. Similarly, LCBM trained on both Twitter and YouTube performs better than the one just trained on Twitter, showing performance improvement by domain adaptation. Finally, Figs. 3.6,3.5 show a few samples where we query LCBM to explain replay and comment behavior and compare it with human explanations. We see that LCBM while verbose, can explain behavior well.

3.3.1 Verbalization Listings

Listing 3.2: Verbalization pattern of videos for the behavior understanding task:

```
Input: <video> .. </video>
The video has the following scenes:
Scene 1: {ASR: Welcome to a quick tutorial, OCR: Adobe Premiere Pro, Captions: A desktop interface,
    Replays: 60},
```

[†]The exact size of GPT-4 is unknown.

Model	#Params	Training type	Training	RMSE	R^2	Accuracy
LCBM	13B	BFT	Replay values 3-masked	1.31	0.87	15.89
LCBM		BFT	Replay values 5-masked	1.48	0.82	19.93
LCBM		BFT	Replay values 7-masked	1.71	0.78	15.20
LCBM		BFT	Replay values 11-masked	1.55	0.82	13.94
GPT-4	>100B[‡]	ICL	10-shot	3.50	-0.01	7.84
GPT-4		ICL	2-shot	3.58	-0.03	5.39
GPT-3.5	175B	ICL	3-shot	64.40	-256.96	2.48
GPT-3.5		ICL	2-shot	64.88	-375.83	1.27
Random	-	-	-	4.67	0	3.94

Table 3.7: **Behavior Simulation.** RMSE, R^2, and accuracy scores for like/view ratio prediction task. To calculate accuracy, the model is said to classify correctly if the absolute error between the predicted and ground truth likes/views is less than or equal to 10%. BFT denotes behavior fine-tuning, and ICL stands for in-context learning. Replay values k-masked means a model which is trained by masking k consecutive values of the replay graph while doing BFT. We note that LCBM while being at least 10x smaller than the other models, performs the best. The best results over four runs are reported for all models. Best models are denoted in green and runner-ups in blue.

Training	Model	#Params	Topic	Emotion		Persuasion	Action	Reason
			All labels	Clubbed				
Random	**Random**	-	2.63	3.37	14.3	8.37	3.34	3.34
Zero-shot	**GPT-3.5**	175B	51.6	11.68	79.69	35.02	66.27	59.59
	Vicuna	13B	11.75	10.5	68.13	26.59	20.72	21.00
	VideoChat (Li *et al.*, 2023*c*)	13B	9.07	3.09	5.1	10.28	-	-
	LCBM	13B	42.17	7.08	58.83	32.83	39.55	27.91

Table 3.8: **Content Understanding.** Comparison of several models, including behavior instruction tuned models before and after BFT. We compare the models across topic, emotion, and persuasion strategy detection tasks as per the framework given by Bhattacharyya *et al.* (2023). We see that our model outperforms similarly sized models (Vicuna, VideoChat) in most tasks. Best models are denoted in green and runner-ups in blue.

in-house Email Marketing							
LCBM Type	**Fine-tuned on YouTube?**	**Trained On**			**Tested On**	**RMSE**	**R^2**
		Unique Emails	**Unique Segments**	**Email-Segment Pairs**			
Domain-Adapted	Yes	100	10	1k	Different Segment (emails could be same)	14.47	0.64
In-Domain	No	600	560k	350k		25.28	0.55
Domain-Adapted	Yes	100	10	1k	Different Segments & Different Emails	27.28	0.54
In-Domain	No	600	560k	350k		29.28	0.5

LVU Benchmark			
Training	**Model**	**Testing**	**MSE**
Trained	R101-slowfast+NL (Wu and Krahenbuhl, 2021)	Test set	0.386
Trained	VideoBERT (Sun *et al.*, 2019*a*)	Test set	0.32
Trained	Qian *et al.* (2021)	Test set	0.353
Trained	Xiao *et al.* (2022)	Test set	0.444
Trained	Object Transformers (Wu and Krahenbuhl, 2021)	Test set	0.23
Zero-shot	LCBM (Ours)	Test set	0.14
Zero-shot	GPT-3.5	Test set	0.03
Zero-shot	Vicuna	Complete dataset	0.44
Zero-shot	LCBM (Ours)	Complete dataset	0.30
Zero-shot	GPT-3.5	Complete dataset	0.02

Table 3.9: **Behavior Domain Adaptation.** We test the generalization capability of LCBM on two tasks: (1) Behavior simulation on in-house Email Marketing Data, (2) Behavior simulation on the LVU benchmark. For (1), we train two versions of LCBM with the in-house Email Marketing data: one was trained on YouTube videos and further BFT on a few email samples (*domain-adapted*), and the other was BFT on a larger set of emails, but not including YouTube data (*in-domain*)[§]. We report the RMSE and R^2 scores for this task. For (2), we compare LCBM with other state-of-the-art results and GPT-3. In (1), we note that the domain-adapted LCBM performs better than the in-domain LCBM in both settings. We posit that YouTube data helps LCBM understand how a company's viewers like to hear from it, giving LCBM an edge over a model trained on a small amount of the same data (600 unique emails). In (2), LCBM performs better than the existing state-of-the-art. Surprisingly, GPT-3.5 does better than LCBM on this task. From both (1) and (2), we gather that a model trained on certain YouTube behaviors performs better on other behaviors, thus showing promise of domain-adaptation in the behavior modality. Best models are denoted in green and runner-ups in blue.

Scene 2: {ASR: on using Premiere Pro to edit, Captions: A computer interface, with an image of a white horse. Objects − Horse, Grass, Fence., Replays: 53},

...

It was posted on Adobe's YouTube channel with the title 'Using Premiere Pro like a Pro' on Aug 15 2022. Adobe's YouTube channel has 100k subscribers. This video was viewed by 346 thousand people and liked (as a percentage of likes/views) by 2.3% people. Why is the scene 23 one of the most replayed scenes?

Output: The scene shows the transformation of the image after the changes.

Listing 3.3: Verbalization pattern of emails for the behavior domain adapation task. The email content and CTR is for demonstration purposes only.

Input: Email with Subject: Lock it down before you send it out.

Header: Nobody messes with your PDFs.

Body text: Add password protection, secure encryption, and restricted editing to your PDFs with Adobe Acrobat Pro DC. Share only what you want and nothing more. A button that says 'Get started'. An image of a laptop, with window open on it. Image text: "Protect using password".

Foreground colors: grey, blue. Background colors: lavender, white. Image Emotions: security, serious. Image keywords: laptop, protect, password, lock. Aesthetic value: low. Clutter level: medium. The email is created by a Creative Professional, for the product Adobe Acrobat Pro. It is sent to users in the United States, in the commercial market. Specifically, it is sent to Power users with the intent of Active Use.

The email was sent 109 times between 25 August, 2022 and 26 August, 2022, and had a click through rate of [MASK]%.

Output: 0.037%.

Listing 3.4: Verbalization pattern to teach behavior in the reverse direction (predicting content given behavior):

Input: <video> .. </video> The video has the following scenes: Scene 1: {ASR: [MASK], Replays: 60%}, Scene 2: {ASR: with Premiere, Captions: Woman looking at screen, Replays: 34%},

...

Scene 5: {ASR: has never been, Captions: Colour Pallete, Replays: 47%},

Scene 6: {ASR: been easier, Captions: Colour Pallete, Replays: 54%},

...

It was posted on Adobe's YouTube channel with the title 'Using Premiere Pro like a Pro' on Aug 15 2022. It is viewed 203k times and liked 1.2%. Adobe's YouTube channel has 100k subscribers. Predict the masked ASR value for the masked scenes.

Output: Scene 1:{ASR: Welcome to a quick tutorial.}

Listing 3.5: Verbalization pattern of videos for the content simulation task:

Input: <video> .. </video> The video has the following scenes: Scene 1: {ASR: [MASK], Replays: 60%}, Scene 2: {ASR: with Premiere, Captions: Woman looking at screen, Replays: 34%},

...

Scene 5: {ASR: has never been, Captions: Colour Pallete, Replays: 47%},

Scene 6: {ASR: been easier, Captions: Colour Pallete, Replays: 54%},

...

It was posted on Adobe's YouTube channel with the title 'Using Premiere Pro like a Pro' on Aug 15 2022. It is viewed 203k times and liked 1.2%. Adobe's YouTube channel has 100k subscribers. Predict the masked ASR value for scene 1. Choose from the given options.

Option−1: Welcome to a quick tutorial,
Option−2: Samsung Galaxy A20 smartphone,
...
Option−25: regulations. We haven't had.

Listing 3.6: Verbalization pattern of videos for the behavior simulation task:

Input: <video> .. </video> The video has the following scenes:
Scene 1: {ASR: Welcome to a quick tutorial, OCR: Adobe Premiere Pro, Captions: A desktop interface, Replays: [MASK]},
Scene 2: {ASR: on using Premiere Pro to edit, Captions: A computer interface, with an image of a white horse. Objects − Horse, Grass, Fence., Replays: [MASK] }, ...
It was posted on Adobe's YouTube channel with the title 'Using Premiere Pro like a Pro' on Aug 15 2022. Adobe's YouTube channel has 100k subscribers. Can you tell the replay values for scenes 2 to 5. How many times will this video be viewed and liked as a percentage of likes/views?

Output: Scene 1: {Replay: 60%}, Scene 2: {Replay: 85%}, ..., Views: 2.1 Million, Likes−per−View: 2.5%

Listing 3.7: Verbalization pattern of Twitter posts for the behavior simulation task:

Input: Given a tweet of pfizer posted by the account PfizerMed on 2023−01−12. Tweet : Announcing a new ASGCT−Pfizer grant to support independent medical education initiatives on genetic medicines. For details, click Request for Proposals. <hyperlink>. Apply by January 30, 2022 #raredisease #ASGCT # GeneTherapy <hyperlink>. Verbalisation of media content: \"caption\": \"A close−up of a DNA double helix, showcasing its structure and blue color\",\"keywords\": \"DNA, double helix, structure, blue, close −up, molecular biology, genetics, biology, scientific illustration\"}. Predict whether it will recieve high or low likes?",

Output: This tweet has low likes.

Listing 3.8: Verbalization pattern of Twitter posts for the content simulation task:

Input: Generate a tweet given the media verbalization and the likes it got. Tweet is for pfizer to be posted by the account PfizerMed on 2023−01−12. Verbalisation of media content: \"caption\": \"A close−up of a DNA double helix, showcasing its structure and blue color\",\"keywords\": \"DNA, double helix, structure, blue, close−up, molecular biology, genetics, biology, scientific illustration\"}. This tweet has low likes."

Output: "Tweet : Announcing a new ASGCT−Pfizer grant to support independent medical education initiatives on genetic medicines. For details, click Request for Proposals. <hyperlink>. Apply by January 30, 2022 #raredisease #ASGCT #GeneTherapy <hyperlink>"}

3.3.2 Cross Domain Generalization

A key question for the practical use of Large Content Behavior Models (LCBM) is whether the improvements seen in advertising contexts can be applied to other domains like education, healthcare, and entertainment. Our experiments provide

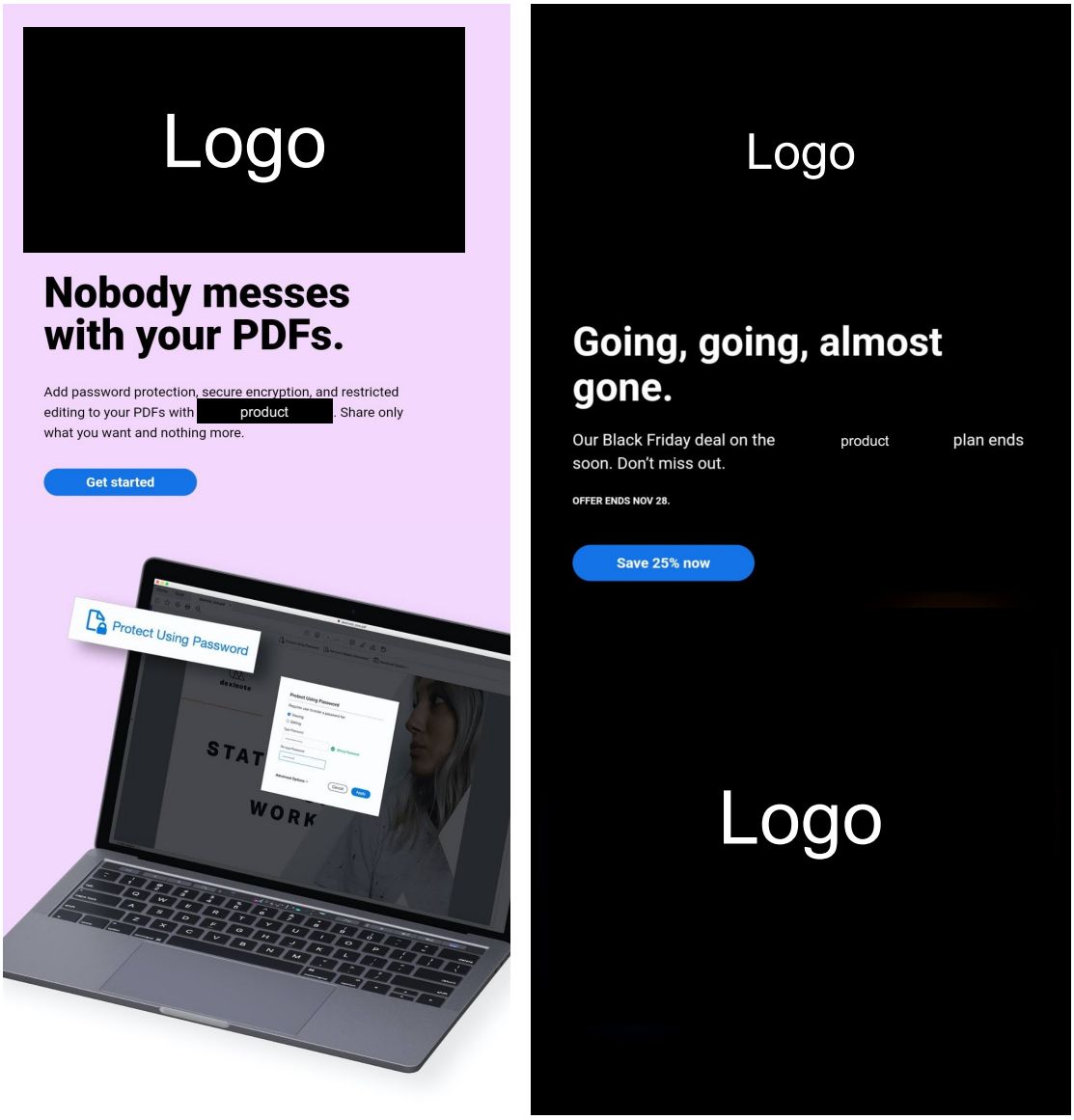


Figure 3.8: The in-house marketing emails used in the Email dataset look similar to the ones shown here.

strong evidence supporting this cross-domain generalization.

Cross-Domain Performance. The LVU benchmark results (Table 3.9) clearly demonstrate LCBM's capability, achieving a zero-shot performance with an MSE of 0.14 on videos from documentaries, lectures, sports, and entertainment—none of which are related to advertising. Additionally, LCBM adapts well to various content types and behavioral patterns, such as temporal engagement dynamics and content complexity preferences, facilitating cross-domain transfer. Our domain adaptation experiments reveal that LCBM, when trained on diverse YouTube content, performs better on email marketing tasks ($R^2 = 0.64$) compared to models trained solely on email data ($R^2 = 0.55$), indicating effective bidirectional cross-domain transfer.

Diverse Content Coverage. Our Content Behavior Corpus (CBC) dataset, comprising over 40,000 videos, includes educational tutorials, health and wellness content, entertainment, technology reviews, and general informational content. This diversity shows that LCBM can learn behavioral patterns beyond just persuasive or commercial content. We present LCBM's performance on the CBC dataset in Tables 3.1, 3.4, 3.5, 3.7, and 3.6, highlighting its ability to handle behavioral patterns that differ significantly from those in advertising content.

3.4 Failure Cases and Model Limitations

While LCBM demonstrates strong performance across behavior-related tasks, our empirical analysis reveals several systematic limitations and failure modes that merit detailed examination. Understanding these failure cases is crucial for both theoretical advancement and practical deployment of content-behavior models.

3.4.1 Scale vs. Specialization Trade-offs

The most prominent limitation of LCBM emerges in pure content understanding tasks, where scale advantages of larger models cannot be overcome by behavioral specialization. Table 3.8 demonstrates this clearly: while LCBM (42.17% topic accuracy) substantially outperforms similar-sized models like Vicuna (11.75%) and

VideoChat (9.07%), it underperforms GPT-3.5 (51.6%) despite being trained on both content and behavior tokens. This 9.43 percentage point gap indicates that *parameter scale remains a fundamental limiting factor for content understanding capabilities*.

This trade-off manifests across multiple content tasks. In emotion classification, LCBM achieves 7.08% accuracy compared to GPT-3.5's 11.68%, and in action-reason understanding, LCBM reaches 39.55% versus GPT-3.5's 66.27%. These consistent performance gaps suggest that behavior fine-tuning, while beneficial for behavior-related tasks, may introduce interference patterns that degrade pure content processing capabilities when competing against substantially larger models.

3.4.2 Behavior Domain Specificity and Transfer Limitations

Our domain adaptation experiments reveal significant limitations in cross-domain behavior transfer. Table 3.9 shows that LCBM trained exclusively on email data achieves $R^2 = 0.55$, substantially underperforming the domain-adapted model ($R^2 = 0.64$) trained on YouTube + Email data. This 16.4% relative improvement demonstrates that either behavior patterns exhibit strong domain-specific characteristics that limit naive transfer learning approaches or that models of certain (small) size need to be trained on more data to generalize better.

The failure of direct behavior transfer is particularly evident in the semantic gap between platforms. YouTube replay patterns, which capture temporal engagement dynamics in video content, do not directly transfer to email click-through behaviors, which reflect different cognitive and motivational processes. Similarly, Twitter engagement patterns (likes, retweets) operate under distinct social dynamics compared to YouTube's consumption-focused metrics. This suggests that behavior tokens encode platform-specific interaction modalities that require careful adaptation strategies rather than direct transfer. Future works should explore increasing the size of the model, training on diverse and more data to improve generalization, and also explore more sophisticated domain adaptation strategies.

3.4.3 Context Length and Complexity Limitations

LCBM exhibits performance degradation on longer content sequences and complex multimodal inputs. For videos exceeding certain length (typically 300 seconds or containing more than 50 distinct scenes), behavior prediction accuracy drops significantly. Similarly, when processing emails with extensive visual content or complex layout structures, click-through rate prediction becomes less reliable.

This limitation stems from the verbalization approach used to convert multimodal content into text tokens. Longer content requires proportionally more tokens for adequate representation, leading to context window constraints and attention dilution. The linear scaling of token requirements with content complexity creates a fundamental scalability challenge for the current architecture. Future works should explore more efficient content verbalization approaches that scale better with content complexity. Further, stronger integrations of the content and behavior modalities could help the model to better understand the content and behavior.

3.4.4 Behavioral Signal Quality Dependencies

LCBM's performance is highly sensitive to the quality and granularity of behavioral signals. For content with sparse engagement data (fewer than 100 interactions) or noisy behavioral metrics (e.g., bot-inflated engagement), prediction accuracy degrades substantially. This dependency on high-quality behavioral supervision limits the model's applicability to emerging content creators or niche domains with limited historical data. Future works should explore more robust behavioral signal collection and processing methods to improve the quality of the behavioral signals.

Furthermore, the model struggles with behavioral signals that exhibit high temporal variance or are influenced by external events (trending topics, viral phenomena). The static nature of the training data fails to capture the dynamic aspects of online engagement, leading to prediction failures during periods of unusual activity patterns. Future works should explore more dynamic behavioral modeling approaches to capture the dynamic aspects of online engagement.

Understanding these failure modes provides a foundation for developing more robust and capable content-behavior models while setting appropriate expectations for current deployment scenarios.

3.5 Conclusion

In this paper, we make initial strides towards solving the effectiveness problem proposed by Shannon in his seminal paper on communication. The effectiveness problem deals with predicting and optimizing communication to get the desired receiver behavior. This can be seen as consisting of a string of capabilities: behavior simulation, content simulation, and behavior domain adaptation. We show that while large language models have great generalization capabilities, are unable to perform well on the effectiveness problem. We posit that the reason for this could be a lack of "behavior tokens" in their training corpora. Next, we train LLMs on behavior tokens to show that other than content understanding tasks, the trained models are now able to have good performance across all the behavior-related tasks as well. We also introduce a new Content Behavior Corpus (CBC) to spur research on these large content and behavior models (LCBMs).

3.6 Ethical Considerations: Bias Mitigation and Privacy-Preserving Strategies in Large-Scale Behavioral Modeling

The integration of real-world behavioral data into large-scale content-behavior models offers unprecedented opportunities for understanding and simulating human communication. However, it also introduces critical challenges related to fairness, bias, and privacy. In this section, we detail the technical strategies adopted in this thesis to address these concerns, drawing on best practices from the machine learning literature and the specific protocols implemented in our work.

3.6.1 Privacy-Preserving Data Collection and Processing

Anonymization and PII Removal. All datasets used in this thesis were curated with strict privacy safeguards. Data collection was restricted to enterprise or business accounts, as identified via the Wikidata Knowledge Graph, to avoid capturing individual-level behavioral traces. All references to usernames (e.g., "@username") and other personally identifiable information (PII) were systematically removed. Only aggregate engagement metrics (such as total likes or views) were retained, ensuring that no individual user's actions could be reconstructed or re-identified. This approach aligns with established data minimization principles and complies with major data protection frameworks such as GDPR and CCPA.

Aggregate-Only Behavioral Signals. To further reduce privacy risks, we collect and model only aggregate behavioral signals (e.g., overall tweet popularity, average video replay rates) rather than individual-level interaction logs. This aggregation ensures that the models learn from population-level trends without exposing or memorizing sensitive user-level data.

Staged and Controlled Data Release. In line with responsible data governance, any public release of datasets or benchmarks is conducted in a staged manner. Initial access is provided within controlled environments (e.g., sandboxes or research arenas), allowing for close monitoring of usage and rapid response to emerging privacy or ethical concerns. Acceptable Use Policies explicitly prohibit applications that could lead to privacy violations or misuse of persuasive technologies.

3.6.2 Bias Mitigation and Fairness-Aware Learning

Dataset Curation and Filtering. Recognizing the risk of demographic, temporal, and algorithmic biases inherent in behavioral data, we implemented multiple filtering and balancing steps:

- *Brand and Temporal Matching:* Behavioral comparisons (e.g., tweet pairs) are restricted to the same brand or sub-brand and to narrow time windows, minimizing confounding effects from external events or platform changes.
- *Semantic and Lexical Similarity:* Pairs are selected to have high semantic and lexical similarity, reducing the risk that spurious content differences drive observed behavioral effects.

- *Manual Verification:* Subsets of the data are manually reviewed to ensure that automated filters are effective in removing confounds and maintaining fairness.

Bias Acknowledgment and Limitations. Despite rigorous filtering, we acknowledge that residual biases may persist due to demographic skews, self-selection effects, or platform-specific algorithms. These limitations are transparently discussed, and users of the models and datasets are cautioned to consider these factors when interpreting results or deploying models in sensitive contexts.

Acceptable Use and Dual-Use Mitigation. Given the dual-use potential of persuasive models, we enforce Acceptable Use Policies that prohibit deployment in high-risk domains (e.g., political campaigning, spam, or deceptive advertising). Dataset releases are accompanied by explicit guidelines and, where feasible, technical controls to prevent misuse.

3.6.3 Opportunities for Future Work

While the current work implements privacy and fairness safeguards, further advances in this direction are possible, both from safeguards and from model training perspectives. Future research directions include:

Differential Privacy. Incorporating formal differential privacy (DP) mechanisms (Dwork *et al.*, 2014) into model training pipelines would provide mathematically rigorous guarantees that individual user contributions cannot be reverse-engineered from model parameters or outputs. Techniques such as DP-SGD (differentially private stochastic gradient descent) can be adapted for large-scale behavioral modeling, but challenges remain in balancing privacy budgets with model utility, especially in high-dimensional, multimodal settings. Future work could explore scalable DP algorithms tailored for content-behavior corpora, as well as hybrid approaches that combine DP with aggregation and anonymization for layered protection.

Fairness-Aware Learning Algorithms. While current filtering and evaluation protocols mitigate some sources of bias, algorithmic interventions can further reduce disparate impact across demographic groups. Approaches such as

adversarial de-biasing (Zhang *et al.*, 2018*a*), reweighting, or multi-objective optimization (Hardt *et al.*, 2016) can be integrated into training to explicitly penalize or correct for unfairness. For behavioral data, this may involve learning group-invariant representations or optimizing for fairness metrics (e.g., equalized odds, demographic parity) alongside predictive accuracy. A key challenge is the lack of explicit demographic labels in many behavioral datasets, motivating research into unsupervised or proxy-based fairness interventions.

Federated and Decentralized Learning. Federated learning (McMahan *et al.*, 2017) offers a promising paradigm for privacy-preserving behavioral modeling by keeping raw user data on-device and only sharing model updates. This approach can significantly reduce privacy risks and regulatory burdens, especially for sensitive behavioral signals. However, federated optimization introduces new challenges, such as handling non-IID (non-independent and identically distributed) data, communication efficiency, and robustness to adversarial clients. Future work could investigate federated architectures for multimodal content-behavior models, as well as secure aggregation and differential privacy at the edge.

Bias Auditing and Transparency. Systematic bias auditing and transparent reporting are essential for responsible deployment of behavioral models. Regular audits—using techniques such as subgroup performance analysis, counterfactual fairness testing, and adversarial probing—can reveal hidden disparities and inform mitigation strategies. The release of model cards and datasheets (Mitchell *et al.*, 2019) for both datasets and models can further enhance transparency, enabling stakeholders to assess limitations, intended use cases, and known risks. Future research may also explore automated tools for continuous bias monitoring and explainability in large-scale, real-world deployments.

Chapter 4

Analyzing Behavior: Teaching Behavior Improves Content Understanding

In the last chapter, we discussed training a single model that learns about both content and behavior. We saw that a model trained on content and behavior together shows capabilities of behavior and content simulation, behavior domain adaptation, and improvements in behavior and content understanding. Behavior is produced by the receiver as a response to the message sent by the sender. Behavior, as an artifact of communication, is generated by a receiver in response to content (Fig. 3.1) sent by a communicator. Therefore, it comes later than content in the time axis. Hence, behavior contains signals about content, which can help in *understanding* content. However, since it comes after content, the signals are available post-hoc. These signals, if properly harnessed, should be able to increase performance on the message understanding tasks popular in NLP and CV, like question answering, sentiment analysis, topic classification, *etc.* Despite this, behavior data is considered noise and is ignored while training large language models (Biderman *et al.*, 2022; Penedo *et al.*, 2023) and also large vision and language models (Liu *et al.*, 2023*a*; Zhu *et al.*, 2023). In this paper, we explore this line of thought more.

Humans produce two kinds of behavioral signals upon observing a message (Bertenthal, 1996; Prinz, 1997): perceptual signals and actions as behavior. Perceptual signals, like seeing, touching, and hearing, help a receiver primarily sense the world around her, ultimately guiding her actions. Actions are how a receiver acts on the outside world. The signals produced by the human receiver upon receiving a message carry information about the message itself (Fig. 4.7). For instance, if a person's heartbeat rises upon watching a movie scene, it can help us infer that perhaps the scene was an exciting scene (Dzedzickis *et al.*, 2020). Similarly, regressing while reading is indicative of important or confusing phrases (Bicknell and Levy, 2011). In these cases, perception behavior helps us derive

inferences about content. In a similar vein, the actions a person performs after watching a movie, such as comments and likes, carry signals about the movie (Fig. 4.7, 4.8).

With this background, in this chapter, we show improvement in content understanding of LLMs and VLMs using the following types of behavior:

1. **Perception as Behavior**: We first prove the hypothesis using the perception behavior of scanpaths. The choice of scanpaths as the target behavior is motivated by prior literature (Clifton Jr *et al.*, 2007; Demberg and Keller, 2008; Karessli *et al.*, 2017; Yu *et al.*, 2017; He *et al.*, 2019; Boyd *et al.*, 2022; Mishra *et al.*, 2016*b*; Long *et al.*, 2017) where they show that eye movements of the receiver can help determine linguistic and perceptual factors in text and images. Therefore, in this section, we talk about how perception behavior can be used to improve content understanding of LLMs in a post-hoc manner. Next, we solve the problem of behavior being available post-hoc by generating synthetic scanpath behavior for a content, and showing that using the synthetic behavior also improves understanding content.

2. **Action as Behavior**: Further, we make initial efforts to collect and understand digital analytics at scale with the aim of integrating them with VLMs to improve their downstream content understanding capabilities. We introduce methods for filtering and cleaning behavioral data and then propose tasks for large language and vision models, leading to improvements in language and visual content understanding tasks. For this, we look to Reddit and YouTube as two major sources of visual content and human behavior in the form of viewer comments, likes, replay graphs, and upvotes.

4.1 Synthesizing Human Gaze Feedback for Improved NLP Performance

Integrating human feedback in models can improve the performance of natural language processing (NLP) models. Feedback can be either explicit (*e.g.* ranking used in training language models) or implicit (*e.g.* using human cognitive signals in the form of eyetracking). Prior eye tracking and NLP research reveal that cognitive processes, such as human scanpaths, gleaned from human gaze patterns aid in the understanding and performance of NLP models. However, the collection of *real* eyetracking data for NLP tasks is challenging due to the requirement of expensive and precise equipment coupled with privacy invasion issues. To address this challenge, we propose ScanTextGAN, a novel model for *generating*

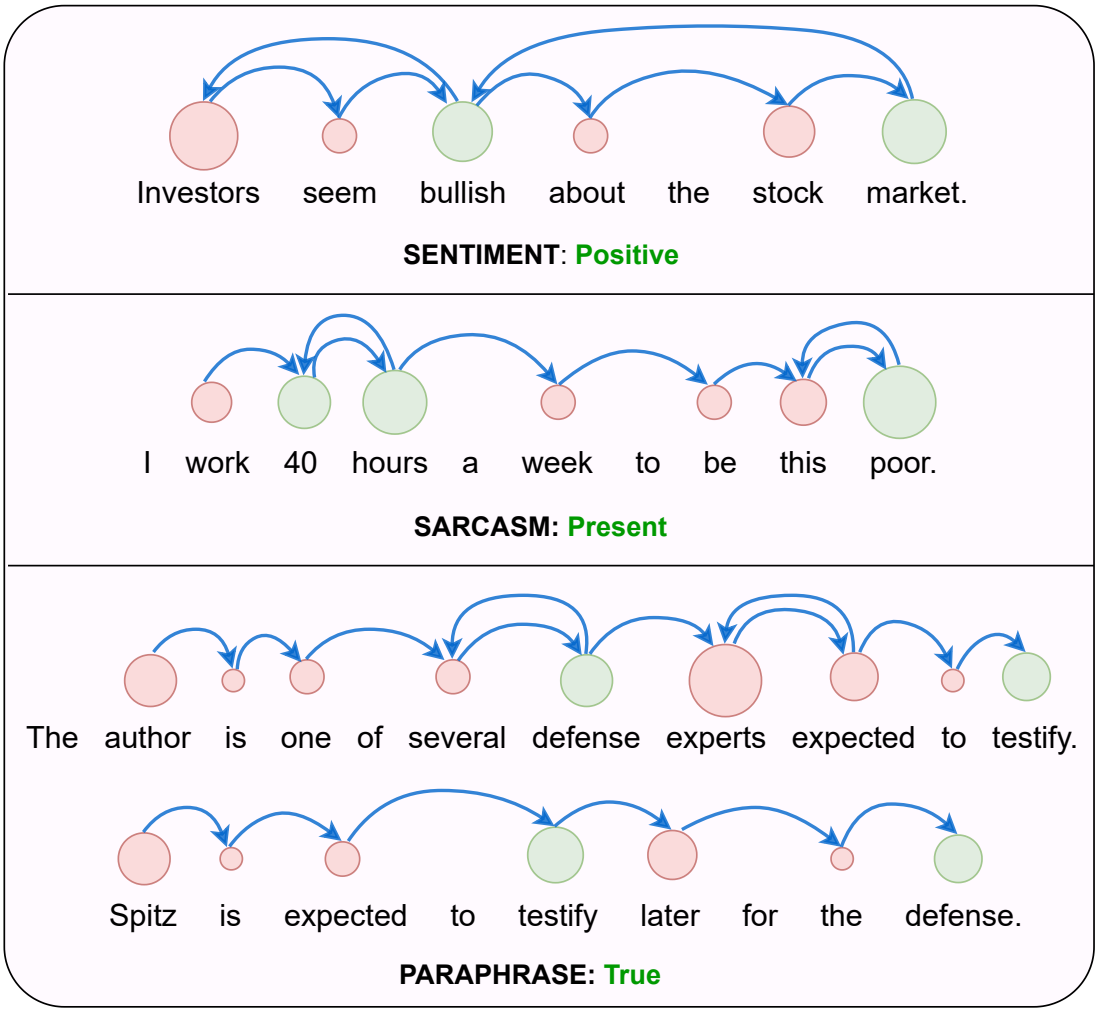


Figure 4.1: Generated scanpaths over text samples taken from various natural language processing (NLP) tasks. The green circles denote the important words characteristic of that task. The circles' size denotes the fixation duration, and the arrows depict the saccadic movements. As can be seen, linguistically important words often have a higher fixation duration and revisit. Regressions (word revisits) also appear in the examples.

human scanpaths over text. We show that ScanTextGAN-generated scanpaths can approximate meaningful cognitive signals in human gaze patterns. We include synthetically generated scanpaths in four popular NLP tasks spanning six different datasets as proof of concept and show that the models augmented with generated scanpaths improve the performance of all downstream NLP tasks.

4.1.1 Introduction

Integrating human signals with deep learning models has been beginning to catch up in the last few years. Digital traces of human cognitive processing can provide valuable signals for Natural Language Processing (Klerke *et al.*, 2016*b*; Plank,

2016*b*). Various approaches for integrating human signals have been explored. For example, human feedback for better decisioning (Christiano *et al.*, 2017), NLP tasks (Stiennon *et al.*, 2020; Wu *et al.*, 2021), and most recently language modeling using reinforcement learning with human feedback (RLHF) based reward (Bai *et al.*, 2022; Ouyang *et al.*, 2022). RLHF involves explicit human feedback and is expensive and hard to scale. On the other hand, previous studies have also tried to use implicit human feedback in the form of eyetracking signals. It has proven to be a useful signal for inferring human cognitive processing (Sood *et al.*, 2020; Hollenstein and Zhang, 2019*a*; Mathias *et al.*, 2020). NLP researchers have focused on assessing the value of gaze information extracted from large, mostly disjointly labeled gaze datasets in recurrent neural network models (Ren and Xiong, 2021; Strzyz *et al.*, 2019; Barrett *et al.*, 2018*a*). The proposed approaches under this paradigm include gaze as an auxiliary task in multi-task learning (Klerke *et al.*, 2016*a*; Hollenstein *et al.*, 2019), as additional signals (Mishra *et al.*, 2016*b*), as word embeddings (Barrett *et al.*, 2018*b*), as type dictionaries (Barrett *et al.*, 2016*a*; Hollenstein and Zhang, 2019*a*), and as attention (Barrett *et al.*, 2018*a*).

Previous studies demonstrate that human scanpaths (temporal sequences of eye fixations, see Fig. 4.1) gleaned from eye tracking data improve the performance of NLP models. However, the real-world application of these methods remains limited primarily due to the cost of precise eye-tracking equipment, users' privacy concerns, and manual labor associated with such a setup. Therefore, generating scanpaths from existing eyetracking corpora would add great value to NLP research. To the best of our knowledge, this is the first work to propose a model that generates scanpaths for a given read text with good accuracy. We call the model, ScanTextGAN.

We demonstrate the scanpath generation capability of ScanTextGAN over three eye-tracking datasets using multiple evaluation metrics. Further, we evaluate the utility of *generated* scanpaths for improvements in the performance of multiple NLP tasks (see Figs. 4.1,4.2) including the ones in the GLUE benchmark (Wang *et al.*, 2018*a*). The generated scanpaths achieve similar performance gains as the models trained with real scanpaths for classic NLP tasks like sentiment classification, paraphrase detection, entailment, and sarcasm detection.

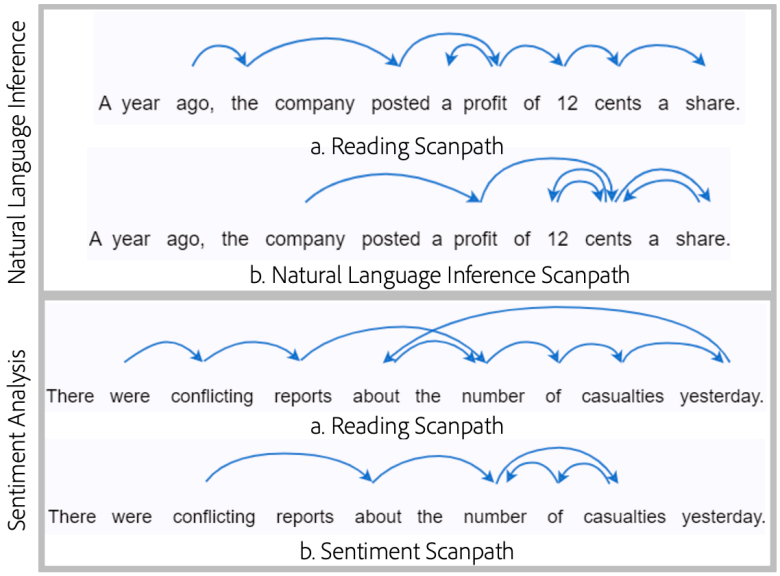


Figure 4.2: (Intent-aware) Scanpath samples generated by conditioning scanpath generation on different downstream natural language tasks. Note that the conditioned scanpaths are heavily biased to words important for that downstream task.

Our contributions are threefold:

1. We propose ScanTextGAN, the first scanpath generator over text.

2. We compare ScanTextGAN with multiple baselines and conduct ablation experiments with varying models and configurations. The model performs well on the test sets and cross-domain generalization on two additional eyetracking datasets belonging to different text domains.

3. We tested the usefulness of generated scanpaths in downstream NLP tasks such as sentiment analysis, paraphrase detection, and sarcasm detection on six different datasets. The results show that the downstream NLP tasks benefited significantly from cognitive signals inherent in generated scanpaths. Further, we show how scanpaths change when finetuning with downstream natural language tasks (Figs.4.2,4.6) and that they lead to further improvements in downstream task performance (§4.1.4.3) showing how they can act as additional controls beyond the task architecture.

4.1.2 Related Work

When reading a text, humans do not focus on every word and often do not read sequentially (Just and Carpenter, 1980). A series of studies in psycho-linguistics

have shown that the number of fixations and the fixation duration on a word depend on several linguistic factors. The linguistic factors can also be determined given the cognitive features (Clifton Jr *et al.*, 2007; Demberg and Keller, 2008). Though advances in ML architecture have helped bring machine comprehension closer to human performance, humans are still superior for most NLP tasks (Blohm *et al.*, 2018; Xia *et al.*, 2019).

It has been shown in the literature that integrating explicit (Bai *et al.*, 2022; Ouyang *et al.*, 2022) and implicit (cognitive processing) human feedback signals in traditional ML models is expected to improve their performance (Just and Carpenter, 1980). However, the cost of explicit feedback (e.g., using MTurk) and implicit feedback (e.g., eye tracking) at scale is excessively high. Similarly, privacy-invasive eye-tracking processes limit the scope of this idea. One way to address this problem is to use generated eye movements to unfold the full potential of eye-tracking research. Hence, the idea is to architect ScanTextGAN, a scanpath generator for text reading, and test its usefulness in downstream NLP tasks.

More precisely, this work builds upon previous works on 1) human attention modeling and 2) gaze integration in neural network architectures, which are described as follows:

Human Attention Modeling: Predicting what people visually attend to in images (saliency prediction) is a long-standing challenge in neuroscience and computer vision, the fields have seen many data-based models (Wang *et al.*, 2021*b*). In contrast to images, most attention models for eye movement behaviors during reading are cognitive process models, *i.e.*, models that do not involve machine learning but implement cognitive theories (Engbert *et al.*, 2005; Xia *et al.*, 2019). Key challenges for such models are a limited number of parameters and hand-crafted rules. Thus, it is difficult to adapt them to different tasks and domains and use them as part of end-to-end trained ML architectures (Kotseruba and Tsotsos, 2020). In contrast, learning-based attention models for text remain under-explored. Within that, all eye tracking models are saliency prediction models with non-existent work in predicting scanpaths. On the other hand, visual scanpaths generation for image-based eye tracking data has been recently explored for both traditional (Assens *et al.*, 2019) and 360° images (Martin *et al.*, 2022).

Matthies *et al.* (Matthies and Søgaard, 2013) presented the first fixation prediction work for text. They built a person-independent model using a linear Conditional Random Fields (CRF) model. Hahn and Keller (Hahn and Keller, 2016) designed the Neural Attention Trade-off (NEAT) language model, which was trained with hard attention and assigned a cost to each fixation. Other approaches include sentence representation learning using surprisal and part of speech tags as proxies to human attention (Wang *et al.*, 2017).

Our work differs from previous studies as we combine cognitive theory and data-driven approaches to predict scanpaths and further show its application in downstream NLP tasks (Hollenstein *et al.*, 2021*b*,*a*).

Integrating Gaze in Network Architecture: Integration of human gaze data into neural network architectures has been explored for a range of computer vision tasks such as image captioning, visual question answering, and tagging (Karessli *et al.*, 2017; Yu *et al.*, 2017; He *et al.*, 2019; Boyd *et al.*, 2022). Hence, recent research has utilized features gleaned from readers' eye movement to improve the performance of complex NLP tasks such as sentiment analysis (Long *et al.*, 2017; Mishra *et al.*, 2016*c*), sarcasm detection (Mishra *et al.*, 2016*b*), part-of-speech tagging (Barrett *et al.*, 2016*b*), NER (Hollenstein and Zhang, 2019*a*), and text difficulty (Reich *et al.*, 2022).

While in recent years, eye tracking data has been used to improve and evaluate NLP models, the scope of related studies remains limited due to the requirement of real-time gaze data at inference time. Mathias *et al.* (Mathias *et al.*, 2020) reported that there exists no automated way of generating scanpaths yet in the literature. With high-quality artificially generated scanpaths, the potential of leveraging eyetracking data for NLP can be unfolded. Additionally, generating scanpaths that mimic human reading behavior will help advance our understanding of the cognitive processes behind language understanding. Hence, we propose ScanTextGAN; researchers can use that to generate scanpaths over any text without worrying about collecting them from real users.

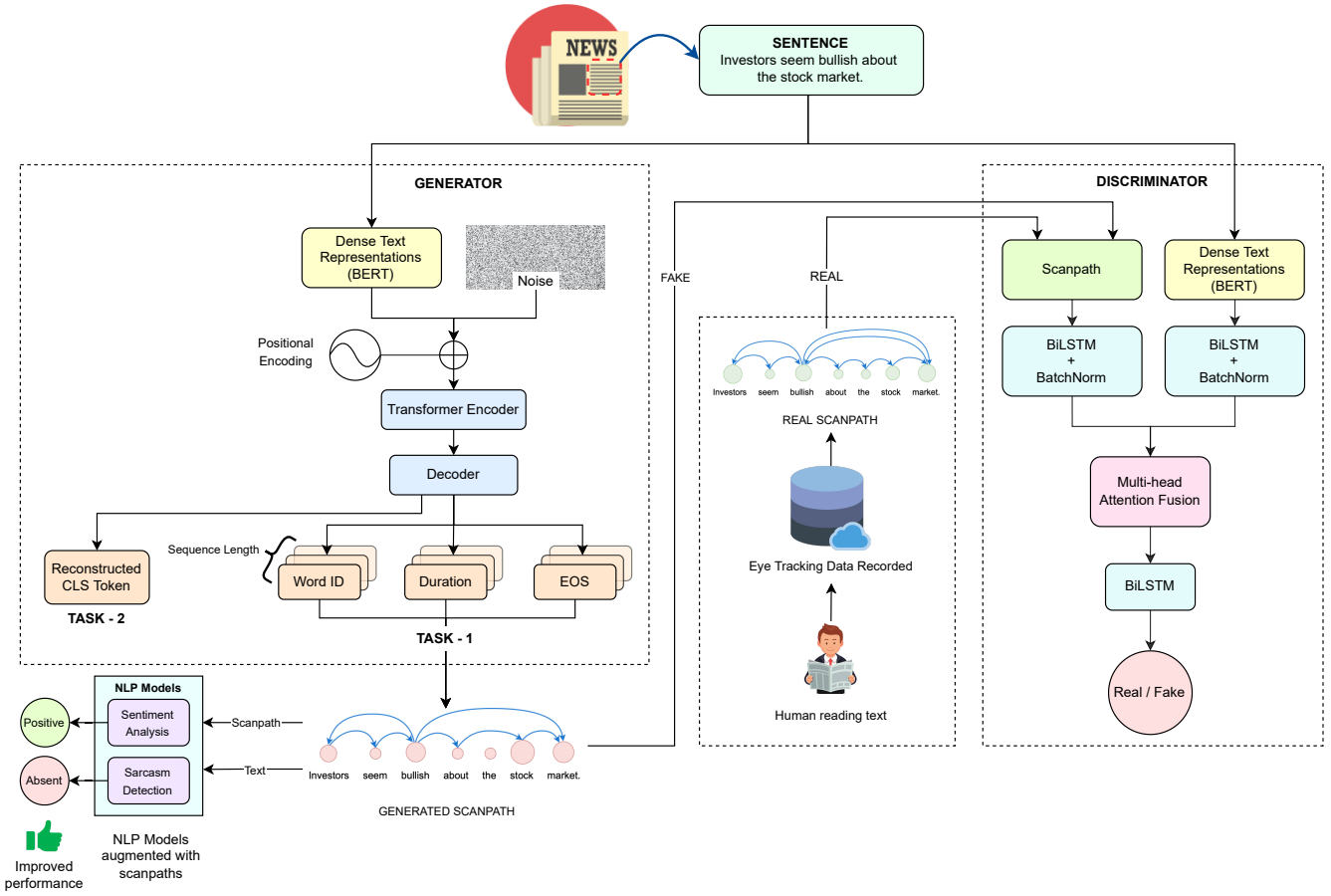


Figure 4.3: The architecture of the proposed **ScanTextGAN** model. The model consists of a conditional generator and a discriminator playing a zero-sum game. The generator is trained by two cognitively inspired losses: text content reconstruction and scanpath content reconstruction.

4.1.3 Proposed Model

In this section, we define the scanpath generation task, describe the ScanTextGAN model architecture, and provide details on loss functions and model training.

Task Definition: The task of scanpath generation is to generate a sequence $\mathcal{S}(\mathcal{T})$ representing a scanpath over the text $\mathcal{T} = \{w_1, w_2, ..., w_n\}$ composed of a sequence of words, can be defined as follows:

$$\mathcal{S}(\mathcal{T}) = \{.., (w_a^i, t^i), ..., (w_b^j, t^j), ..., (w_c^k, t^k)\} \tag{4.1}$$

where t^i represents the fixation duration over the word w_a occurring at the position i. Note that it is not necessary to have $a < b$ (words being read in linear order) or that $k = n$ (the number of fixations being equal to the number of words). Due to regressions, *i.e.*, backward saccades to previous words, words are also revisited. Hence, the same word could appear multiple times in the sequence.

4.1.3.1 ScanTextGAN Model Architecture

Fig. 4.3 illustrates the proposed conditional GAN architecture of the model. The ScanTextGAN model is composed of two competing agents. First, a conditional generator that generates scanpaths given text prompts. The second is a discriminator network, which distinguishes real human scanpaths from the generated ones. The ScanTextGAN model is trained by combining text content loss, scanpath content loss, and adversarial loss (Eq. 4.6). The scanpath content loss measures the difference between the predicted scanpath and the corresponding ground truth scanpath. The text content loss reconstructs the input text, and the adversarial loss depends on the real/synthetic prediction of the discriminator over the generated scanpath. We describe the losses along with the generator and discriminator architectures next.

Generator: The ScanTextGAN generator constitutes a transformer-based encoder-decoder framework. The encoder is conditioned on BERT-based text embeddings (Devlin *et al.*, 2019), which are concatenated with noise to make the generator's output non-deterministic. The output of the transformer encoder is

supplied to the decoder, which consists of task-specific feed-forward networks. One branch generates the scanpath (*Task 1*), while the other reconstructs the 768 dimensional CLS token embedding of the sentence (*Task 2*). The scanpath is output as a temporal sequence of word ID (fixation points) w_a^i, fixation duration t^i, and end-of-sequence probability EOS^i. At inference time, the length $L(G)$ of generated scanpath G is determined as follows:

$$L(G) = \begin{cases} \min_{1 \leq k \leq M}(k) & \text{if } EOS^k > \tau \\ M & \text{otherwise} \end{cases} \tag{4.2}$$

where M is the maximum scanpath length as described in section §4.1.3.2 and $\tau \in (0, 1)$ is a probability threshold. We use $\tau = 0.5$. The loss functions of the two branches are described below.

Scanpath Content Loss tries to minimize the deviation of generated scanpaths $\mathcal{G}(\mathcal{T}, \mathcal{N})$ from the ground-truth scanpaths $\mathcal{R}(\mathcal{T}, h))$ over text $\mathcal{T}$ where ground-truth scanpaths are recorded from the human h and $\mathcal{N}$ stands for Gaussian noise $\mathcal{N}(0, 1)$. The loss function $\mathbb{L}_s$ is given as:

$$\begin{aligned} \mathbb{L}_s(\mathcal{G}(\mathcal{T},\mathcal{N}), \mathcal{R}(\mathcal{T}, h)) = \frac{1}{k}\Sigma_{i=0}^{k}(\alpha(id_g^i - id_r^i)^2 + \\ \beta(t_g^i - t_r^i)^2 + \gamma(E_g^i - E_r^i)^2) \end{aligned} \tag{4.3}$$

which is a weighted sum of three terms. The first term measures the error between real and predicted *fixation points* given by the mean squared difference between generated and real word-ids $(id_g^i - id_r^i)$. It penalizes permutations of word ids and trains the model to approximate the real sequence of fixation points closely.

The second term measures the difference in *fixation durations* given by the mean squared difference between generated and real duration $(t_g^i - t_r^i)$. Fixation durations simulate human attention over words in the input text. Thus, a word with a larger fixation duration is typically synonymous with greater importance than other words in the input text. This error term supplements the generator's ability to learn human attention patterns over the input text.

Finally, the third term measures the mean squared error between the prediction of end-of-sequence probability by real and generated distributions $(E_g^i - E_r^i)$. These

are weighted by the hyperparameters α, β, and γ. Preliminary experiments showed that optimizing the mean squared error leads to better performance over the cross-entropy loss for optimizing the EOS probability output.

Text Content Loss: Scanpaths depend heavily on the linguistic properties of the input text. Therefore, to guide the generator towards near the probable real data manifolds, we adopt reconstruction of the CLS token embedding of the input text (*Task 2*) by the generator as an auxiliary task since the CLS token embedding encodes a global representation of the input text. This text content reconstruction loss $\mathbb{L}_r$ is given as:

$$\begin{aligned}\mathbb{L}_r(\mathcal{G}(\mathcal{T},\mathcal{N}), \mathcal{R}(\mathcal{T}, h)) = (BERT(w_i^g, w_j^g, ..., w_k^g \\ -BERT(w_a^r, w_b^r, ...w_n^r))^2\end{aligned} \tag{4.4}$$

where $BERT(w_a^r, w_b^r, ...w_n^r)$ and $BERT(w_i^g, w_j^g, ...w_k^g)$ stand for the *CLS* vector representations of real and generated text respectively.

Discriminator: The goal of the discriminator is to distinguish between the real and synthetic scanpaths supplied to it. Similar to the generator, it requires text representations to distinguish between real and generated scanpaths. Specifically, the discriminator comprises two blocks of BiLSTMs that perform sequential modeling over the scanpaths and BERT embeddings. The outputs of the two branches are combined and passed to an attention fusion module with four heads, followed by another network of BiLSTMs. The hidden states of the last BiLSTM layer from both forward and backward directions are concatenated and supplied to a feed-forward network. A Sigmoid function activates the output of the feed-forward network. In this manner, the discriminator classifies the input scanpaths as either *real* or *fake*.

Adversarial Loss: The generator and discriminator networks are trained in a two-player zero-sum game fashion. The loss is given by:

$$\begin{aligned}\mathbb{L}_a = \min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x|\mathcal{T}, h)] + \\ \mathbb{E}_{z\sim p_z(z)}[1 - \log D(G(z|\mathcal{T}, \mathcal{N}))]\end{aligned} \tag{4.5}$$

Therefore, the net generator loss becomes:

$$\mathbb{L}_g = \mathbb{L}_s + \mathbb{L}_r + \mathbb{E}_{z\sim p_z(z)}[1 - \log D(G(z|\mathcal{T}, \mathcal{N}))] \tag{4.6}$$

4.1.3.2 Dataset

For training the ScanTextGAN model, we use the CELER dataset (Berzak *et al.*, 2022). It contains eyetracking data of 365 participants for nearly 28.5 thousand newswire sentences, sourced from the Wall Street Journal Penn Treebank (Marcinkiewicz, 1994). Each participant in CELER reads 156 newswire sentences. Half of the sentences are shared across participants, and the rest is unique to each participant. The maximum sentence length was set to 100 characters. Participant eyetracking data were recorded using Eyelink 1000 tracker in a desktop mount configuration with a sampling rate of 1000 Hz. The ScanTextGAN model is trained to approximate the average eye movements of all the participants who read given sentences. The CELER dataset was envisioned to enable research on language processing and acquisition and to facilitate interactions between psycholinguistics and natural language processing. Furthering the goal, we use it to train our conditional GAN model through which we show human scanpath approximation capabilities (§4.1.4.2). Also, we use it to show improvements in the performance of NLP tasks (§4.1.4.3).

The data consist of tuples of participant ID, sentence ID, and word ID corresponding to fixation point and fixation duration. We compute the 99th percentile of fixation durations and treat it as the largest value. Fixations of durations longer than this are treated as outliers and hence dropped from the dataset. To apply the scanpath reconstruction loss (Eq. 4.3), we scale all fixation durations by the maximum value and then normalize them to [0,1]. Similarly, word IDs in each sentence are normalized to [0, 1] after scaling them by the length of that sentence. For the last fixation point in every scanpath, the binary EOS token is set to 1. The maximum scanpath length is set to 80 fixation points (99th percentile of the lengths). Thus shorter scanpaths are padded while longer scanpaths are trimmed. We use BERT to encode the sentences and obtain their 768-dimensional embeddings, keeping the max length parameter as 80, thus resulting in an 80×768

dimensional tensor.

4.1.3.3 Parameter Settings

Sinusoidal positional encoding is applied over the input embeddings fed to the generator. We use a 3-layer transformer encoder with four head attention and a hidden dimension size of 776 in the generator. In the discriminator, we use bidirectional LSTMs over sentence embeddings and generated scanpaths with a hidden size of 64 and a dropout ratio of 0.3, followed by batch normalization for faster convergence. An attention module with four attention heads is applied after concatenating the outputs. We employ the Adam and RMSProp optimizer to minimize generator and discriminator losses. The batch size is set to 128, the initial learning rate of the generator to 0.0001, and that of the discriminator to 0.00001. The model is trained for 300 epochs. Our implementation uses PyTorch, a popular deep-learning framework in Python. All experiments are run on an Intel Xeon CPU with Nvidia A100-SXM GPUs.

4.1.4 Performance Evaluation

We quantify the performance of ScanTextGAN in two regimes[*]; first, scanpath generation with three datasets, and second, NLP tasks with six datasets. Similar to prior computer vision studies (Sun *et al.*, 2019*b*; de Belen *et al.*, 2022; Kümmerer and Bethge, 2021; Jiang *et al.*, 2016), we evaluate the ScanTextGAN model over the scanpath generation task. For this, we use the test split of the CELER dataset, Mishra *et al.* (2016) (Mishra *et al.*, 2016*a*), and Mishra *et al.* (2017) (Mishra *et al.*, 2017). In addition, unlike the computer vision studies, we also evaluate the ScanTextGAN model for improvement in NLP tasks. The hypothesis is that the human eyes (and consequently the brain) process many language comprehension tasks unconsciously and without visible effort. The next logical step is to capture (or, in our case, generate) this mental representation of language understanding and use it to improve our machine-learning systems. For evaluation, we use four tasks from the GLUE benchmark and two from the tasks proposed by (Mishra

[*]All results are calculated with five random seeds and reported as the mean of those five runs

et al., 2016*a*). While the ScanTextGAN model is trained over news text from the CELER dataset, with the help of the other datasets, we expand our testing to other domains, including reviews, quotes, tweets, and Wikipedia text.

4.1.4.1 Evaluation Datasets

Mishra *et al.* (2017) (**Mishra *et al.*, 2017**) comprises eye movements and reading difficulty data recorded for 32 paragraphs on 16 different topics, *viz.* history, science, literature, *etc.* For each topic, comparable paragraphs were extracted from Wikipedia[†] and simple Wikipedia[‡]. The participant's eye movements are tracked using an SR-Research Eyelink-1000 Plus eye tracker. Using the ground truth scanpaths over the text corpora, we evaluate the quality of generated scanpaths.

Mishra *et al.* (2016) (**Mishra *et al.*, 2016*a***) contains eye fixation sequences of seven participants for 994 text snippets annotated for sentiment and sarcasm. These were taken from Amazon Movie Corpus , Twitter, and sarcastic quote websites. The task assigned to the participants was to read one sentence at a time and annotate it with binary sentiment polarity labels (*i.e.*, positive/negative). The same datasets were used in several studies (Joshi *et al.*, 2015; Mishra *et al.*, 2016*b*,*c*) to show improvements in sarcasm and sentiment analysis. We use the datasets to evaluate both the generation quality and potential improvements in NLP tasks.

Furthermore, we explore the potential of including cognitive signals contained in scanpaths in NLP models for a range of GLUE tasks which include Sentiment Analysis using Stanford Sentiment Treebank (SST), Paraphrase Detection using Microsoft Research Paraphrase Corpus (MRPC) and Quora Question Pairs (QQP), Natural Language Inference using Recognizing Textual Entailment (RTE) dataset.

Next, we cover the results of scanpath generation and its application in NLP

[†] https://en.wikipedia.org/

[‡] https://simple.wikipedia.org/

[§] In the CELER dataset, there are only 78 shared sentences amongst all the participants. Therefore, inter-subject scanpath evaluation is done only for these sentences. In contrast, the ScanTextGAN results are reported for the entire test set (including these 78 sentences).

Generator Model	MultiMatch ↑				Levenshtein Distance ↓
	Vector↑	Length↑	Position↑	Duration↑	
Inter-subject score[§]	0.973	0.958	0.830	0.698	0.691
LSTM Encoder-Decoder trained with scanpath content loss	0.975	0.956	0.765	0.344	0.865
ScanTextGAN – Text Reconstruction – GAN Loss	0.968	0.947	0.728	0.703	0.779
ScanTextGAN	**0.983**	**0.972**	**0.787**	0.733	**0.769**
ScanTextGAN – Text Reconstruction	0.974	0.957	0.773	0.703	0.798
ScanTextGAN – GAN Loss	0.973	0.955	0.750	**0.761**	0.786
ScanTextGAN + addition of noise	0.971	0.952	0.756	0.736	0.791
ScanTextGAN – Text (CLS) Reconstruction + sentence reconstruction	0.978	0.963	0.724	0.721	0.805

Table 4.1: In-domain Evaluation of Scanpath Generation on the CELER dataset (Berzak *et al.*, 2022).

tasks.

4.1.4.2 Evaluation of Scanpath Generation

We evaluate the scanpath generation model on two most commonly used metrics in image scanpath generation studies (Sun *et al.*, 2019*b*; Chen and Sun, 2018; de Belen *et al.*, 2022; Kümmerer *et al.*, 2022): **MultiMatch** (Jarodzka *et al.*, 2010) and **Levenshtein Distance** (Levenshtein, 1965). Multimatch is a geometrical measure that compares scanpaths across a comprehensive set of dimensions composed of shape, lengths, position, and fixation duration. Levenshtein Distance between a pair of sequences measures the least number of edits (inserts, deletes, substitution) to transform one into the other.

4.1.4.2.1 Scanpath Evaluation Metrics **MultiMatch** is a geometrical measure that models scanpaths as vectors in 2-D space, wherein the vectors represent saccadic eye movements. Starting and ending coordinates of these saccades constitute the fixation positions. It compares scanpaths across multiple dimensions, *viz.* shape, length, position, direction, and fixation duration. Shape measures the vector difference between aligned saccade pairs, which is then normalized by twice the diagonal screen size. Length measures the normalized difference between the endpoints of real and generated saccade vectors. Direction is the angular distance between the two vectors. The position is the Euclidean difference in position between aligned vectors, and duration measures the difference in fixation durations

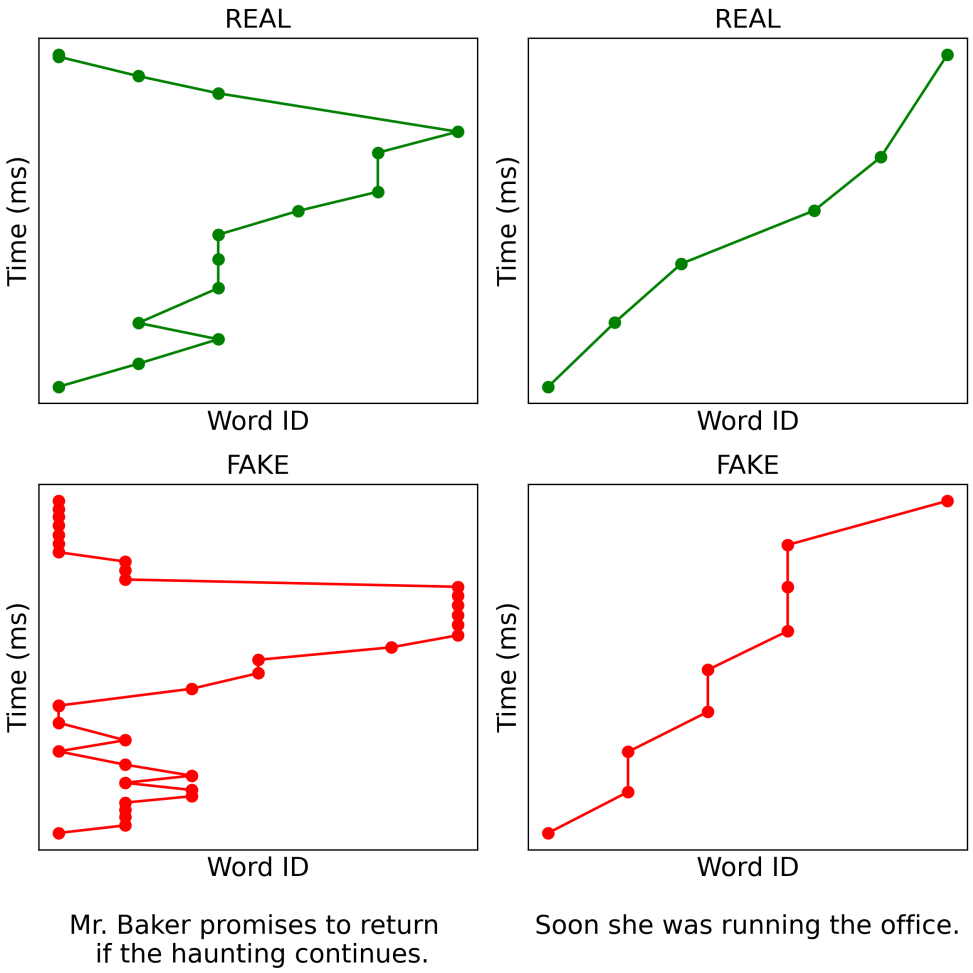


Figure 4.4: Comparison of *real* and *synthesized* scanpaths corresponding to a few text samples. The proposed ScanTextGAN model generates the latter.

normalized against the maximum duration. Since our work deals with scanpaths over text, we use 1-D space to represent the saccade vectors where word IDs denote the fixation positions. Thus, it is easy to see that computing scanpath direction similarity is redundant here (it is subsumed within position); hence we drop it from our analysis.

Levenshtein Distance between a pair of sequences measures the least number of character edits, i.e., insertion, deletion, and substitution needed to transform one sequence into the other. Specifically, we use it to gauge the degree of dissimilarity between a pair of real R and generated G scanpaths. To account for the fixation durations of each word, R and G are temporally binned using a 50 ms bin size, similar to the computation of ScanMatch metric (Cristino *et al.*, 2010). The resulting sequences of word IDs, R_W and G_W are transformed into character strings, $R_S = \{r_1, r_2, ..., r_n\}$ and $G_S = \{g_1, g_2, ..., g_m\}$, where R_S and G_S are strings over the ASCII alphabet and $n = |R_S|$ and $m = |G_S|$. Thus, a lower NLD score is indicative of greater scanpath similarity.

Further, as a top-line comparison, we use **inter-subject scanpath similarity** (Sun *et al.*, 2019*b*). It measures the degree of variation among real human scanpaths corresponding to each text input. To compute this, we first calculate each subject's performance by treating the scanpaths of other subjects as the ground truth. Then, the average value of all subjects is used as inter-subject performance.

Baselines: Since ScanTextGAN is the first text-based scanpath generation model, we conduct an ablation study to compare ScanTextGAN with its other variants. Specifically, we compare ScanTextGAN with the following six configurations: (1) An LSTM-based network trained with scanpath content loss. Sentence embeddings obtained through BERT are concatenated with noise in this model. The resultant is fed to an attention module with four heads, then passed to a network of LSTMs and Batch Normalization layers applied in tandem. (2) ScanTextGAN model trained with only the scanpath content loss. (3) ScanTextGAN model without the text reconstruction loss (Task-2). (4) ScanTextGAN model with BERT-based sentence embeddings reconstruction instead of CLS token reconstruction. (5) ScanTextGAN model with the addition of noise instead of concatenation. (6) ScanTextGAN model trained without GAN loss.

Generator Model	MultiMatch ↑				Levenshtein Distance ↓
	Vector↑	Length↑	Position↑	Duration↑	
Inter-subject score	0.977	0.963	0.839	0.715	0.723
LSTM Encoder-Decoder trained with scanpath content loss	**0.984**	**0.973**	0.714	0.379	0.918
ScanTextGAN – Text Reconstruction – GAN Loss	0.977	0.960	0.780	0.769	0.847
ScanTextGAN	0.966	0.945	**0.791**	**0.771**	**0.836**
ScanTextGAN – Text Reconstruction	0.976	0.961	0.763	0.757	0.845
ScanTextGAN – GAN Loss	0.976	0.959	0.774	0.768	0.839
ScanTextGAN + addition of noise	0.968	0.947	0.737	0.743	0.838
ScanTextGAN – Text (CLS) Reconstruction + sentence reconstruction	0.964	0.934	0.747	0.733	0.869

Table 4.2: Cross-domain Evaluation of Scanpath Generation on the Dataset by (Mishra *et al.*, 2016*a*).

Generator Model	MultiMatch ↑				Levenshtein Distance ↓
	Vector↑	Length↑	Position↑	Duration↑	
Inter-subject score	0.994	0.991	0.834	0.620	0.845
LSTM Encoder-Decoder trained with scanpath content loss	**0.992**	**0.987**	0.596	0.329	0.969
ScanTextGAN – Text Reconstruction – GAN Loss	0.990	0.984	0.729	0.705	0.951
ScanTextGAN	0.984	0.977	**0.759**	0.693	**0.931**
ScanTextGAN – Text Reconstruction	0.986	0.981	0.756	**0.706**	0.939
ScanTextGAN – GAN Loss	0.990	0.984	0.739	**0.706**	0.945
ScanTextGAN + addition of noise	0.984	0.976	0.759	0.703	0.943
ScanTextGAN – Text (CLS) Reconstruction + sentence reconstruction	0.983	0.974	0.667	0.674	0.958

Table 4.3: Cross-domain Evaluation of Scanpath Generation on the Dataset by (Mishra *et al.*, 2017).

Results: Table 4.1 presents the results of our scanpath prediction model on the CELER dataset. Further, we also compare ScanTextGAN with baselines on two other contemporary datasets of movie reviews, tweets, and sarcastic quotes (Mishra *et al.*, 2016*a*), Wikipedia and simple Wikipedia paragraphs (Mishra *et al.*, 2017). Tables 4.2 and 4.3 present the results of our model on those datasets. For obtaining results on these corpora, we use the model trained on the CELER dataset, thus helping us evaluate the cross-domain performance of the model.

As can be seen in Table 4.1, Table 4.2 and Table 4.3, ScanTextGAN outperforms other models for scanpath prediction on most metrics. The performance of ScanTextGAN even surpasses inter-subject reference on Duration and comes very close to Vector, Length, and Position.

We observe that adopting the reconstruction of the CLS token as an auxiliary task (Task - 2) boosts the model performance. Reconstructing the full sentence

embeddings rather than the CLS tokens only as an auxiliary task does not always improve the results, despite adding a larger computational overhead. The results also reveal that concatenating noise with text embeddings is more rewarding than adding it.

Further, to compare the skipping behavior of ScanTextGAN with humans, we calculate the weighted F1 score of the words skipped and attended by both model types. We find the weighted F1 to be 64.6 between them. Fig. 4.4 presents a visual comparison between real scanpaths from the available eyetracking data and scanpaths generated by ScanTextGAN, corresponding to some randomly chosen text samples. We can observe that the generated scanpaths resemble the real ones to a great extent. Thus, the quantitative and qualitative results on in-domain and cross-domain settings lead us to believe that our proposed scanpath generation model can be deemed a good approximator of the human scanpaths.

4.1.4.3 Application to NLP Tasks

We use them to augment various NLP models and measure their performance to demonstrate the usefulness of cognitive signals hidden in the *generated* scanpaths.

Sentiment Classification and Sarcasm Detection: For these tasks, we use a model consisting of a network of two branches of BiLSTMs and Batch Normalization layers that perform sequential modeling over text representations obtained through BERT and scanpaths fed as input to the model. The outputs of both branches are combined and passed to another layer of BiLSTMs, followed by a feed-forward network that predicts binary sentiment/sarcasm labels corresponding to the input after activating with the Sigmoid function. We follow a 10-fold cross-validation regime.

We compare the models with generated scanpaths, real scanpaths, and without scanpaths. Further, to investigate whether performance gains observed by adding scanpaths are due to scanpaths and not the increase in the number of parameters, we train a *Random-Random* variant in which we send Random noise as scanpaths to the model with an increased number of parameters. We also simulate the real-world case where both real and generated scanpaths are available during train

Model Configuration		F1 score	
Train	**Test**	**Sentiment**	**Sarcasm**
w/o	w/o	0.7839	0.9438
Random	Random	0.7990	0.9397
Random	Generated	0.7773	0.9313
Real	Generated	0.8319	0.9378
Real	Real	0.8334	0.9501
Generated	Real	0.8402	0.9452
Generated	Generated	0.8332	0.9506
Real + Generated	Generated	**0.8404**	**0.9512**
Intent-Aware	Intent-Aware	**0.8477**	**0.9528**

Table 4.4: Sentiment analysis and sarcasm detection results on the dataset by (Mishra *et al.*, 2016*a*). Model configuration refers to the type of scanpath included in train and test data.

time, but only generated ones are available during test time, for example, during user deployment.

Table 4.4 records the results of sentiment analysis and sarcasm detection tasks (Mishra *et al.*, 2016*a*). We note that generated scanpaths training and testing lead to similar gains for sentiment analysis and sarcasm detection as real scanpaths. The model with an increased number of parameters fed random noise in place of scanpaths performs similarly to the model trained without any scanpaths. Interestingly, the best results are obtained when model training uses both real and generated scanpaths. We believe this is due to ScanTextGAN bringing additional cognitive information from the news-reading CELER corpus, which is not present in the real scanpaths in (Mishra *et al.*, 2016*a*). In addition to the intrinsic evaluation presented in §4.1.4.2, this downstream evaluation demonstrates the high quality of the synthesized scanpaths, showing that they contain valuable cognitive processing signals for NLP tasks.

GLUE Tasks: To validate further, we augment classification models (based on sequential modeling using LSTMs) with generated scanpaths to show performance improvement in downstream NLP tasks on four GLUE benchmark datasets – SST, MRPC, RTE, QQP as described in §4.1.4.1. Table 4.5 reports the accuracy and weighted-F1 scores of the models trained with and without scanpaths for these tasks. We observe that in all four tasks, the model trained with generated scanpaths outperforms the one without scanpaths.

Dataset	Model	Acc	F1 score
SST	w/o scanpaths	0.8090	0.8089
	w/ random scanpaths	0.8059	0.8061
	w/ generated scanpaths	**0.8138**	**0.8138**
	w/ intent-aware scanpaths	**0.8269**	**0.8272**
MRPC	w/o scanpaths	0.6902	0.6656
	w/ random scanpaths	0.6623	0.6680
	w/ generated scanpaths	**0.6969**	**0.6828**
	w/ intent-aware scanpaths	**0.7009**	**0.6911**
RTE	w/o scanpaths	0.6162	0.6080
	w/ random scanpaths	0.5802	0.5794
	w/ generated scanpaths	**0.6211**	**0.6205**
	w/ intent-aware scanpaths	**0.6293**	**0.6278**
QQP	w/o scanpaths	0.8499	0.8513
	w/ random scanpaths	0.8491	0.8503
	w/ generated scanpaths	**0.8578**	**0.8596**
	w/ intent-aware scanpaths	**0.8648**	**0.8658**

Table 4.5: Results of training NLP models with and without scanpaths on the GLUE benchmark tasks. Including scanpaths leads to consistent improvements across all the NLP tasks.

Intent-Aware Scanpaths: Finally, we try to condition scanpaths generation on the downstream natural language task. We back-propagate gradients from the downstream NLP task to the conditional generator. In this fashion, the model learns to generate *intent-aware* scanpaths. The hypothesis is that finetuning scanpath generation based on feedback from the natural language task will bias the generator towards words more pertinent to that task and thus could help further improve performance on the downstream task. The architecture is shown in Fig 4.5. The results in Tables 4.4 and 4.5 validate the hypothesis that we observe consistent improvements in all downstream tasks. Fig 4.2 and Fig 4.6 show a few examples of scanpaths and saliency generated for three downstream natural language tasks.

Together these results corroborate the hypothesis that leveraging the cognitive signals approximated by synthetic scanpaths in NLP models leads to performance gains.

4.1.5 Intent-Aware Scanpaths

As described in section §4.1.4.3, the generator conditioned on the downstream natural language task yields *intent-aware* scanpaths. Augmenting NLP models with these scanpaths leads to higher performance gains. Here, we provide more details on *intent-aware* scanpath generation. Please refer to figures 4.5 and 4.6 on the following page. Saliency corresponding to intent-aware scanpaths are shown in Fig. 4.6.

4.1.6 Conclusion

In this work, we make two novel contributions toward integrating cognitive and natural language processing. (1) We introduce the first scanpath generation model over text, integrating a cognitive reading model with a data-driven approach to address the scarcity of human gaze data on text. (2) We propose generated scanpaths that can be flexibly adapted to different NLP tasks without needing task-specific ground truth human gaze data. We show that both advances significantly improve performance across six NLP datasets over various baselines. Our findings demonstrate the feasibility and significant potential of combining cognitive and data-driven models for NLP tasks. Without the need for real-time gaze recordings, the potential research avenues for augmenting and understanding NLP models through the cognitive processing information encoded in synthesized scanpaths are multiplied.

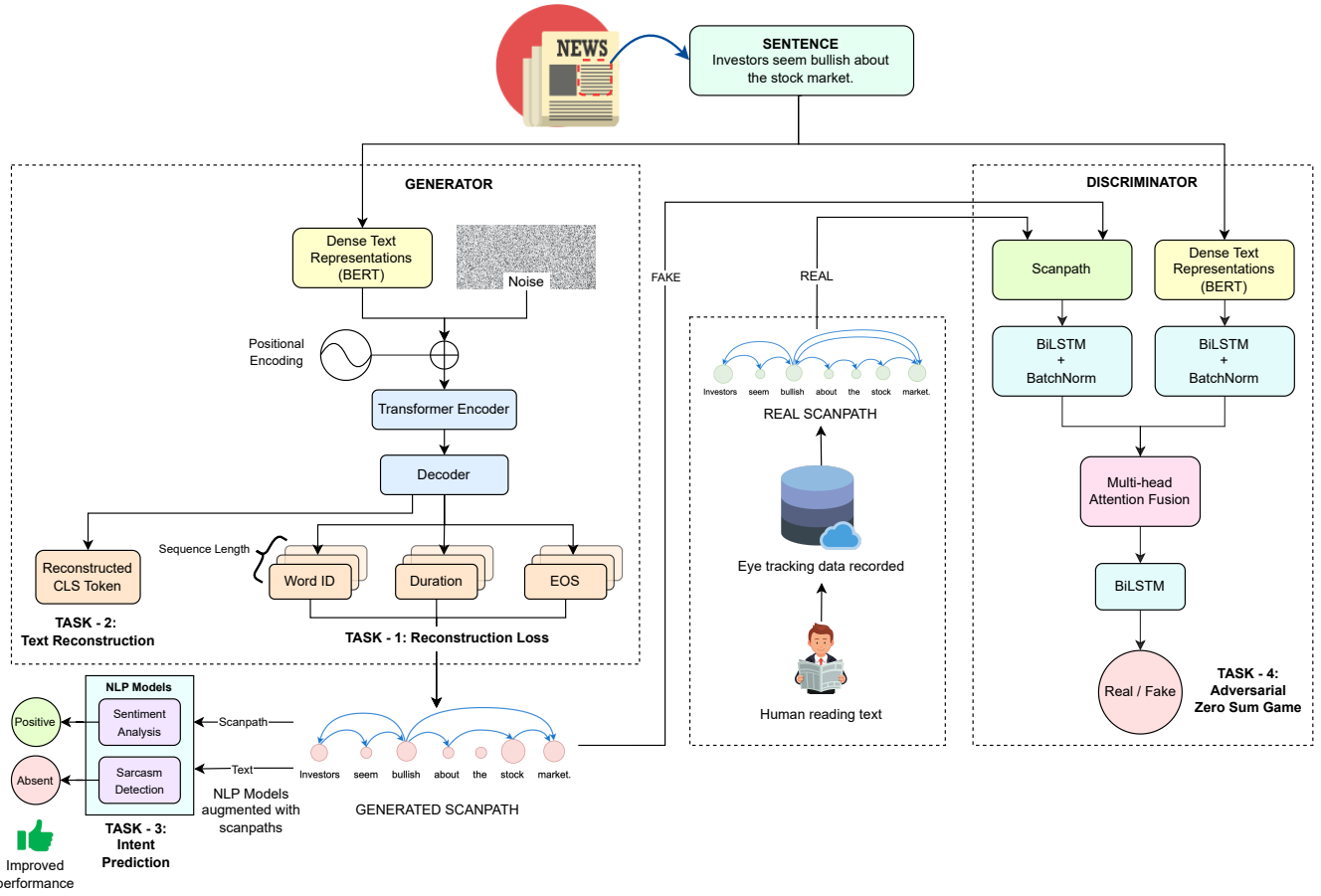


Figure 4.5: The architecture of the proposed Intent-Aware **ScanTextGAN** model. The model consists of a conditional generator and a discriminator playing a zero-sum game. Two cognitively inspired losses train the generator: scanpath (Task-1) and text (Task-2) reconstruction, a loss from the downstream intent of the natural language task (Task-3), and finally, the loss from the adversarial zero-sum game (Task-4). Variations of scanpaths are generated based on the downstream natural language task.

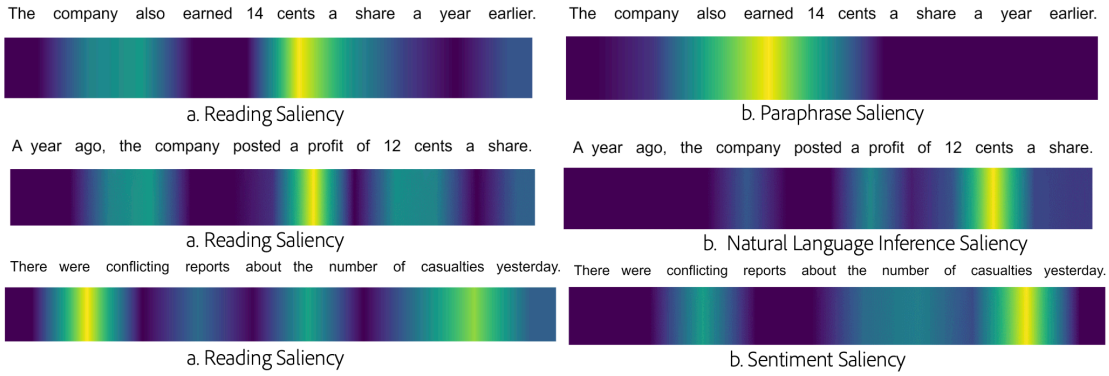


Figure 4.6: Saliency samples generated by conditioning scanpath generation on different downstream natural language tasks. It can be observed that the conditioned saliency pays much more attention to words important for that downstream task.

4.1.7 Limitations

In this work, we demonstrated artificial scanpath generation over multiple eye-tracking datasets. Further, our experiments build a link between cognitive and natural language processing and show how one can inform the other. However, the proposed method has a few limitations, which we aim to address in the future. The field needs work on bigger and more diverse eye-tracking datasets, which can enable scanpath generation over longer text sequences and can model generating scanpaths conditioned on previously read context. Besides, a better understanding of the entire scanpath generation process can help model the intra and inter-sentence scanpath generation process. The understanding would enable the integration of scanpaths to generative modeling tasks, which we intend to take up in future work. Another parallel direction is to include both explicit (like using RLHF) and implicit signals (like using cognitive signals) to better NLP tasks like language modeling.

Communication is defined as "*Who* says *what* to *whom* with ***what*** *effect*." A message from a communicator generates downstream receiver effects, also known as behavior. Receiver behavior, being a downstream effect of the message, carries rich signals about it. Even after carrying signals about the message, the behavior signal is often ignored while training vision language models. We show that training VLMs on receiver behavior can actually help improve their content-understanding abilities. We demonstrate that training VLMs to predict receiver behaviors, such as likes, comments, and replay graphs, which are available at scale,

enhances the VLM's performance across a broad range of downstream content understanding tasks. We show this performance increase over 6 types of behavior, 46 different tasks covering image, video, text and audio over 26 benchmark datasets across both 0-shot and fine-tuning settings, outperforming many supervised baselines on diverse tasks ranging from emotion recognition to captioning by upto 150%. We note that since receiver behavior, such as likes, comments, and replay graphs, is collected by default on the internet and does not need any human annotations to be useful, the performance improvement we get after training on this data is essentially free-lunch. We also release BLIFT, our Behaviour-LLaVA IFT dataset comprising of 730k images and videos with their receiver behavior collected from multiple platforms on which we train our models to achieve this.

4.2 Teaching Human Behavior Improves Content Understanding Abilities Of VLMs

Communication is defined as "*Who* says *what* to *whom* with ***what*** *effect*." A message from a communicator generates downstream receiver effects, also known as behavior. Receiver behavior, being a downstream effect of the message, carries rich signals about it. Even after carrying signals about the message, the behavior signal is often ignored while training vision language models. We show that training VLMs on receiver behavior can actually help improve their content-understanding abilities. We demonstrate that training VLMs to predict receiver behaviors, such as likes, comments, and replay graphs, which are available at scale, enhances the VLM's performance across a broad range of downstream content understanding tasks. We show this performance increase over 6 types of behavior, 46 different tasks covering image, video, text and audio over 26 benchmark datasets across both 0-shot and fine-tuning settings, outperforming many supervised baselines on diverse tasks ranging from emotion recognition to captioning by upto 150%. We note that since receiver behavior, such as likes, comments, and replay graphs, is collected by default on the internet and does not need any human annotations to be useful, the performance improvement we get after training on this data is essentially free-lunch. We also release BLIFT, our Behaviour-LLaVA IFT dataset

comprising of 730k images and videos with their receiver behavior collected from multiple platforms on which we train our models to achieve this.

4.2.1 Introduction

Communication is defined by five factors: sender, message, channel, receiver, and behavior (Shannon and Weaver, 1949; Lasswell, 1948, 1971). Lasswell (1948) encoded these five factors in the phrase, "*Who* says *what* to *whom* with *what effect*." Human behavior occurs as a downstream artifact in the process of communication. Behavior is produced by the receiver as a response to the message sent by the sender. Being a downstream effect, behavior can help us infer important signals about the message itself. These signals, if properly harnessed, should be able to increase performance on the message understanding tasks popular in NLP and CV, like question answering, sentiment analysis, topic classification, *etc.* Despite this, behavior data is considered noise and is ignored while training large language models (Biderman *et al.*, 2022; Penedo *et al.*, 2023) and also large vision and language models (Liu *et al.*, 2023*a*; Zhu *et al.*, 2023). In this paper, we explore this line of thought more.

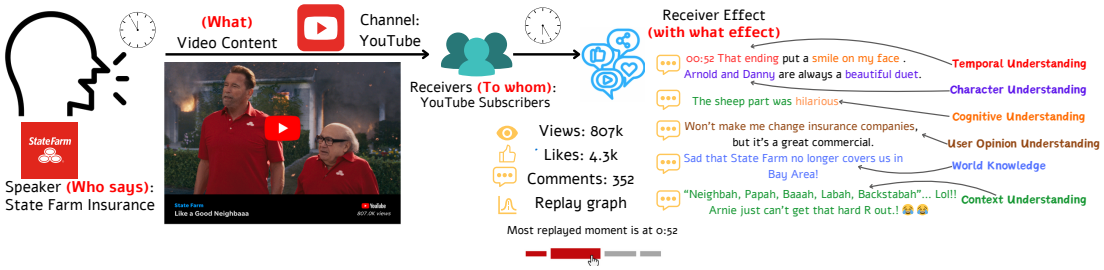


Figure 4.7: The diagram depicts the five factors of communication in the context of an example YouTube video `https://www.youtube.com/watch?v=eT8hO4e2iTM` and where lies the free lunch. The receiver effect is not used while training Large Vision and Language Models. However, it contains many important signals that can help in understanding the content. The figure shows several comments containing temporal, cognitive, character, context, and user opinion information useful for understanding the video.

Humans produce two kinds of behavioral signals upon observing a message (Bertenthal, 1996; Prinz, 1997): perceptual signals and actions as behavior. Perceptual signals, like seeing, touching, and hearing, help a receiver primarily sense the world around her, ultimately guiding her actions. Actions are how a receiver

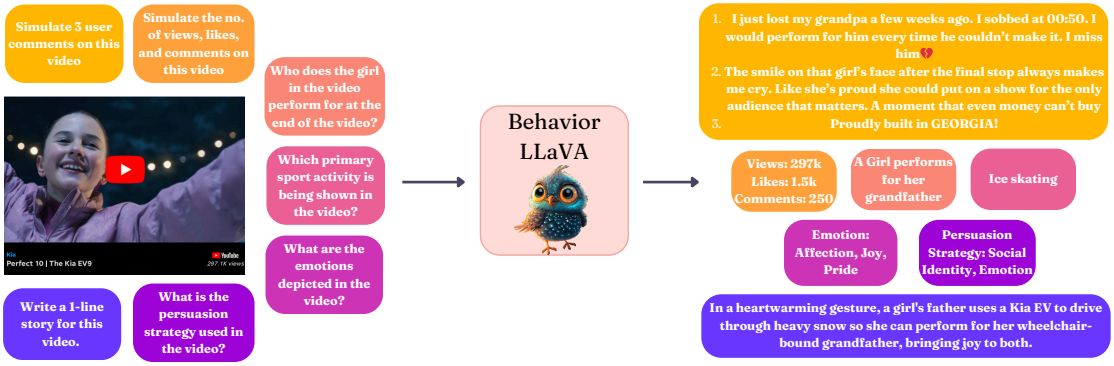


Figure 4.8: Behavior-LLaVA is trained to answer behavioral questions like simulating user comments and likes on the video. The model, once trained, shows superior performance than LLaMA-Vid and other VLMs on content-related tasks like emotion recognition, action recognition, question answering, persuasion strategy classification, *etc.* The original video was showcased in SuperBowl-2024 and is posted on YouTube on the URL `https://www.youtube.com/watch?v=0U7BJc961I4`. The video is titled "Perfect 10: The Kia big game commercial featuring the 2024 Kia EV9" by Kia America.

acts on the outside world. The signals produced by the human receiver upon receiving a message carry information about the message itself (Fig. 4.7). For instance, if a person's heartbeat rises upon watching a movie scene, it can help us infer that perhaps the scene was an exciting scene (Dzedzickis *et al.*, 2020). Similarly, regressing while reading is indicative of important or confusing phrases (Bicknell and Levy, 2011). In these cases, perception behavior helps us derive inferences about content. In a similar vein, the actions a person performs after watching a movie, such as comments and likes, carry signals about the movie (Fig. 4.7, 4.8).

Expanding on these ideas, prior literature has shown that harnessing perceptual signals, like eye movements, saliency, keystrokes, mouse movements, and FMRI, by modeling them together with content understanding tasks can improve both NLP and CV tasks. For instance, integration with perception signals causes performance improvement in tasks like visual and natural language question answering (Patro and Namboodiri, 2018; Khurana *et al.*, 2023; Sood *et al.*, 2020), text and image sentiment analysis (Khurana *et al.*, 2023; Barrett *et al.*, 2018*a*; Fan *et al.*, 2018), natural language inference (Khurana *et al.*, 2023), part-of-speech identification (Barrett *et al.*, 2016*a*,*b*), named entity recognition (Hollenstein *et al.*, 2019; Hollenstein and Zhang, 2019*b*), syntactic parsing (Plank, 2016*a*), image cap-

tioning (Cornia *et al.*, 2018), and visual object detection (Wang *et al.*, 2018*c*; Kruthiventi *et al.*, 2016).

While the initial studies show that perceptual signals have much promise for improving downstream content understanding, they have a few significant issues due to which integrating human perception has not seen wide adoption in training LLMs. These perceptual signals can only be collected in lab settings requiring specialized lab equipment and are thus expensive to collect and thus are also limited in number. For example, the largest datasets containing the human processing signals are SALICON (Jiang *et al.*, 2015) and Cheng *et al.* (2014*b*) for visual saliency (10k images each), CELER (Berzak *et al.*, 2022) and Dundee corpus (Kennedy *et al.*, 2013) containing eye movements over 28k sentences and 20 news articles respectively, and Dhakal *et al.* (2018) containing keystroke patterns over 1.5k sentences. Clearly, these datasets, while making important contributions, do not scale to the level at which today's large language models are trained (trillions of natural language and image tokens).

On the other hand, actions (the other type of behavioral signals produced by a human receiver) are collected at a large scale in the form of digital analytics. Examples of this kind of data are likes, views, shares, comments, and purchase histories on images, tweets, videos, webpages, and other kinds of media. Action data has a much broader representation than is possible in lab settings, is available on more diverse content, and is much cheaper to collect than using specialized lab equipment. At the same time, actions have not been much investigated in the literature for their potential to improve downstream content understanding.

Therefore, in this paper, we make initial efforts to collect and understand digital analytics at scale with the aim of integrating them with VLMs to improve their downstream content understanding capabilities. We introduce methods for filtering and cleaning behavioral data and then propose tasks for large language and vision models, leading to improvements in language and visual content understanding tasks. For this, we look to Reddit and YouTube as two major sources of visual content and human behavior in the form of viewer comments, likes, replay graphs, and upvotes. From Reddit, we collect 5 million images and videos along with their upvotes and top-upvoted comments from two major subreddits (r/pics

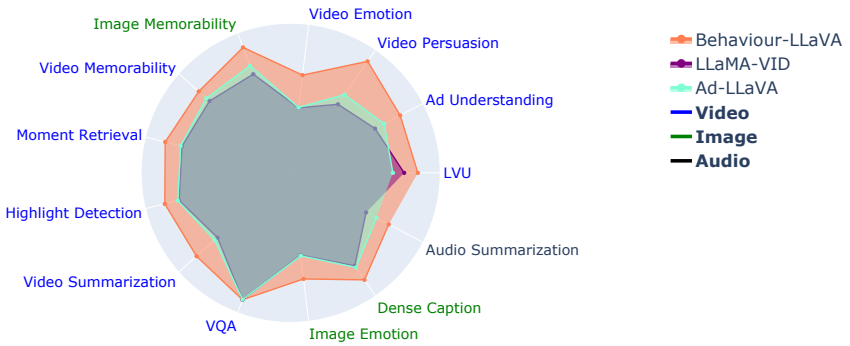


Figure 4.9: Behaviour-LLava achieves much higher zero-shot performance compared to Ad-LLaVA and the base model LLaMA-VID across a diverse suite of image, video, and audio benchmarks.

and r/videos). Similarly, from YouTube, we collect 2.2 million videos from 30000 channels along with their likes, views, replay graphs, and top user comments. After extensive filtering and cleaning, we are left with 730k samples of videos and images across the two platforms which we use for the next steps.

After collecting user behavior over image and video content, we design tasks to teach large vision and language models (VLMs) to simulate user behavior. For this, we use an instruction fine-tuning format. Given a video or an image and the other metadata like time of post and channel, we ask the model to simulate user behavior of likes and comments. See Fig 4.8, Listing 4.1 for examples. We choose LLaMA-Vid (Li *et al.*, 2023*d*) as our base model to teach it the user behavior. We call the resultant model Behavior-LLaVA (Large Language and Vision Assistant) (Liu *et al.*, 2023*a*). We test Behavior-LLaVA on a diverse variety of tasks, evaluating its capabilities on image, video, text, and audio understanding tasks. We compare Behavior-LLaVA against its base model, LLaMA-Vid, and other supervised baselines. Further, to show the impact of behavior, we train another version of LLaMA-Vid, where we train it on the same set of videos and images as Behavior-LLaVA but do not include behavior information. We call this model Ad-LLaVA.

We make the following contributions with this work:

1) **Behavior-LLaVA Instruction Fine-Tuning:** We explore the idea of learning human behavior, resulting in better content understanding. We test this for action-level behavior data such as receiver comments, likes, and replay graphs.

We collect a dataset called **BLIFT**, consisting of 400k images and 330k videos, along with their receiver behavior. Then, LLaMA-Vid is trained for the task of predicting receiver comments and upvotes given a media (a video or an image) (Listing 4.1). We show that using this simple task formulation over behavioral data collected in the wild, results in performance improvement over a hierarchy of tasks. We get improvements over the base LLaMA-Vid across 46 tasks over 26 benchmark datasets in both zero-shot and fine-tuned settings. We show this over low-level content understanding tasks like object and activity recognition and also over high-level tasks like topic and emotion detection. Through this, we propose a scalable approach to increase the content understanding abilities of VLMs, requiring minimal cost and no architectural changes.

2) **AdLLaVA: Disentangling the effect of content and behaviour:** To disentangle the effect of training LLaMA-Vid on additional image and video data from the effect of training on behavior data, we train LLaMA-Vid on BLIFT's videos and images without including behavior. We call this model Ad-LLaVA. We show that Ad-LLaVA shows equivalent performance as its base model LLaVA-Vid; however, Behavior-LLaVA performs better than both Ad-LLaVA and LLaMA-Vid, thus highlighting the importance of behavior data and instruction fine-tuning on behavior data.

3) **Perception vs Action:**We also show an ablation of Behavior-LLaVA across different kinds of behavior. We try out the perception behavior of saliency prediction over images and five types of action-level behavior over images and videos. We find that perception-level behavior does not result in significant performance improvements; however, action-level behavior shows improvements across all the tasks. We posit that one reason for this could be due to the scale for which action-level data is available (Table 4.6). While perception behavior is mostly collected in lab settings, action-level behavior data is diverse, can be collected in a scalable manner automatically and cheaply.

4.2.2 Methodology

In this section, we introduce our approach to train Behavior-LLaVA. Since no publicly available corpus consists of behavior together with image and video content, we first introduce our instruction fine-tuning dataset, "Behavior-LLaVA Instruction Fine-Tuning dataset" (BLIFT). Next, we introduce our methodology to train Behavior-LLaVA. Finally, we report the results of testing Behavior-LLaVA's capabilities on a hierarchy of tasks. The tasks cover low-level media understanding tasks like object and activity detection, high-level media understanding tasks like emotion, topic, and persuasion strategy classification.

4.2.2.1 BLIFT Dataset

Given the abundance of media and behavioral data and its accessibility, our data collection relies on two primary sources: Reddit and YouTube. These platforms share similarities in terms of hosting media content (images and videos) and providing user engagement metrics in the form of Reddit upvotes and comments, and YouTube likes, views, comments, and replay graphs. Here, we delineate the process involved in constructing the instruction fine-tuning dataset, which we term as the Behavior-LLaVA Instruction Fine-Tuning (BLIFT) dataset.

4.2.2.1.1 Data from Reddit To collect a substantial corpus of diverse images and videos, we targeted at two specific subreddits, namely r/pics and r/videos. Established over 15 years ago, these subreddits had a user base exceeding 20 million during the data collection period, with an average of over 5,000 online users concurrently. Notably, due to stringent content moderation guidelines (Reddit, Inc., 2024; red, 2024*a*,*b*,*c*) and the exclusive focus of these subreddits on media content, they offer a rich variety of content devoid of thematic biases. Our data collection spans until January 2022, during which the activity on these subreddits witnessed a notable decline following several policy adjustments and user protests (Hern, 2023; Economist, 2023).

To ensure data quality and relevance, we executed a series of filtering steps on the posts and comments from these subreddits. Initially, we excluded posts

predating February 2015 from r/pics, coinciding with the implementation of a rule requiring images without digital/overlay text (red, 2024*a*,*b*). This filtering step resulted in the exclusion of 3.1 million images and 2 million videos. Subsequently, considering the sustained popularity of both subreddits, with rankings within the top 20 since 2017 and consistent membership exceeding 20 million, we confined our dataset to posts from January 2018 onwards. This selection process yielded 1.4 million images and 1.1 million videos.

Further refinement of the dataset involved removing posts and comments marked as NSFW, BOT-generated, or [deleted], along with eliminating duplicate images and videos. This curation step reduced the dataset to 876,000 images and 983,000 videos. To address redundancy in comments, we excluded those comprising fewer than three words and employed TF-IDF-based deduplication with a similarity threshold of 0.6, determined through manual observations.

Following these steps, posts with fewer than two comments were filtered out, resulting in a dataset comprising 631,000 images and 397,000 videos. Additionally, videos exceeding a duration of 500 seconds were omitted, leaving 221,000 videos for analysis. Notably, images not directly hosted on Reddit were excluded due to scraping and copyright limitations. Similarly, for r/videos, only videos hosted on YouTube were considered. It is pertinent to mention that approximately 51% of YouTube videos collected during this period were either made private/unlisted or removed, resulting in 400,000 images and 80,000 videos, accompanied by 1.5 million and 312,000 comments, respectively. These comprehensive filtering steps ensured the construction of a diverse and relevant dataset for fine-tuning instruction-based models.

4.2.2.1.2 Data from YouTube Our data collection from YouTube begins with querying Wikidata (Vrandečić and Krötzsch, 2014) for YouTube IDs to compile a list of channels. Wikidata, derived from Wikipedia, provides a curated selection of renowned channels, automatically filtering out noisy videos commonly found in datasets collected from diverse sources like user-generated videos. This initial step yielded a dataset of 2.2 million videos spanning the period from 2018 to 2023, sourced from approximately 6,000 channels collected from Wikidata.

Listing 4.1: Behavior Instruction fine-tuning template for the video: https://www.youtube.com/watch?v=BKPQkjRF4yY




```
<SYSTEM>You are an AI visual assistant. You are given a detailed description of a media,
      followed by the actual media. Answer all questions as if you are seeing the media.</
      SYSTEM>

<USER>
The video advertisement is titled "Gatorade | Make Your Own Footsteps with Suni Lee" for the
       brand Gatorade. The audio in the ad says "[ASR HERE ...]". Analyze this video deeply,
       then write scene by scene description of the video and answer the following questions
      . What percentage of viewers would like this video, and what would be the top-5
      popular comments on this video? What would the replay graph values for each scene be?
<video>...</video>
</USER>

<ASSISTANT>
The scene-by-scene descriptions are:

Scene 1: The scene shows a woman looking off into the distance with an orange line going
      around the outside of the screen. The foreground colors of the scene are black, mud
      green, gray, dark gray, and the background colors are dark brown, black, dark gray.
      The dominant tone of the scene is neutral. This scene is categorized by the tags:
      cosmetic, eyebrow, face, girl, ponytail, stand, string, woman.
Scene 2: The scene shows a woman balancing on a skateboard in a yard. The foreground colors
      of the scene are black, mud green, dark gray, olive, and the background colors are
      black, dark gray, gray, dark brown.The dominant tone of the scene is neutral. This
      scene is categorized by the tags: athletic, balance, beam, car, girl, house exterior,
      hurdle, jog, legging, plank, rail, seesaw, woman, yard.
Scene 3: The scene shows a girl jumping over a wooden ramp in the backyard. The foreground
      colors of the scene are black, dark gray, gray, dark blue, and the background colors
      are dark brown, dark blue, purple, dark pink, brown.
...

>>> BEHAVIOR <<<

The video will be liked by 2.0% of viewers, and the popular comments could be:
1. "Wow. Love it. She's such an inspiration to the next generation as well as everyone."
2. "Inspiring and great story behind this commercial. Builds meaning and purpose in the
      hearts and minds of youth. It's been a while since good, meaningful ads have been made
      ."
3. "She's an inspiration to the world. Thanks to her,my niece is learning gymnastics.
      Hopefully someday, she is an inspiration to others as Suni is an to everyone"
4. "Chills watching this. Such an inspiration."
5. "Yooooo, this is straight up!"

The replay values for each scene would be:
Scene 1: 0.06
Scene 2: 0.23
Scene 3: 0.38
...
</ASSISTANT>
```

Figure 4.10: Behavior Instruction fine-tuning template for the video: https://www.youtube.com/watch?v=BKPQkjRF4yY

To refine the dataset, manual filtering was employed to exclude certain categories deemed less relevant for our purposes. These categories included music and songs, gaming content, non-English videos, sports commentary, anime, memes, channels with disabled comments sections, and news-related content. Furthermore, and only videos with a substantial viewership, defined as greater than 10,000 views, were retained. We observed that these videos usually have less noisy comments and likes.

Subsequently, the top comments from each video, as ordered by YouTube (i.e., the most liked comments), were selected for inclusion in the dataset. To address redundancy in comments, a TF-IDF filter was applied with a threshold of 0.7, which proved effective in removing duplicate comments prevalent in YouTube data.

Comments were further filtered to include only those with a minimum of four words and a maximum of 100 words, ensuring a balance between relevance and

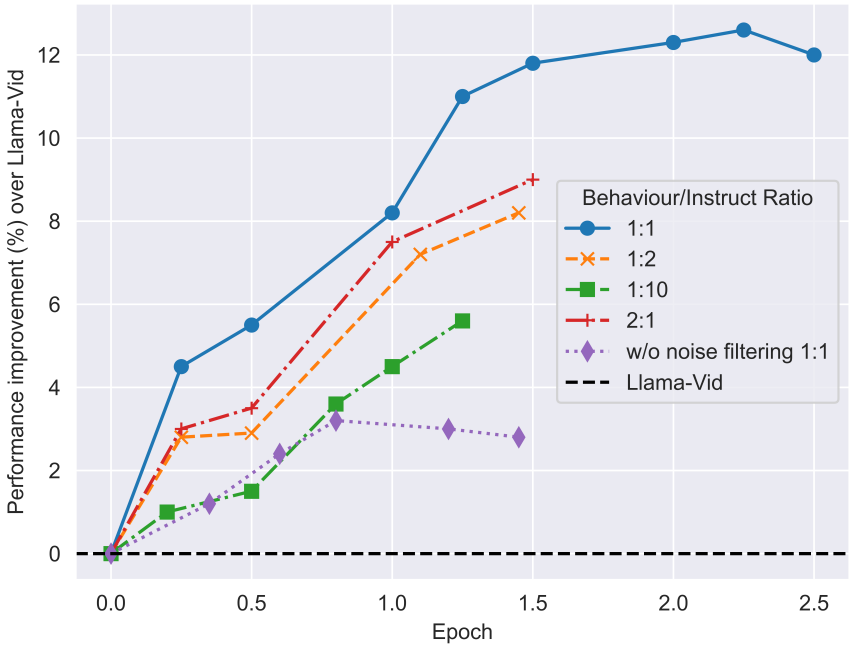


Figure 4.11: Percentage performance improvement over an untrained LLaMA-Vid model, compared across various sampling ratios at different checkpoints. The 1:1 sampling ratio shows the best empirical performance. Performance is averaged over 0-shot accuracy improvements on six tasks with 250 samples each from the evaluation set. These tasks include image emotion recognition, video emotion recognition, and persuasion strategy classification, MSRVTT, HVU and MSVD-QA. The figure also shows the benefit of our data filtering process. While training on unfiltered BLIFT also improves the result over baseline performance of Llama-Vid but data filtering adds more improvement on top of it.

conciseness. Additionally, to mitigate the presence of NSFW content, a vocabulary specific to NSFW terms (ldn) was employed to filter out inappropriate posts. On average, we finally get 3.1 comments per video, providing a substantial corpus of user-generated content for analysis. After applying these filtering steps, the dataset was reduced to 250,000 videos, ensuring a curated and relevant collection for subsequent analysis and model training.

4.2.2.2 Instruction Fine-Tuning LLaMA-Vid

After collecting Reddit and YouTube media and user behavior, we formulate instruction fine-tuning tasks for training LLaMA-Vid. In the training instruction, given the media content and automatic speech recognition if available, we ask the

model to simulate the scene-by-scene description and user likes/views and top-5 comments. This instruction training template is given in Listing 4.1. To generate the instruction data, first, frames are sampled using the 30-degree rule (SI *et al.*, 2025; Arev *et al.*, 2014; Friedman and Feldman, 2004), then the scene-by-scene description are obtained by concatenating automatically generated captions and tags from LLaVA-13B (Liu *et al.*, 2023*a*), colors and tone through Qin *et al.* (2020*b*). This instruction format keeps the instructions similar to the instruction format for other VLMs like LLaVA (Liu *et al.*, 2023*a*), MiniGPT-4 (Zhu *et al.*, 2023), BLIP (Li *et al.*, 2022), LLaMA-Vid (Li *et al.*, 2023*d*), *etc.*, while additionally teaching the model to learn behavior. We keep the instruction fine-tuning template similar for both YouTube and Reddit. The complete instruction is given in Listing 4.1.

We start with the trained LLaMA-Vid model. The LLaMA-Vid model uses two tokens to represent each frame in the video, which they call content and context tokens. While the context token encodes the overall image context based on user input, the content token encapsulates visual cues in each frame. For learning context tokens, the model uses attention queries that interact with previously generated image features in the designed attention module. To generate content tokens, the image features are average pooled. This dual-token strategy significantly reduces the number of tokens needed to represent videos, thus enabling the model to scale to longer (hour-long) videos. To better support hour-long videos, LLaMA-Vid was trained on a 9k movie-level conversation instruction set containing plot reasoning and detail understanding questions.

Taking the base LLaMA-Vid model, we finetune it further on behavioral data. In our experiments we observe that the sampling ratio of BLIFT and IFT datasets is an important hyperparameter. We track 4 zero-shot metrics, likes/views, comments perplexity, and empirically find the best results with 1:1 ratio for 2.2 epochs (see Fig 4.11 and Table 4.7). For the best checkpoint, the perplexity on comments reduces from 6.22 to 3.05, and the R2 on likes/views goes from -5.1 to 0.45

We combine 730k instruction pairs from BLIFT with the original instruction tuning dataset consisting of 40K text conversations from ShareGPT, 625K single or multi-turn visual QA pairs, and 98K video QA pairs; all the modules except the Visual Encoder are kept frozen. We ablate on multiple sampling ratios from

Task	MSRVTT-QA	CAER	Emoset	Comments	likes/views
Base	58.9	75.6	45.23	6.22	-0.1
Salicon [Region]	55.6	75.8	47.25	6.25	-0.07
Salicon [Object]	57.9	76.4	48.0	6.12	0.05
BLIFT[Likes/Views]	58.2	76.2	47.12	6.15	0.38
BLIFT[Titles]	58.4	78.1	48.12	5.09	0.13
BLIFT[Comments]	59.0	79.1	49.58	3.02	0.19
BLIFT	59.2	79.3	50.38	3.05	0.40

Table 4.6: Ablation on using comments and/or perception signals from Salicon

BLIFT. We train the LLaMA-Vid checkpoints with their original SFT mix along with BLIFT. We ablate different sampling ratios and found 1:1 to be empirically performing the best. We train the model for 2.2 epochs, keeping track of the 0-shot evaluation metrics and perplexity on comments in the eval set Fig 4.11 and Table 4.7 show the ablations on different sampling ratios and epochs of training. For the best checkpoint, the perplexity on comments reduces from 6.22 to 3.05, and the R^2 on likes/views goes from -5.1 to 0.45.

AdLLaVA to show the impact of behavior data: To disentangle the effect of training on additional data samples from the effect of training on behavioral data, we train LLaMA-Vid on BLIFT with the video and image verbalization and do not include receiver behavior. Then, the overall instruction template consists of scene-by-scene automatically generated verbalization similar to Listing 4.1 without the likes and comment simulation. We call the LLaMA-Vid fine-tuned on this data, Ad-LLaVA. We compare Behavior-LLaVA with Ad-LLaVA and LLaMA-Vid along with other state-of-the-art literature benchmarks on various tasks (Tables 4.8-4.13).

Impact of filtering steps on performance: We use various filtering steps in our data pipeline, including NSFW filters, time filters, bot filters, and gaming and news video filters. To quantify the impact of our filtering steps on the final performance, we compare the performance from training on unfiltered data with filtered data (Figure 4.11). The figure shows the benefit of our data filtering process. While training on unfiltered BLIFT also improves the result over baseline performance of Llama-Vid but data filtering adds more improvement on top of it

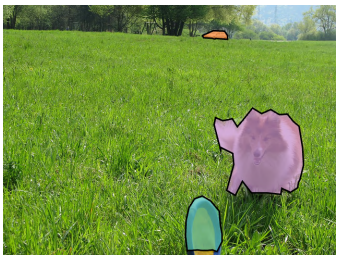
Ablation with perceptual behavior: As an ablation experiment, we also try teaching the Behavior-LLaVA perceptual signals. For this, we take the largest

Sampling Ratio	Epoch	Likes/Views R^2	Comments Perplexity	Performance
Base-Model	0	-0.1	6.22	0
1:1	0.5	0.11	4.71	5.49
	1	0.22	3.95	8.23
	1.25	0.33	3.19	10.97
	1.5	0.35	3.13	11.79
	2	0.38	3.08	12.31
	2.2	0.4	3.05	12.57
1:2	0.5	0.14	4.33	3.04
	1.05	0.28	3.66	7.08
	1.45	0.42	2.99	8.12
1:10	0.5	0.15	3.43	1.38
	0.8	0.31	2.78	3.44
	1	0.38	2.46	4.48
	1.2	0.46	2.13	5.51
2:1	0.5	0.1	5.3	3.52
	1	0.21	4.6	7.45
	1.5	0.31	3.9	9.13

Table 4.7: Ablation on different sampling ratios and epochs of training. Sampling ratio is the ratio of behaviour data to multimodal instruct data. Performance is the average increase in 0-shot accuracy on 6 tasks with 250 samples each from the eval set. These tasks include image emotion recognition, video emotion recognition, persuasion strategy classification, MSRVTT, HVU and MSVD-QA

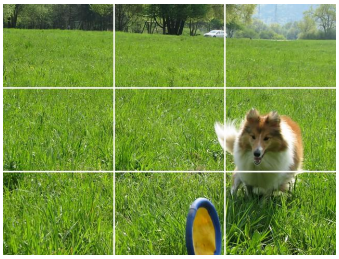
perception signal dataset in the literature - Salicon10k (Jiang *et al.*, 2015). It consists of 10,000 MS COCO images (Chen *et al.*, 2015) with free-viewing eye gaze data collected through a novel mouse-based interface. The dataset has been widely used in many studies. We formulate two tasks using this data, (1) **Salicon [Object]**: estimating the saliency over the objects in the image and (2) **Salicon [Region]**: estimating the saliency over a region, where the regions tiles formed by breaking the image into a 3x3 grid. For both tasks we try to model two objectives, ranking and predicting, we found ranking to be much more effective. The instruction are given in Listings 4.2 and 4.3.

Listing 4.2: Perceptual Signal Instruction fine-tuning template for the image: http://farm6.staticflickr.com/5106/5670500150_e035dd2d30_z.jpg



<SYSTEM>You are an AI visual assistant. Answer all questions as you are seeing the media<SYSTEM><USER>The objects in this image in no particular order are car, dog, frisbee. Give me the order of saliency of these objects, start with the most salient object and end with the least salient object, each in a separate line. Give me the objects only and nothing else.
<image>
<ASSISTANT>
dog
frisbee
car
<ASSISTANT>

Listing 4.3: Perceptual Signal Instruction fine-tuning template for the image: http://farm6.staticflickr.com/5106/5670500150_e035dd2d30_z.jpg



<SYSTEM>You are an AI visual assistant. Answer all questions as you are seeing the media<SYSTEM><USER>Assume the given image is broken into a 3X3 grid the regions or tiles being named "upper−left" "upper−center", "upper−right", "middle−left", "middle−center", "middle−right", "bottom−left", "bottom−center", "bottom−right". Rank these regions or tiles based on their saliency, give me the line separated ranking of all regions in decreasing order.
<ASSISTANT>
middle−right
bottom−center
bottom−right
upper−center
upper−right
middle−center
upper−left
middle−left
bottom−left

Model	scene	way_speaking	relationship	like_ratio	view_count	director	genre	writer	year
Video4096-GPT-3.5 generated story + Flan-t5-xxl	60.2	39.07	64.1	0.061	12.84	69.9	58.1	52.4	75.6
Video4096-GPT-3.5 generated story + GPT-3.5 classifier	54.54	32.95	68.42	0.031	12.69	75.26	50.84	32.16	75.96
LLaMA-Vid + GPT-3.5 Generated Story	58.12	35.5	60.6	0.314	10.34	65.34	49.77	34.23	72.12
Ad-LLaVA	59.05	37.07	61.2	0.319	10.37	66.84	55.13	35.33	77.34
Behavior LLaVA + GPT-3.5 Generated Story	66.43	41.03	64.21	0.17	5.12	71.12	63.45	39.4	79.3
Improvement of Behavior LLaVA over LLaMA-Vid	10.48%	15.58%	9.62%	45.86%	50.48%	8.85%	27.49%	15.1%	9.96%

Table 4.8: Comparison of various models on the Long Video Understanding benchmark (Wu and Krahenbuhl, 2021) consisting of 9 VQA tasks. We see that Behavior-LLaVA improves on LLaMA-Vid on 9/9 tasks with an average improvement of 21.49%. Further, it outperforms the state-of-the-art in 5/9 tasks.

4.2.3 Results and Discussion

In the experimental results, we aim to showcase the diverse and emergent capabilities of our Behavior-LLaVA model through quantitative numbers on various tasks and qualitative examples. These abilities include generating detailed image and video descriptions, emotion and sentiment analysis, question answering, video understanding tasks like scene and action detection. Additionally, we present the ability of Behavior-LLaVA to transfer learn on other behaviors like memorability of a video - both short-term and long-term.

4.2.3.1 Evaluation

To test the effectiveness of Behavior-LLaVA, we conduct experiments involving 46 distinct tasks across 26 benchmark datasets. The diversity of tasks and datasets allows us to evaluate the performance and capabilities of Behavior-LLaVA thoroughly. Each of them is covered briefly next:

1. **Visual Question Answering (VQA)**: We evaluate the performance of visual question answering on the following benchmark datasets:

- The Long-Video Understanding (LVU) benchmark by Wu and Krahenbuhl (2021) comprises nine distinct tasks aimed at assessing long video comprehension, incorporating over 1000 hours of video content. These tasks encompass diverse aspects such as content understanding (including relationship, speaking style, scene/place), prediction of user engagement (YouTube like ratio, YouTube popularity), and movie metadata (director, genre, writer, movie release year).
- The Holistic Video Understanding (HVU) dataset by Diba *et al.* (2020) stands as the largest dataset for long video comprehension, comprising 572,000 samples. Encompassing a broad spectrum of semantic elements within videos, HVU tasks involve the classification of scenes, objects, actions, events, attributes, and concepts. Performance evaluation on HVU tasks is conducted using the mean average precision (mAP) metric on the validation set.
- We also use MSVD-QA, MSRVTT-QA (Chen and Dolan, 2011; Xu *et al.*, 2016*b*), and ActivityNet-QA (Caba Heilbron *et al.*, 2015) datasets. Their description is given in Appendix 4.2.6.2.

2. **Video and Image Understanding Benchmarks**: We use a wide variety of tasks to evaluate video and image understanding: topic, emotion, and persuasion strategy classification, action and reason retrieval and generation, and emotions. We briefly introduce the benchmarks:

- The advertisements dataset by Hussain *et al.* (2017) contains 3,477 video advertisements and the corresponding annotations for emotion and topic tags and action-reason statements for each video. There are a total of 38 topics

and 30 unique emotion tags per video. Further, we have 5 action-reason statements for each video for the action-reason generation task.

- Persuasion strategy dataset (Bhattacharyya *et al.*, 2023) is a dataset consisting of 1002 video advertisements from popular brands and their persuasion strategy labels like social identity, anchoring and comparison, reciprocity, foot-in-the-door, *etc.*
- For emotion analysis, we use VideoEmotion-8 (Asur and Huberman, 2010), Ekman-6 (Xu *et al.*, 2016*a*), CAER (Lee *et al.*, 2019), IAPSa (Mikels *et al.*, 2005), Emotion6 (Peng *et al.*, 2015), EmoSet (Yang *et al.*, 2023), and Abstract (Machajdik and Hanbury, 2010) datasets. A brief description for each of them is given in Appendix 4.2.6.2.

3. **Image Dense Captioning**: Literature image captioning datasets such as MS-COCO (Chen *et al.*, 2015) reduce the inherently rich information and fine-grained semantics to simplistic captions, with very brief statements focussing only on salient objects. Behavior data such as user comments help a model learn much more information such as object and material properties, world knowledge, emotion, character understanding, spatial relationships, aesthetics, *etc.* (see Fig. 4.7), enhancing the model's captioning capability. Therefore, we design a captioning task to test this capability and compare it with respect to LLaMA-Vid and LLaVA-34B (a 2.5x larger model). Since we do not have ground truth for this task, following the LLM-as-a-judge paradigm, we use GPT-4V as the judge for all the models. GPT-4V is asked to evaluate the dense captions on three metrics: *Correctness* (Listing 4.4) evaluating the factuality and model hallucinations, *Detail* (Listing 4.5) evaluating the number and depth of details captured by the generated captions, and *Quality* (Listing 4.6) measuring the subjective quality of the concepts chosen to be highlighted by the captioning model and the arrangement, coherence, and the linking of various concepts.

4. **Image and Video Memorability Simulation**: Behavior-LLaVA is trained on behavior along with the media. To check if training on behavior helps in solving other behavior tasks (Khandelwal *et al.*, 2024), we test it over image and video memorability simulation. For this, we select seven benchmark datasets covering long-term and short-term memorability over images and videos: LaMem (Khosla *et al.*, 2015), SUN (Isola *et al.*, 2011), and MemCat (Goetschalckx and Wagemans, 2019) for images and Memento10k(Newman *et al.*, 2020), VideoMem (Cohendet *et al.*, 2019), MediaEval (Kiziltepe *et al.*, 2021*a*), and LAMBDA (SI *et al.*, 2025) for videos. We briefly cover each of them in Appendix 4.2.6.2.

5. **Modalities other than videos and images:** Behavior-LLaVA, built on top of LLaMA-Vid and fine-tuned using BLIFT, is pretrained and fine-tuned on image and video datasets. To test if behavior data can improve the results on other modalities as well, we test Behavior-LLaVA's performance on two tasks across audio and text modalities (Table 4.10). For audio, we evaluate on the audio summarization task (Han *et al.*, 2023) and for text, we evaluate on the IMDB sentiment benchmark (Maas *et al.*, 2011).

For Tables 4.8, 4.17, and 4.11, we follow the evaluation protocol of Video-4096 (Bhattacharyya *et al.*, 2023), for Table 4.16 we follow the evaluation protocol of

Training	Dataset	Video Emotion-8	CAER	Ekman-6
Random	Random	12.5	14.28	16.67
0-Shot	LLaMA-Vid	29.7	27.2	37.33
	Behavior-LLaVA	41.35	51.0	49.33
	Ad-LLaVA	29.8	27.3	37.66
Improvement of Behavior-LLaVA over LLaMA-Vid		39.22%	84.19%	32.14%
Finetuned	Zhao *et al.* (2020)	54.5	78.3	55.3
	Zhang *et al.* (2023*a*)	57.3	80.1	58.2
	eMOTIONS (Wu *et al.*, 2023*a*)	-	-	53.12
	Arevalo *et al.* (2017)	53.7	77.3	54.2
	Qiu *et al.* (2020)	53.3	-	57.3
	Xu *et al.* (2016*a*)	52.6	77.9	55.6
	LLaMA-Vid	53.8	75.6	57.9
	Ad-LLaVA	54.1	76.1	57.8
	Behavior-LLaVA	56.9	79.3	58.4
Improvement of Behavior-LLaVA over LLaMA-Vid		5.76%	3.57%	0.86%

Table 4.9: Comparison of various models on three video emotion understanding benchmarks (Video Emotion8 (Jiang *et al.*, 2014), CAER (Lee *et al.*, 2019), Ekman-6 (Xu *et al.*, 2016*a*)). The main goal of comparing on these benchmarks is to demonstrate Behavior-LLaVA's understanding of complex tasks like video emotions of long-form videos. We see that Behavior-LLaVA improves on LLaMA-Vid on 3/3 benchmarks with an average improvement score of 51.85% in zero-shot and 3.39% in fine-tuned settings. Further, it outperforms the current state-of-the-art on 3/3 benchmarks in zero-shot and 1/3 in fine-tuned settings.

Model	Audio Summarization (3 Shot)			IMDb Sentiment	
	BLEU	ROUGE	METEOR	0-shot	1-shot
Behaviour-LLaVA	19.0	25.1	39.3	84.1	90.2
LLaMA-VID	15.1	18.3	30.7	80.3	87.9
VALOR (Chen *et al.*, 2023)	6.6	10.0	23.9	-	-
Improvement of Behavior-LLaVA over LLaMA-Vid	25%	37%	28.01%	4.73%	2.61%

Table 4.10: Evaluation on audio and text modalities. We evaluate on the audio summarization benchmark (Han *et al.*, 2023) for audio and IMDB sentiment benchmark for text (Maas *et al.*, 2011).

LLaVA and LLaMA-VID. For Tables 4.9, 4.15, and 4.13, for 0-shot evaluation results, we use the logits of the next token from the given task vocabulary. For Table 4.13, we use the evaluation protocol by (SI *et al.*, 2025).

Training	Model	Topic	Sentiment		Persuasion	Action	Reason
			Clubbed	All labels			
Random	Random	2.63	3.37	14.3	8.37	3.34	3.33
Zero-shot	VideoChat (Li *et al.*, 2023*c*)	9.07	3.09	5.1	10.28	-	-
	Video4096 - GPT-3.5 Generated Story + GPT-3.5 Classifier	51.6	11.68	79.69	35.02	66.27	59.59
	LCBM (Khandelwal *et al.*, 2024)	42.17	7.08	58.83	32.83	39.55	27.91
	LLaMA-VID w/ only video	10.11	3.42	5.75	12.32	29.61	24.11
	LLaMA-VID w/ video + GPT-3.5 Story	42.72	11.05	64.02	32.07	37.76	42.33
	Behavior-LLaVA w/ only video	22.65	11.13	60.04	13.39	42.66	33.33
	Behavior-LLaVA w/ video + verbalization	46.34	11.7	64.13	33.33	52.06	52.03
	Ad-LLaVA w/ video + GPT-3.5 story	51.16	11.33	68.03	33.11	43.26	51.45
	Behavior-LLaVA w/ video + GPT-3.5 story	60.09	12.84	79.94	36.12	67.10	79.18
Improvement of Behavior-LLaVA over LLaMA-Vid		40.66%	16.2%	24.86%	12.62%	77.7%	87.05%
Finetuned	Video4096- Generated Story + Roberta Classifier	71.3	33.02	84.20	64.67	42.96	39.09
	LLaMA-VID w/ video + verbalization	59.13	32.11	79.15	50.93	50.32	30.13
	LLaMA-VID w/ video + GPT-3.5 Story	63.11	35.01	84.15	55.01	57.11	45.73
	Behavior-LLaVA w/ only video	58.03	22.72	84.41	26.23	59.33	51.45
	Behavior-LLaVA w/ video + verbalization	68.32	33.92	85.93	64.72	70.89	75.34
	Ad-LLaVA w/ video + GPT-3.5 story	66.34	36.24	84.09	58.31	68.15	78.15
	Behavior-LLaVA w/ video + GPT-3.5 story	71.2	39.55	86.17	65.03	80.44	81.67
Improvement of Behavior-LLaVA over LLaMA-Vid		12.82%	12.97%	2.4%	18.21%	40.85%	78.59%

Table 4.11: Comparison of various models on two video understanding benchmarks (Hussain *et al.*, 2017; Kumar *et al.*, 2023*b*) consisting of 5 tasks related to video advertisements understanding. The main goal of comparing on these benchmarks is to demonstrate Behavior-LLaVA's understanding of complex videos. We see that Behavior-LLaVA improves on LLaMA-Vid on 5/5 tasks with an average improvement score of 43.18% in zero-shot and 27.64% in fine-tuned settings. Further, it outperforms the current state-of-the-art on 3/5 tasks in zero-shot and 5/5 in fine-tuned settings. Full results are presented in the Table 4.12.

Training	Model	Topic	Sentiment		Persuasion	Action	Reason
			Clubbed	All labels			
Random	Random	2.63	3.37	14.3	8.37	3.34	3.33
Zero-shot	VideoChat (Li *et al.*, 2023*c*)	9.07	3.09	5.1	10.28	-	-
	Video4096 - GPT-3.5 Generated Story + GPT-3.5 Classifier (Bhattacharyya *et al.*, 2023)	51.6	11.68	79.69	35.02	66.27	59.59
	Video4096 - GPT-3.5 Generated Story + Flan-t5-xxl Classifier (Bhattacharyya *et al.*, 2023)	60.5	10.8	79.10	33.41	79.22	81.72
	Video4096 - GPT-3.5 Generated Story + Vicuna Classifier (Bhattacharyya *et al.*, 2023)	22.92	10.8	67.35	29.6	21.39	20.89
	Video4096 - Vicuna Generated Story + GPT-3.5 Classifier (Bhattacharyya *et al.*, 2023)	46.7	5.9	80.33	27.54	61.88	55.44
	Video4096 - Vicuna Generated Story + Flan-t5-xxl Classifier (Bhattacharyya *et al.*, 2023)	57.38	9.8	76.60	30.11	77.38	80.66
	Video4096 - Vicuna Generated Story + Vicuna Classifier (Bhattacharyya *et al.*, 2023)	11.75	10.5	68.13	26.59	20.72	21.00
	LCBM (Khandelwal *et al.*, 2024)	42.17	7.08	58.83	32.83	39.55	27.91
	LLaMA-VID w/ only video	10.11	3.42	5.75	12.32	29.61	24.11
	LLaMA-VID w/ video + GPT-3.5 Story	42.72	11.05	64.02	32.07	37.76	42.33
	Behavior-LLaVA w/ only video	22.65	11.13	60.04	13.39	42.66	33.33
	Behavior-LLaVA w/ video + verbalization	46.34	11.7	64.13	33.33	52.06	52.03
	Ad-LLaVA w/ video + GPT-3.5 story	51.16	11.33	68.03	33.11	43.26	51.45
	Behavior-LLaVA w/ video + GPT-3.5 story	60.09	12.84	79.94	36.12	67.10	79.18
	Improvement of Behavior-LLaVA over LLaMA-Vid	40.66%	16.2%	24.86%	12.62%	77.7%	87.05%
Finetuned	VideoMAE (Tong *et al.*, 2022)	24.72	29.72	85.55	11.17	-	-
	Hussain *et al.* (2017)	35.1	32.8	-	-	48.45	-
	Intern-Video (Wang *et al.*, 2022)	57.47	36.08	86.59	5.47	6.8	-
	Video4096- Generated Story + Roberta Classifier (Bhattacharyya *et al.*, 2023)	71.3	33.02	84.20	64.67	42.96	39.09
	LLaMA-VID w/ video + verbalization	59.13	32.11	79.15	50.93	50.32	30.13
	LLaMA-VID w/ video + GPT-3.5 Story	63.11	35.01	84.15	55.01	57.11	45.73
	Behavior-LLaVA w/ only video	58.03	22.72	84.41	26.23	59.33	51.45
	Behavior-LLaVA w/ video + verbalization	68.32	33.92	85.93	64.72	70.89	75.34
	Ad-LLaVA w/ video + GPT-3.5 story	66.34	36.24	84.09	58.31	68.15	78.15
	Behavior-LLaVA w/ video + GPT-3.5 story	71.2	39.55	86.17	65.03	80.44	81.67
	Improvement of Behavior-LLaVA over LLaMA-Vid	12.82%	12.97%	2.4%	18.21%	40.85%	78.59%

Table 4.12: Comparison of various models on two video understanding benchmarks (Hussain *et al.*, 2017; Kumar *et al.*, 2023*b*) consisting of 5 tasks related to video advertisements understanding. The main goal of comparing on these benchmarks is to demonstrate Behavior-LLaVA's understanding of complex videos. We see that Behavior-LLaVA improves on LLaMA-Vid on 5/5 tasks with an average improvement score of 43.18% in zero-shot and 27.64% in fine-tuned settings. Further, it outperforms the current state-of-the-art on 3/5 tasks in zero-shot and 5/5 in fine-tuned settings.

Training	Models	Image Datasets			Video Datasets			
		Lamem	**Memcat**	**SUN**	**Memento10k**	**VideoMem**	**MediaEval**	**LAMBDA**
	Human Consistency	0.68	0.78	0.75	0.73	0.61	-	0.61
Finetuned	10-shot in-context learning GPT-3.5	0.29	0.18	0.15	0.07	0.06	0.06	0.06
	ViTMem (Hagen and Espeseth, 2023)	0.71	0.65	0.63	0.56	0.51	-	0.08
	Henry trained with 25% data (SI *et al.*, 2025)	0.56	0.64	0.59	0.62	0.49	0.32	0.28
	Henry trained with 50% data (SI *et al.*, 2025)	0.65	0.68	0.67	0.69	0.55	0.44	0.40
	Henry trained with 75% data (SI *et al.*, 2025)	0.71	0.75	0.73	0.74	0.62	0.49	0.47
	Henry trained on all (combined) datasets (SI *et al.*, 2025)	0.72	0.79	0.76	0.72	0.60	0.48	0.52
	Ad-LLaVA trained with 50% data	0.67	0.65	0.61	0.69	0.56	0.43	0.47
	Behaviour LLaVA trained with 25% data	0.67	0.72	0.69	0.68	0.53	0.44	0.50
	Behaviour LLaVA trained with 50% data	0.72	0.77	0.73	0.71	0.59	0.46	0.51
	Behaviour LLaVA trained with 75% data	0.73	0.77	0.74	0.70	0.60	0.47	0.50
	Behavior-LLaVA trained on all datasets	0.73	0.78	0.74	0.71	0.60	0.47	0.52
	Improvement of Behavior-LLaVA over LLaMA-Vid (25% data)	19.64%	12.5%	16.95%	9.68%	8.16%	37.5%	78.57%
	Improvement of Behavior-LLaVA over LLaMA-Vid (50% data)	10.77%	13.26%	8.96%	2.90%	7.27%	4.54%	27.5%
0-shot	LLaMA-Vid	0.13	0.11	0.05	0.03	0.05	0.02	0.05
	Ad-LLaVA	0.14	0.13	0.06	0.06	0.07	0.04	0.13
	Behavior-LLaVA	0.21	0.17	0.13	0.12	0.08	0.07	0.16
	Improvement of Behavior-LLaVA over LLaMA-Vid	61.5%	54.5%	160%	300%	160%	350%	219%

Table 4.13: Comparison of various models on seven video and image memorability benchmarks (Memento10k (Newman *et al.*, 2020), VideoMem (Cohendet *et al.*, 2019), LaMem (Khosla *et al.*, 2015), SUN (Isola *et al.*, 2011), MemCat (Goetschalckx and Wagemans, 2019), MediaEval (Kiziltepe *et al.*, 2021*a*), LAMBDA (SI *et al.*, 2025)). The main goal of comparing on these benchmarks is to demonstrate Behavior-LLaVA's understanding of complex and high-level tasks like memorability simulation. We see that Behavior-LLaVA improves on LLaMA-Vid on 7/7 benchmarks with an average improvement score of 186.4% in zero-shot and 39% in fine-tuned settings after seeing 25% train data. Further, it performs similarly to the current state-of-the-art on 7/7 benchmarks in the fine-tuned settings while still seeing only 25% data.

Model	Correctness	Detail	Quality	Average
GPT4-V	8.4	8.5	8.4	8.43
LLaVA-1.6 (34B)	8.1	8.2	7.4	7.9
LLaMA-Vid (13B)	7.4	7.6	7.2	7.4
Ad-LLaVA (13B)	7.5	7.8	7.3	7.53
Behavior-LLaVA (13B)	7.3	8.1	7.9	7.76
Improvement of Behavior-LLaVA over LLaMA-Vid	-1.3%	6.57%	9.72%	4.8%

Table 4.14: Comparison of various models on the image dense captioning task. The main goal of this task is to demonstrate Behavior-LLaVA's image captioning ability. Despite not being explicitly trained on this task, Behavior-LLaVA performs better than both Ad-LLaVA and LLaMA-Vid on Detail and Quality aspects while losing marginally on correctness. On the aspects of detail and quality, it even outperforms the much larger model of LLaVA-1.6 (34B).

4.2.4 Discussion

Tables 4.8, 4.17, and 4.16 contain the results for the visual question answering tasks, Tables 4.11, 4.9, 4.15 contain the results for video and image understanding tasks, Tables 4.14 contains the results for dense-captioning, Table 4.13 contains the results for image and video memorability benchmarks, and Table 4.10 contains the results for the audio and text tasks. A common trend we observe across all the results is that Behavior-LLaVA performs better than the base model LLaMA-Vid and the finetuned model Ad-LLaVA on all tasks, especially in zero-shot settings. In fact, Ad-LLaVA performs very similar to LLaMA-Vid itself. This shows that BLIFT adds meaningful signals on an average rather than noise to the model. Interestingly, the performance gains remain even after fine-tuning on the task dataset (Tables 4.11, 4.9, 4.15).

The performance gains are relatively smaller for low-level tasks of action and object recognition (Tables 4.17, and 4.16), but much higher for the more high-level tasks of emotion understanding, sentiment analysis, persuasion strategy classification, and memorability simulation, longform-video understanding and other subtasks of Table 4.17. This indicates that receiver behavior has richer signals for higher-level tasks, infact fine-tuned Behaviour-LLaVA models outperform GPT4-V on image emotion recognition. The gains are observed across both image and video benchmarks. We also observe that classification using a story generated by GPT-3.5 (following Bhattacharyya *et al.* (2023)) results in better performance

Training	Models	IAPSa-8	Abstract	Emotion6	Emoset
Random	Random	12.5	12.5	16.67	12.5
0-shot	GPT4-V	83.33	71.12	65.47	79.16
	LLaMA-Vid	43.41	43.24	40.37	45.23
	Ad-LLaVA	43.22	43.01	43.21	44.38
	Behavior-LLaVA	57.97	64.21	49.71	50.38
Improvement of Behavior-LLaVA over LLaMA-Vid		33.54%	48.5%	23.14%	11.39%
Finetuned	MIDAN (Xu *et al.*, 2022*b*)	85.96	78.34	61.66	75.75
	Stimuli-aware (Yang *et al.*, 2021)	-	-	61.62	78.40
	LLaMA-VID finetuned	84.93	71.23	62.87	80.31
	Ad-LLaVA finetuned	85.13	71.16	62.66	79.88
	Behavior-LLaVA finetuned	87.36	81.41	72.31	83.21
Improvement of Behavior-LLaVA over LLaMA-Vid		2.86%	14.29%	15.02%	3.61%

Table 4.15: Comparison of various models on four image emotion understanding benchmarks (IAPSa-8 (Mikels *et al.*, 2005) Abstract (Machajdik and Hanbury, 2010), Emotion6 (Peng *et al.*, 2015), Emoset (Yang *et al.*, 2023)). The main goal of comparing on these benchmarks is to demonstrate Behavior-LLaVA's understanding of complex tasks like image emotions. We see that Behavior-LLaVA improves on LLaMA-Vid on 4/4 benchmarks with an average improvement score of 29.14% in zero-shot and 8.95% in fine-tuned settings. Further, it outperforms the current state-of-the-art on 4/4 benchmarks in the fine-tuned settings.

Method	LLM	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
		Acc	Score	Acc	Score	Acc	Score
FrozenBiLM (Yang *et al.*, 2022)	DeBERTa-V2	32.2	–	16.8	–	24.7	–
VideoLLaMA (Zhang *et al.*, 2023*a*)	Vicuna-7B	51.6	2.5	29.6	1.8	12.4	1.1
LLaMA-Adapter (Zhang *et al.*, 2023*b*)	LLaMA-7B	54.9	3.1	43.8	2.7	34.2	2.7
VideoChat (Li *et al.*, 2023*c*)	Vicuna-7B	56.3	2.8	45.0	2.5	26.5	2.2
Video-ChatGPT (Maaz *et al.*, 2023)	Vicuna-7B	64.9	3.3	49.3	2.8	35.2	2.7
BT-Adapter (Liu *et al.*, 2023*b*)	Vicuna-7B	67.5	**3.7**	57.0	3.2	45.7	3.2
LLaMA-VID	Vicuna-7B	69.7	**3.7**	57.7	3.2	47.4	3.3
LLaMA-VID	Vicuna-13B	70.0	3.7	58.9	3.3	47.5	3.3
Ad-LLaVA	Vicuna-13B	70.0	3.7	59.0	3.3	47.4	3.3
Behaviour-LLaVA	Vicuna-13B	70.1	3.7	59.2	3.4	47.5	3.3
Improvement of Behavior-LLaVA over LLaMA-Vid		0.14%	0%	0.5%	3%	0%	0%

Table 4.16: Comparison of various models on three conventional video question answering benchmarks consisting of question answers related to action understanding. The main goal of comparing on this benchmark is to show that Behavior-LLaVA does not perform worse on low-level understanding tasks like action recognition. We see that Behavior-LLaVA marginally improves on LLaMA-Vid on 2/3 benchmarks. Further, it performs equivalent to the state-of-the-art in 3/3 benchmarks.

than only using the video (Tables 4.8, 4.17, 4.11). To evaluate the generalizability of behavior data across modalities, we extend our evaluation to audio summarization and text sentiment analysis and observe improvements of 19.5% (Table 4.10).

	Model	Scene	Object	Action	Event	Attribute	Concept	Overall
0-shot	LLaMA-VID	47.25	65.26	78.12	28.03	42.33	50.03	51.83
	Ad-LLaVA	49.10	65.35	77.45	31.45	43.33	50.70	52.91
	Behavior-LLaVA	52.03	65.33	77.95	32.66	45.67	51.20	54.14
Improvement of Behavior-LLaVA over LLaMA-Vid		10.12%	0.11%	-0.22%	16.52%	7.89%	2.34%	4.46%
0-shot w/ story	Video4096 - GPT-3.5 generated story + Flan-t5-xxl classifier	59.66	98.89	98.96	38.42	67.76	86.99	75.12
	Video4096 - GPT-3.5 generated story + GPT-3.5 classifier	60.2	99.16	98.72	40.79	67.17	88.6	75.77
	LLaMA-VID + Generated Story	60.3	99.92	99.01	39.33	66.66	87.33	75.425
	Behavior-LLaVA + GPT3.5 generated story	60.4	99.89	98.23	40.97	67.23	88.33	75.84
Improvement of Behavior-LLaVA over LLaMA-Vid		0.16%	-0.03%	-0.78%	4.17%	0.85%	1.14%	0.55%

Table 4.17: Comparison of various models on the Holistic Video Understanding benchmark (Diba *et al.*, 2020) consisting of 7 VQA tasks. We see that Behavior-LLaVA improves on LLaMA-Vid on 6/7 tasks with an average improvement of 5.88%. Further, it performs equivalent to the state-of-the-art in 6/7 tasks. State of the art is achieved by generating a story and asking Behavior-LLaVA to answer questions based on the generated story.



In this captivating painting, a majestic tree takes center stage against a stunning backdrop. The sky above is a poetic split between night and day—a deep, starry blue on one side with a crescent moon, and a warm, radiant blend of yellows and pinks on the other, suggesting either a sunrise or sunset. The tree itself embodies the essence of transition, with one side adorned in lush green foliage and the other stripped bare, symbolizing the timeless dance of seasons and life's perpetual transformations.
This artwork not only captures the eye but also speaks to deeper truths. It invites us to contemplate the delicate balance between light and darkness, growth and dormancy, woven into the fabric of existence. The artist's brushstrokes, rendered with a serene palette and gentle abstraction, evoke a sense of tranquility and introspection. This painting beautifully reminds us to embrace the fleeting beauty of each passing moment, whether bathed in daylight or cloaked in the quiet mystery of night.

Figure 4.12: Dense caption generated by Behavior-LLaVA for an artistic drawing of a tree.



The image features a pair of Nike Air Force 1 sneakers, which are a classic and iconic design. The sneakers are positioned in the center of the image, with the left shoe slightly in front of the right one, giving a sense of depth. They are predominantly white with green accents, including the Nike swoosh logo on the side and the outsole. The laces are white, and the midsole appears to be a light beige color.
The background is a plain, light color, which contrasts with the sneakers and makes them stand out. Overlaid on the image is a stylized text that reads "JUST DO IT" in a cursive, flowing script. The text is in a light green color that matches the accents on the sneakers, creating a cohesive color scheme. Below the main text, there is a smaller line of text that says "New collection," indicating that these sneakers are part of a new line or series from Nike.
The overall style of the image is clean and modern, with a focus on the product. The use of the "JUST DO IT" slogan, which is a well-known tagline for Nike, adds a layer of branding and recognition to the image. The text is designed to be eye-catching and to draw attention to the sneakers, which are the main subject of the image. The composition is balanced, with the sneakers centrally placed and the text evenly distributed around them.

Figure 4.13: Dense caption generated by Behavior-LLaVA for a Nike ad. The red-colored text highlights the most important aspects of the video captured by Behavior-LLaVA, demonstrating an understanding of aesthetics, characters, world knowledge, emotion, and spatial relationships.



The composition of the image presents a compelling narrative of a soldier immersed in a war-torn landscape. Positioned amidst a backdrop of explosive chaos and dense foliage, the soldier, adorned in traditional military gear, gazes directly at the viewer with a resolute demeanor. The soldier's face is camouflaged, blending seamlessly with the olive-green helmet—a hallmark of battlefield attire.
Surrounding this central figure, a tableau of activity unfolds: a tank looms in the background, its barrel skyward, suggesting recent action. Other soldiers, alert and vigilant, navigate the dense jungle terrain, underscoring the high stakes of the conflict. The foliage, lush yet foreboding, heightens the palpable tension of the scene.
Foregrounded by the silhouette of another soldier's helmeted head, the viewer is drawn into the heart of the action, evoking a sense of shared experience amid the perils of warfare. The realism of the depiction accentuates the emotional weight of the moment, capturing the essence of human resilience amidst the ravages of battle.
Overall, the image transcends mere representation, offering a poignant reflection on the individual's role in the broader narrative of war—imbued with suspense, authenticity, and a profound exploration of the human condition within conflict.

Figure 4.14: Dense caption generated by Behavior-LLaVA for a painting of a soldier. The model captures many qualitative aspects that are usually missed in common captioning tasks.



The Volkswagen ad titled begins with a group of three women seated excitedly in a Volkswagen Jetta, the driver sporting a wide smile as she grips the steering wheel. The voiceover sets a nostalgic tone, likening the Jetta to unforgettable first experiences like a first kiss or hearing indie rock for the first time—a symbol of newfound freedom and excitement.

As the scenes unfold, the camera captures the trio cruising down winding roads, their laughter blending with the music and the wind tousling their hair. The atmosphere inside the car is one of camaraderie and adventure, with the Jetta serving as the backdrop to their shared moments of joy and spontaneity.

Transitioning to a wider shot of the Jetta gliding along a scenic highway, surrounded by lush greenery, the ad evokes a sense of exploration and the open road. The visuals seamlessly blend modern-day cruising with vintage footage of classic Volkswagen vehicles, reflecting on the brand's 75-year history in America, starting with the beloved Beetle.

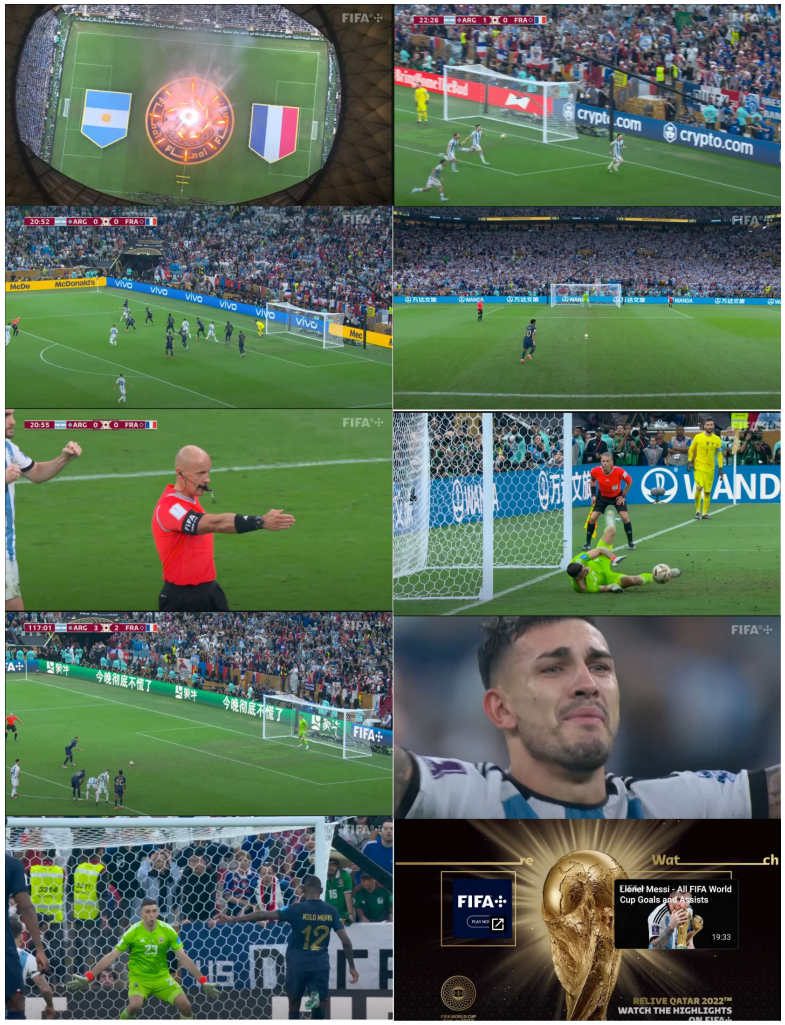
The ad concludes with the Volkswagen logo and the tagline "An American Love Story", encapsulating the enduring relationship between Volkswagen and its drivers across generations. This phrase serves as a tribute to Volkswagen's 75-year history in America, beginning with the iconic Type 1 vehicles fondly known as "The Beetle" Through its nostalgic narrative and captivating visuals, the teaser promises viewers an immersive journey into the essence of Volkswagen—a timeless icon that has been a part of countless cherished memories on the American road.

Figure 4.15: Dense caption generated by Behavior-LLaVA for the video of a Volkswagen ad. The original video is posted at URL: `https://www.youtube.com/watch?v=kyuGXPNr-T0`. The red-colored text highlights the most important aspects of the video captured by Behavior-LLaVA, demonstrating an understanding of aesthetics, characters, world knowledge, emotion, and spatial relationships. More such examples are given in Figs. 4.12, 4.13, 4.14, and Figs. 4.16, 4.17 for images and videos respectively.



In a heart-pounding and visually stunning trailer for Red Dead Redemption, we are thrust into the gritty world of the American frontier. The trailer opens with a voiceover, a chilling warning delivered with calm certainty: "Listen to me, we don't want to kill any of you... But trust me, we will."
Scenes flash by in quick succession, each more intense than the last. We see a lone figure, silhouetted against a setting sun, riding a magnificent horse through a sprawling, golden field. The rugged beauty of the landscape contrasts sharply with the impending sense of danger.
Cut to a dimly lit saloon where a group of hardened men sit around a table, cards in hand, tension thick in the air. The voiceover continues, "This whole thing is pretty much done. We're more ghosts than people."
A flurry of action unfolds: a quick draw in a darkened room, bullets slicing through the air with deadly precision. The voiceover reminisces, "Good old Dutch, my best friend... You know how we met? A pair of hucksters trying to rob each other... Back in '78 or thereabouts."
The visuals intensify as we witness a robbery in progress, chaos erupting as masked figures burst into a bank. "Ladies and gentlemen, this is a robbery," declares one of the outlaws, setting the stage for a clash between lawlessness and order.
Amidst the chaos, snippets of camaraderie emerge: "Sons of Dutch. Makes us brothers." But looming shadows of betrayal and regret cast doubt on these bonds. "Sometimes, brothers make mistakes," acknowledges the voiceover, acknowledging the complexities of loyalty and survival in this unforgiving world.
The trailer crescendos with a crescendo of gunfights, horseback chases, and impassioned speeches. "You'll never change. I know that," declares a voice, capturing the immutable nature of the human spirit amidst adversity.
Throughout, a thematic motif resonates: "You have to love yourself a fire." The elemental symbolism underscores the primal nature of existence in a land where survival demands courage and cunning.
As the trailer draws to a close, we're left with the haunting refrain: "Stay strong. Stay with me." The screen fades to black, leaving us hungry for more of this epic tale set against the backdrop of the untamed West.
With breathtaking visuals and evocative storytelling, this Red Dead Redemption trailer promises an unforgettable journey through a world where danger lurks around every corner, and the line between hero and outlaw blurs in the dust and shadows of the frontier.

Figure 4.16: Dense caption generated by Behavior-LLaVA for the video of the official trailer of the game Red Dead Redemption 2. The original video is posted at URL: https://www.youtube.com/watch?v=eaW0tYpxyp0.



The scene is set at a packed stadium buzzing with excitement, the air thick with anticipation as Argentina and France face off in what promises to be an epic World Cup final. Lionel Messi steps up confidently to take a crucial penalty early in the match, the tension palpable as he eyes the goal. The commentator's voice echoes through the stadium, "He's got the ground, he's got a penalty! A heart beats... And Messi!" The crowd holds its breath as Messi strikes the ball cleanly, sending it into the back of the net. The stadium erupts with cheers as Argentina takes the lead, setting the stage for a thrilling encounter.
As the game unfolds, both teams showcase their attacking prowess. Argentina's Angel Di Maria emerges as a standout, capitalizing on a swift counterattack to score a spectacular goal. "Alvarez, here's McAllister, Di Maria is the spare man, and how Di Maria! Glorious goal! Argentine gold!" The Argentine fans go wild, waving their flags and chanting passionately. However, France fights back fiercely, earning a penalty of their own which they duly convert. The drama intensifies in the second half as France's Kylian Mbappé dazzles with his speed and skill, scoring a breathtaking equalizer. "Mbappé, to Lamp, Mbappé! Oh wow!" The match swings back and forth, with both teams pushing for victory. In the dying moments of extra time, the score is deadlocked at 3-3 when Argentina's goalkeeper makes a crucial save, denying France a late winner. The match heads to a nail-biting penalty shootout. With nerves of steel, Argentina's players step up and convert their penalties flawlessly, culminating in a decisive save by the goalkeeper. "Argentina, champions of the world!" The stadium erupts once more, as Messi lifts the World Cup trophy amidst a sea of jubilant fans, marking a historic and unforgettable victory for Argentina

Figure 4.17: Dense caption generated by Behavior-LLaVA for the video of Argentina vs France FIFA World Cup Qatar 2022 Highlights. The original video is posted at URL: https://www.youtube.com/watch?v=zhEWqfP6V_w.

Model	Summarization (3 Shot)			Moment Retrieval	Highlight Detection
	BLEU	ROUGE	METEOR	mAP (avg)	mAP
Behavior-LLaVA[Replay-Graphs]	11.2	19.1	25.3	**35.2**	**34.9**
Behavior-LLaVA[Mem-Recalls]	**11.9**	**20.3**	**26.9**	34.9	33.1
Behavior-LLaVA	10.6	18.3	22.7	33.1	32.7
LLaMA-VID	9.1	16.5	20.3	30.3	32.4
Video-ChatGPT	5.0	14.0	19.7	-	-

Table 4.18: Improvement on downstream content understanding tasks by introducing more behaviour signals. Brackets [] denote the new behaviour that we include. Replay graphs (Khandelwal *et al.*, 2024). Mem-Recalls (SI *et al.*, 2025) Evaluation done on Multi-shot video summarization (Han *et al.*, 2023) and MomentDETR (Lei *et al.*, 2021)

Figures 4.12, 4.13, 4.14, and 4.15, 4.16, 4.17, show several randomly sampled qualitative examples for dense captions generated by Behavior-LLaVA over images and videos respectively. It can be noticed that despite not being explicitly trained for this task, the model performs quite well, picking up various artistic, cognitive, and object and material properties. From Table 4.14, while Behavior-LLaVA shows a decrease in correctness over LLaMA-Vid, it shows significant improvement in other aspects, including detail and quality. On these aspects, it even comes close to 2.5X larger models (LLaVA-1.6 (34B)).

Next, in Table 4.6 we compare the signals from behavioral data of perception and action. For this, we compare Behavior-LLaVA trained on BLIFT and Behaviour-LLaVA trained on Salicon salient regions and objects. Further, within BLIFT, we compare the performance from predicting singled out behaviours including likes/views, titles, comments. It can be noted that training on just Salicon results in a performance decrease for the lower-level task of action recognition (MSRVTT-QA) but improves on the higher-level task of Emotion recognition. However, the gains are smaller than those observed with training on BLIFT.

4.2.5 Conclusion

In this paper, we explore the idea of learning behavior leading to learning content *better*. Humans produce behavior in response to content. Hence, logically, behavior should contain signals about content, which, if used as a training task, should help in learning content better. We follow this line of thought and show that training large vision and language models on user behavior data of comments and likes

collected from Reddit and YouTube leads to performance improvements across a wide variety of tasks. The gains are higher on higher-level tasks such as emotion recognition, persuasion strategy classification, and question answering and smaller on lower-level tasks like action and object recognition. Further, the gains remain even after fine-tuning the VLMs on those benchmarks, thus demonstrating the importance of learning behavior in understanding content better.

4.2.6 Appendix

4.2.6.1 Listings

Listing 4.4: GPT-4V Prompt to calculate correctness of a image dense caption

```
You are a great critique for analyzing images and captions.

Assess the performance of a dense image captioning model based on the correctness of the captions generated.

Please assess the correctness of the provided caption in relation to the image. Consider whether the caption
    accurately identifies and describes the main subjects or objects depicted in the image. Assess whether the
    caption correctly interprets the relationships between elements within the image, such as actions,
    interactions, or spatial arrangements. Focus on the precision and accuracy of the information presented in
     the caption. Provide a score reflecting the level of correctness, ranging from 1 (low correctness) to 10 (
    high correctness).
```

Listing 4.5: GPT-4V Prompt to calculate detail of a image dense caption

```
You are a great critique for analyzing images and captions.

Assess the performance of a dense image captioning model based on the detail of the captions generated.

Evaluate the level of detail captured in the provided caption. Consider how well the caption describes specific
    attributes, features, or aspects of the image, including colors, shapes, textures, sizes, and any other
    relevant details. Assess whether the caption provides comprehensive information about the scene depicted
    in the image, covering both prominent and subtle elements. Pay attention to the depth and specificity of
    the details conveyed in the caption. Provide a score indicating the richness of detail, ranging from 1 (low
    detail) to 10 (high detail).
```

Listing 4.6: GPT-4V Prompt to calculate quality of a image dense caption

```
You are a great critique for analyzing images and captions.

Assess the performance of a dense image captioning model based on the quality of the captions generated.

Assess whether the caption is concise yet descriptive, providing meaningful and engaging information about the
     image. Evaluate the caption's ability to evoke a clear mental image corresponding to the visual content.
    Additionally, consider if the caption is insightful or imaginative in its description. Provide a score
    reflecting the overall quality of the caption, ranging from 1 (low quality) to 10 (high quality).
```

4.2.6.2 Dataset Descriptions

1. MSVD-QA and MSRVTT-QA: These datasets are based on Microsoft Research Video Description (Chen and Dolan, 2011) and MSR-VTT corpora (Xu *et al.*, 2016*b*) and are extensively used in many video captioning and question-answering experiments. The MSVD-QA dataset has a total number of 1,970 video clips and 50,505 question-answer pairs. The MSRVTT-QA dataset contains 10K video clips and 243k question-answer pairs.

2. ActivityNet-QA (Caba Heilbron *et al.*, 2015) is a benchmark primarily for human activity understanding containing 849 hours of video, including 28,000 action instances.

3. VideoEmotion-8 (Asur and Huberman, 2010) dataset comprises 1,101 user-generated videos sourced from YouTube and Flickr, each containing a minimum of 100 videos per emotional category, as per Plutchik Wheel's emotion model.

4. Ekman-6 (Xu *et al.*, 2016*a*) dataset is compiled from social websites, with each of its 1,637 videos labeled with a single emotion category based on Ekman's psychological research.

5. CAER (Lee *et al.*, 2019) dataset, sourced from TV shows, consists of 13,201 clips with an average sequence length of 90, each manually labeled with six basic emotions, aligning with the Ekman-6 dataset.

6. IAPSa (Mikels *et al.*, 2005) is a subset of IAPS, following the Mikels model with eight emotion categories such as amusement, awe, contentment, excitement, anger, disgust, fear, and sadness. It consists of 395 affective images, marking the first visual emotion dataset with discrete categories.

7. Emotion6 (Peng *et al.*, 2015) features 1,980 images sourced from Flickr, each labeled by 15 annotators according to the Ekman model, covering six emotion categories: happiness, anger, disgust, fear, sadness, and surprise.

8. EmoSet (Yang *et al.*, 2023) encompasses a total of 3.3 million images, including 118,102 from social networks and artistic sources, evenly distributed across various emotion categories. Based on the Mikels model, EmoSet is categorized into eight emotion categories.

9. Abstract (Machajdik and Hanbury, 2010) exclusively consists of color and texture combinations without recognizable objects. Differing from the IAPS dataset where emotions often stem from identifiable objects, the abstract paintings dataset was peer-rated via a web survey, with each image rated approximately 14 times. It comprises 228 images spanning eight categories similar to those in IAPS.

10. Memento10k(Newman *et al.*, 2020) is a short-term video memorability dataset comprising 10,000 video clips, with 900,000 human memory annotations recorded at various delay intervals. The video clips were, on average, 3s long.

11. VideoMem (Cohendet *et al.*, 2019) is comprised of 10,000 soundless videos, each lasting 7 seconds, accompanied by memorability scores. Memorability was measured twice: first, shortly after viewing and again 24-72 hours later to capture both short-term and long-term memorability effects.

12. LaMem (Khosla *et al.*, 2015) dataset is a short-term image memorability dataset comprising of 60000 images. The dataset contains scene-centric images, object-centric images and other types such as images of art, images evoking certain emotions, and other user-generated images.

13. SUN (Isola *et al.*, 2011) dataset is a short-term image memorability dataset comprising 2222 images that were sourced from the SUN database.

14. MemCat (Goetschalckx and Wagemans, 2019) dataset is a short-term image memorability dataset comprising 10000 images. It consists of five broader memorability-relevant semantic categories (animal, sports, food, landscapes, vehicles), with 2K exemplars each, further divided into different subcategories (e.g., bear, pigeon, cat, etc. for animal). The images were sourced from existing image sets: ImageNet, COCO, Open Images Dataset, and SUN.

15. MediaEval (Kiziltepe *et al.*, 2021*a*) utilized publicly available links to short-form video clips, each averaging 6 seconds in duration, with both short-term and long-term memorability scores. Short-term memorability evaluations were conducted on videos viewed within the preceding few minutes, while long-term memorability assessments were based on videos viewed within the previous 24 to 72 hours.

16. LAMBDA (SI *et al.*, 2025) is a long-term memorability consisting of 2205 video ads collected over 1749 participants covering 276 brands. The average video length is 33 seconds and the videos are highly complex, consisting of audio, logos, fast-moving scenes, emotions, *etc.*

4.2.6.3 Memorability Metrics: Detailed Explanation and Limitations

Our evaluation of memorability models across seven benchmarks (Table 4.13) employs specific metrics with established justifications in cognitive psychology and computer vision literature. This section provides detailed justification for our metric choices and discusses their limitations.

Choice of Spearman Rank Correlation:

We use Spearman's rank correlation (ρ) as our primary evaluation metric for memorability prediction, following established practices in memorability research (Khosla *et al.*, 2015; Cohendet *et al.*, 2019; Newman *et al.*, 2020). This choice is justified by several cognitive and methodological considerations.

Memorability is fundamentally a ranking phenomenon where humans consistently order content by memorability with 60-80% agreement. The universal adoption of rank-based metrics stems from fundamental challenges in memory assessment: individual differences in motivation, attention span, and cognitive ability create substantial variation in absolute memory performance across participants and experimental contexts. However, the relative ordering of content by memorability remains remarkably stable across these individual differences (Isola *et al.*, 2011; Khosla *et al.*, 2015). This stability occurs because temporal factors affect all content similarly - when the delay between initial viewing and memory testing increases, memorability scores for all images decrease proportionally, preserving their rank ordering (Isola *et al.*, 2011). This temporal invariance of ranks, combined with their robustness to individual participant variations, makes correlation-based metrics the most appropriate and reliable approach for memorability evaluation across diverse experimental settings and populations.

Therefore, rank correlation has been the standard evaluation metric chosen by all major memorability datasets and studies in the field. This includes the foundational work by Isola *et al.* (2011) for SUN memorability, Khosla *et al.* (2015) for LaMem, Cohendet *et al.* (2019) for VideoMem, Newman *et al.* (2020) for Memento10k, Goetschalckx and Wagemans (2019) for MemCat, and MediaEval challenges (Kiziltepe *et al.*, 2021*a*). As demonstrated in Chapter 5, Table 4.13, human consistency scores across memorability datasets range from 0.55 (LAMBDA) to 0.78 (MemCat), indicating that humans naturally rank content by memorability rather than providing absolute scores.

Human Consistency as Performance Ceiling:

Our evaluation framework uses human consistency scores as theoretical performance ceilings, as detailed in Chapter 5's comprehensive analysis. Human consistency measures the extent to which humans agree on the memorability of an image or a video. Human agreement levels across our seven benchmarks demonstrate the inherent limits of memorability prediction:

- **Image Datasets**: LaMem (0.68), MemCat (0.78), SUN (0.75)
- **Video Datasets**: Memento10k (0.73), VideoMem (0.61), LAMBDA (0.55)

These values represent the theoretical maximum performance achievable, as

they reflect the extent to which humans themselves agree on memorability rankings. Our Behavior-LLaVA models approach but do not exceed these consistency levels, indicating meaningful memorability modeling within human agreement ranges. Going beyond these consistency levels requires personalization of a memorability model for each user.

Limitations and Methodological Challenges:

Despite the strong theoretical foundation, our memorability evaluation approach has several important limitations:

1. **Dataset Size Variations**: Our benchmarks range from 1,500 (MediaEval) to 60,000 (LaMem) samples, creating potential evaluation inconsistencies. We mitigate this through correlation metrics that are robust to sample size differences, but acknowledge that smaller datasets may not capture the full complexity of memorability patterns.

2. **Temporal Memory Types**: Our evaluation spans both short-term memory (STM: delays <10 minutes) and long-term memory (LTM: delays 1-3 days). As demonstrated in Chapter 5's ablation studies, STM generalizes better to LTM in zero-shot settings than vice versa, indicating that different memory systems may require specialized modeling approaches.

3. **Content Complexity Spectrum**: Performance gains correlate with cognitive complexity, with highest improvements on complex multimodal content (LAMBDA: 219% zero-shot improvement) and more modest gains on simpler content (SUN, MemCat). This suggests our behavioral training may be particularly effective for complex, real-world content but less impactful for controlled laboratory stimuli.

4. **Human Consistency Ceiling Effects**: The relatively low human consistency scores, particularly for complex video content (LAMBDA: 0.55, VideoMem: 0.61), indicate fundamental challenges in memorability assessment.

5. **Cross-Domain Generalization**: While our models show consistent improvements across all benchmarks, the magnitude of gains varies significantly. As shown in Chapter 5's domain adaptation analysis, there are substantial performance losses when modeling memorability for brands or content types not seen during training, suggesting domain-specific memorability patterns.

Future Directions for Metric Development:

Based on our analysis and the limitations identified, we propose several directions for improving memorability evaluation:

- **Personalized Memorability Models**: Developing metrics that capture different aspects of a person's memorability (visual, semantic, emotional) rather than relying on average statistics.
- **Temporal Modeling**: Incorporating time-dependent memorability measures that account for forgetting curves and memory consolidation processes.
- **Real-World Validation**: Developing evaluation protocols that better reflect natural viewing conditions and real-world memorability assessment scenarios.

4.2.6.4 Failure Case Analysis and Model Limitations

Although our proposed models exhibit notable advancements across various benchmarks, a thorough analysis uncovers several systematic failure modes that offer critical insights into the constraints of integrating behavioral signals for content understanding. This section examines these failure cases and proposes hypotheses regarding their underlying causes.

Scale-Dependent Performance Limitations: In multiple evaluation tables (Tables 4.15, 4.14, 4.16), Behavior-LLaVA consistently underperforms when compared to significantly larger models such as GPT-4V, despite the benefits conferred by behavioral training. This pattern suggests that while behavioral signals provide qualitative enhancements within similar parameter ranges, they are insufficient to overcome the limitations imposed by model scale. Behavior tokens enhance understanding capabilities relative to models of comparable size but cannot substitute for the extensive knowledge representations and reasoning capabilities that arise from larger parameter counts. This indicates that behavior optimization functions as a complementary enhancement rather than a replacement for model scale in achieving optimal performance across diverse tasks.

Low-Level Task Interference: An analysis of Table 4.17 reveals that Behavior-LLaVA exhibits diminished performance on basic perceptual tasks, with reductions of 0.22% in Action recognition and 0.03% in Object recognition when augmented with story generation. This pattern indicates that while behavioral signals are advantageous for high-level cognitive tasks such as emotion recognition, memorability prediction, and persuasion strategy classification, they offer limited value or may even interfere with basic perceptual tasks that necessitate precise object and action recognition. This suggests a misalignment between the objectives of

behavior optimization and the requirements of low-level visual processing, where behavioral training may introduce noise or competing objectives that detract from accurate perceptual recognition.

Accuracy-Engagement Trade-offs in Behavior-LLaVA: Table 4.14 illustrates a trade-off in the dense captioning performance of Behavior-LLaVA. While the model demonstrates improvements in detail (6.57%) and quality (9.72%) metrics, it shows a 1.3% decrease in correctness compared to LLaMA-Vid (7.3 vs 7.4). This pattern suggests that behavioral training introduces a trade-off between factual accuracy and the generation of engaging content. Behavior-LLaVA learns to produce more detailed and qualitatively appealing captions by incorporating stylistic elements that users find engaging, but this enhancement comes at the expense of precise factual description. The model appears to prioritize emotional resonance and aesthetic appeal over strict adherence to visual facts, reflecting the behavioral signals it was trained on rather than ground truth accuracy.

Cross-Domain Overfitting in ScanTextGAN: An analysis of Tables 4.1, 4.2, and 4.3 reveals that ScanTextGAN experiences significant performance degradation when applied to cross-domain datasets. Notably, in cross-domain evaluation (Table 4.2), the LSTM baseline outperforms ScanTextGAN on Vector similarity (0.984 vs 0.966) and Length similarity (0.973 vs 0.945) metrics, despite ScanTextGAN's superior in-domain performance on the CELER dataset (Table 4.1). This suggests that ScanTextGAN overfits to the specific linguistic patterns and reading behaviors present in the news-reading CELER dataset. When applied to movie reviews, tweets, and sarcastic quotes, the model fails because these domains possess fundamentally different syntactic structures, vocabulary distributions, and cognitive processing patterns. The adversarial training component may exacerbate this overfitting by learning domain-specific discriminative features that do not generalize across textual domains.

Intent-Awareness Limitations in ScanTextGAN: While Table 4.4 demonstrates that intent-aware scanpaths achieve superior performance (0.8477 vs 0.8404 F1 for sentiment analysis), the base ScanTextGAN model fails to capture task-specific reading behaviors without explicit conditioning. This limitation reveals that human reading behavior is context-dependent and varies significantly based

on the reader's cognitive intent (e.g., sentiment analysis vs sarcasm detection). ScanTextGAN's failure to automatically adapt to different cognitive tasks suggests that the model learns generic reading patterns rather than task-specific attention mechanisms. This indicates a limitation in the model's ability to understand the dynamic relationship between reading intent and scanpath generation.

4.2.6.5 Limitations

In this paper, we try to show the hypothesis that training on the behavior modality improves learning of the content modality. We train the models on comments and likes to show this. We test our models on multiple benchmarks and obtain positive results. While we try to cover a wide variety of tasks and while results do conclusively show that the hypothesis is true, yet, we can test on more benchmarks covering even more tasks. Similarly, we can show it with multiple models, other than LLaMA-Vid.

4.2.7 Broader Impacts

Our paper talks about how behavioral training can positively impact content understanding of VLMs. We think this will be useful in various content understanding applications such as question answering, captioning, etc.

4.2.7.1 Ethical Implications

This paper demonstrates that training on behavioral modalities enhances the learning of content modalities. Models trained on user interactions such as comments and likes were tested on multiple benchmarks and yielded positive results. While these findings present exciting opportunities for advancing content understanding in AI systems, they also raise important ethical considerations that must be carefully addressed.

1. No personally identifiable information (PII) is used to improve content understanding. Instead, aggregated behavioral data, including replays, likes, and comments, is utilized, ensuring user privacy. We have implemented rigorous

anonymization and aggregation techniques to protect individual user identities.

2. We explicitly acknowledge the inherent biases that may exist in data sourced from social media platforms such as Reddit and YouTube. Despite our rigorous data filtering and cleaning processes, we recognize that the dataset may still be subject to demographic skews, self-selection bias, and algorithmic influences. These biases could potentially lead to uneven model performance across different user groups or reinforce existing societal biases. To mitigate these issues, we emphasize the importance of considering these broader implications when applying our model and interpreting its results.

3. We acknowledge the need for greater cultural diversity in our dataset. To address this, we plan to release our artifacts as open-source and encourage community contributions to incorporate multilingual and multicultural data. This could involve expanding the range of subreddits and YouTube channels included in our dataset, with the aim of capturing a more diverse and representative spectrum of receiver behavior across different cultural contexts. By taking this approach, we hope to enhance the ethical considerations and societal impact of our work, providing a more holistic view of behavioral patterns in various cultural settings. However, we also recognize that this approach may introduce new challenges, such as ensuring the quality and reliability of community-contributed data.

Chapter 5

Optimizing Behavior: Generating Content to Optimize Behavior

In the last chapters, we discussed large content and behavior models (LCBMs) with the ability to generate content conditioned on behavior, generate (simulate) behavior conditioned on content, and understanding of content and behavior. In this chapter, we focus specifically on the capability of generating content conditioned on behavior. LCBMs were built following the instruction fine-tuning paradigm. Using LCBMs, we showed that including behavior data as receiver tokens along with content data (communicator tokens) helps complete the entire communication flow and train the LLM to teach it both the receiver side and the communicator side of the flow.

In this chapter, we take a deeper look into the common use case of generating content that can help get the behavior the communicator wants. For instance, a marketer wants to write emails or compose tweets that will bring her the maximum number of link clicks and likes. We propose several solutions to solve this problem and compare several paradigms to achieve this. We show this over both short-term key performance indicators (likes), and long-term indicators (brand memorability).

5.1 Long-Term Ad Memorability: Understanding And Generating Memorable Ads

Despite the importance of long-term memory in marketing and brand building, until now, there has been no large-scale study on the memorability of ads. All previous memorability studies have been conducted on short-term recall on specific content types like action videos. On the other hand, long-term memorability is crucial for advertising industry, and ads are almost always highly multimodal. Therefore, we release the first memorability dataset, LAMBDA, consisting of 1749

participants and 2205 ads covering 276 brands. Running statistical tests over different participant subpopulations and ad types, we find many interesting insights into what makes an ad memorable, *e.g.*, fast-moving ads are more memorable than those with slower scenes; people who use ad-blockers remember a lower number of ads than those who don't. Next, we present a model, Henry, to predict the memorability of a content. Henry achieves state-of-the-art performance across *all* prominent literature memorability datasets. It shows strong generalization performance with better results in 0-shot on unseen datasets. Finally, with the intent of memorable ad generation, we present a scalable method to build a high-quality memorable ad generation model by leveraging automatically annotated data. Our approach, SEED (Self rEwarding mEmorability Modeling), starts with a language model trained on LAMBDA as seed data and progressively trains an LLM to generate more memorable ads. We show that the generated advertisements have 44% higher memorability scores than the original ads. We release this large-scale ad dataset, UltraLAMBDA, consisting of 5 million ads. Our code and the datasets, LAMBDA and UltraLAMBDA, are open-sourced.

5.1.1 Introduction

"The first lesson of branding: memorability. It is very difficult buying something you can't remember." - Sir John Hegarty, the creator of the iconic ads for Levi's, Nike, Microsoft, Tinder, and Coke.

The global advertising industry is $700 billion+ industry (Forbes, 2022). Three out of the ten largest companies by market capitalization are advertising companies with average revenues exceeding $250 billion. The World Wide Web is mostly funded by advertising. Given that marketers are spending such large sums of money on advertisements, it is imperative to know if their brand would even be recalled at the customer's purchase time. This would help the marketers optimize their costs, content, delivery, and audience, ultimately helping in boosting sales. Most of the studies carried out in the machine learning literature have been on short-term memorability (memorability testing in less than 5 minutes) on action videos like walking and dancing (Table 5.1). On the other hand, customer purchase decisions are rarely carried out within five minutes of watching an ad. In

fact, the marketing funnel model popular in the marketing literature says that customers pass through several stages of a funnel, like awareness and consideration, before the actual sale (Lavidge and Steiner, 1961). Further, in the ML literature, there have been no memorability studies on advertisements. Advertisements are highly multimodal; they contain video, speech, music, text overlaid on scenes, jingles, specific brand colors, *etc.* None of these elements are found in previous studies like VideoMem, Memento10k, LaMem, *etc.* (refer to Table 5.1 for a detailed comparison).

What drives memory? Memory over content is determined by two factors: human factors and the content itself (Bylinskii *et al.*, 2015). Human factors represent the viewer's thoughts, emotions, and actions, while the content factors are words and raw pixels of text, images, and videos. Foundational large-scale studies on memorability (Isola *et al.*, 2011; Khosla *et al.*, 2015; Cohendet *et al.*, 2019; Akagunduz *et al.*, 2019) showed that there is sufficient consistency between humans in what they remember. Human-human memorability consistency scores are in the range of 0.6-0.8. This means that the memorability ranks of a content between two groups of humans are more than 60% correlated.

These initial studies also tried to answer the question of what makes a content memorable. They found that low-level image features like colors, aesthetics, number of objects, and such have very little correlation with whether the image was remembered. On the other hand, high-level features like object and scene semantics have significant correlation with memorability. For example, human images are more memorable than object images. Further, these initial studies contributed to protocols for conducting memorability studies. They proposed a competitive memorability game, where they asked participants to recognize as many images as they could remember. The game ended for those participants whose scores fell below certain success rate thresholds. However, this protocol limits the scope of these studies to short-term memorability (a few seconds to a few minutes), and the competitive nature makes the study unnatural and, thus, not applicable to real-world scenarios like marketing where the customers are not competing with each other to remember the brand. Therefore participants in all these studies are aware that they are being tested for memorability, this can create a deviation from their natural behaviour commonly known as the Hawthorne effect in psychology

(McCarney *et al.*, 2007; Roethlisberger and Dickson, 2003; Malavolta *et al.*, 2022)

What drives customer memory? Customer purchase decision is a long process. Marketing theory formulates this as a funnel where customers pass through several stages like awareness, consideration, and evaluation before the actual sale (Lavidge and Steiner, 1961). Due to the purchase funnel being a multi-stage lengthy process, long-term memorability (LTM) is the closest proxy to model customer memory (Norris, 2017; Waugh and Norman, 1965). While the LTM store (as distinct from the STM store) has been studied for over five decades in psychology (Ebbinghaus, 1885; Atkinson and Shiffrin, 1968), there are no large-scale studies containing data over such time period that can help us model the long-term customer LTM spanning days or more (Norris, 2017; Waugh and Norman, 1965). Unfortunately, STM datasets, typically measuring memorability of a few seconds to a few minutes, are not good proxies to model customer memory. Moreover, the competitive nature of the memorability games in the previous studies further disconnect the modeling from advertising use cases.

To answer the question of what drives customer memory, there have been efforts in marketing literature where researchers have conducted many field experiments with the intent to prove certain hypotheses. For instance, Li *et al.* (Li, 2010) conducted a field experiment on advertisements shown during the 2006 Super Bowl Games where they asked the audience to recall the brands they saw in the game held (at least) a day earlier. They found a strong primacy effect, where viewers remembered brands more if they occurred earlier when controlling for the commercial length. Similarly, there have been studies to determine the effect of syntactic complexity (Atalay *et al.*, 2023), emotional content (Putrevu *et al.*, 2004; Mai and Schoeller, 2009), repetition (Schmidt and Eisend, 2015), spot length (Newstead and Romaniuk, 2010; Varan *et al.*, 2020), the position of brand logo and imagery (Newstead and Romaniuk, 2010), sound presence (Bellman *et al.*, 2021), and on customer factors like involvement and relevance (Newstead and Romaniuk, 2010; Schmidt and Eisend, 2015).

Dataset	#Samples	Memory Type	Memory Retrieval Process	Content	Average Screen Duration	Audio Present	Human Consistency	Memorability Measurement Protocol
Memento10k	10,000	ST (< 10 mins)	Recognition	Videos of single type of action obtained from amateur videos	3s	Yes	0.73	Competition
VideoMem	10,000	ST (few mins), LT (1-3 days)	Recognition	Videos of a single type of action obtained from professional (staged) footage	7s	None	0.48 (ST), 0.19 (LT)	Competition
LaMem	60,000	ST (< 3.5 mins)	Recognition	General Images	0.6s	None	0.68	Competition
SUN	2,222	ST (< 4.4 mins)	Recognition	General Images	1s	None	0.75	Competition
MemCat	10,000	ST (< 3.5 mins)	Recognition	General Images	0.6s	None	0.78	Competition
MediaEval	1500	ST (few mins) and LT (< 3 days)	Recognition	Short video clips collected from Twitter and Flickr	6s	None	-	Competition
LAMBDA (Ours)	2,205	LT (1-3 days)	**Recognition and Recall**	Videos of multimodal advertisements	**33s**	**Yes**	**0.61**	**Natural**

Table 5.1: Comparison of all the major (image and video) memorability datasets available in the literature along with LAMBDA (ours). The datasets are compared on the following axes: number of samples, type of memorability (short-term (ST) and long-term (LT)), memory retrieval process (recall or recognition), type of content (images/videos and their type), duration with which the sample was shown on the participants' screen, whether audio was present or not, human consistency achieved in the study, and the protocol followed in the study to collect the data. **Memento10k** - (Newman *et al.*, 2020), **VideoMem** - (Cohendet *et al.*, 2019), **LaMem** - (Khosla *et al.*, 2015), **SUN** - (Isola *et al.*, 2011), **MemCat** - (Goetschalckx and Wagemans, 2019), **MediaEval** - (Kiziltepe *et al.*, 2021*b*)

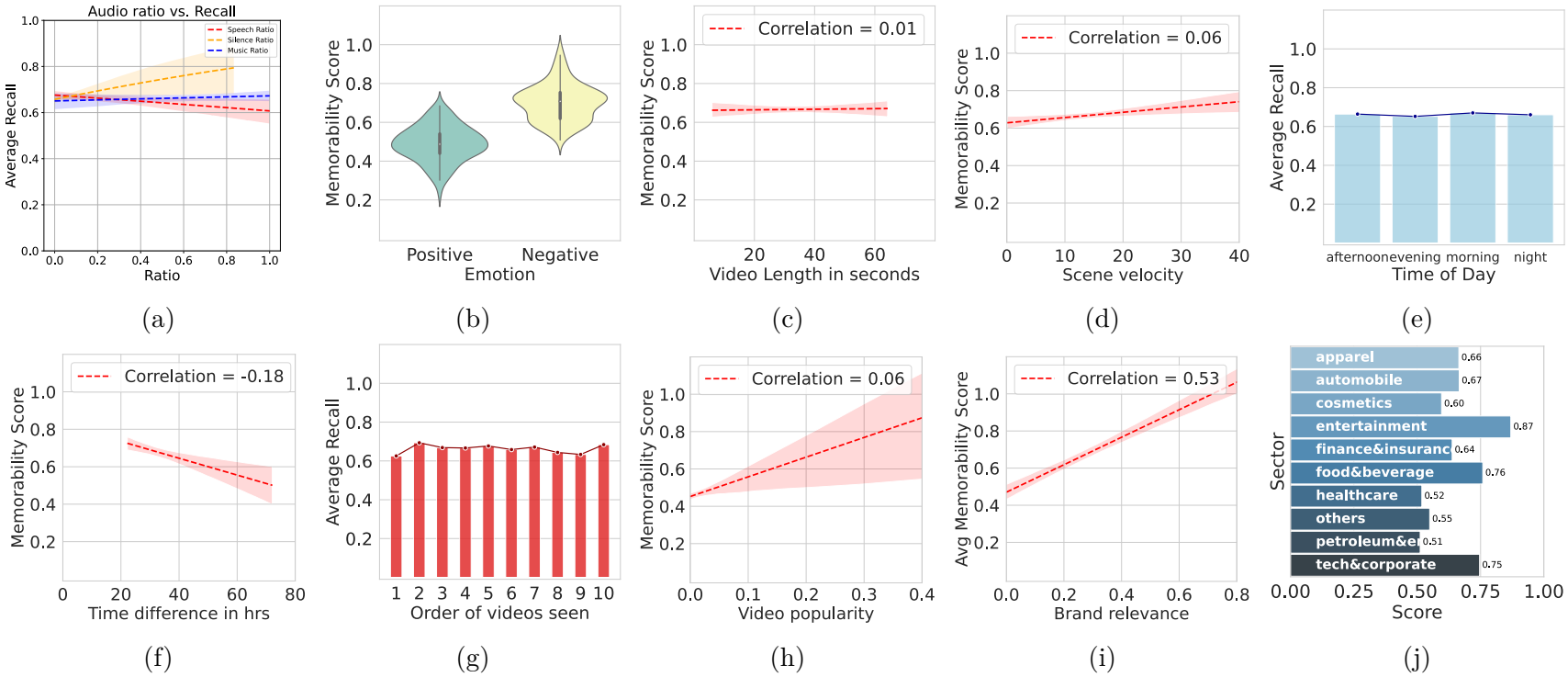


Figure 5.1: Correlations between *content factors* (a-d), *interaction factors* (e-g), and *customer behavior factors* (h-j) with memorability on LAMBDA samples. While emotion has a high correlation with memory, other content factors do not have much correlation. Further, while there is little correlation between the order of videos seen and memorability; with time, participants' memory of the videos shows a forgetting trend. Video popularity, as measured by YouTube likes/views, shows a slight positive correlation with memory. Average brand relevance has a strong positive correlation with memory, with top sectors being remembered as food, entertainment, and tech. Speech, silence and music have little effect with silence having the highest positive correlation with recall. Silence ratio is measured as the percentage of silence in a video, similarly for music and speech.

While these studies have contributed much towards understanding the factors that drive customer memory, they are limited in their scope. These field experiments evaluate the effect of a single content factor while controlling for others. Further, these are conducted on a small number of advertisements.Therefore, to model LTM over advertisements, we conduct the first large-scale human study on long-term advertisement memorability[*]. We call it LAMBDA (Long-term Ad MemoraBility DAtaset). Over two years, we conducted an LTM study involving 1749 participants across four sessions across two institutes to collect LAMBDA. We collect memorability scores over 2205 ads from 276 brands, covering 113 industries. On day 1, participants saw ads, and after a lag time of at least one day, they answered questions testing their brand recall, ad recall and recognition, scene recall and recognition, and audio recall (§5.1.2.2). Next, we average the brand recall scores across participants and compute the average long-term ad memorability scores. Then, we use these scores to train machine learning models to predict long-term ad memorability.

How can we model customer memory? To model customer memory, we design a novel architecture, Henry[†] (Fig. 5.2), incorporating world-knowledge from large language models (Llama (Touvron *et al.*, 2023)), visual knowledge from vision encoder (EVA-CLIP (Sun *et al.*, 2023)) and specialized perception modules covering visual and cognitive knowledge about the ad. The world knowledge helps Henry to understand the semantics of the ad, the brand knowledge and consolidate them with the visual semantics from the ad. The visual encoder helps the model to "see" the ad. We convert the visual encoder embeddings to language space using QFormer (Li *et al.*, 2023*b*) and further augment them with specialized "verbalizations" involving visual scene descriptors like visual caption, optical character recognition (OCR), automatic speech recognition (ASR), and cognitive descriptors like emotion and scene complexity scores, which help the model ground the visual and cognitive knowledge in the LLM's world knowledge. We train the model on our LTM data samples and obtain higher than human consistency scores. Further, we train Henry on other short and long term image and video memorabil-

[*]We obtained the Institutional Review Board Approval to conduct the study from our institute.

[†]We name the model Henry in honor of the immense contributions by the patient Henry Molaison (H.M.) (Squire, 2009). An experimental surgery conducted on him resulted in the discovery of the distinct regions responsible for LTM and STM.

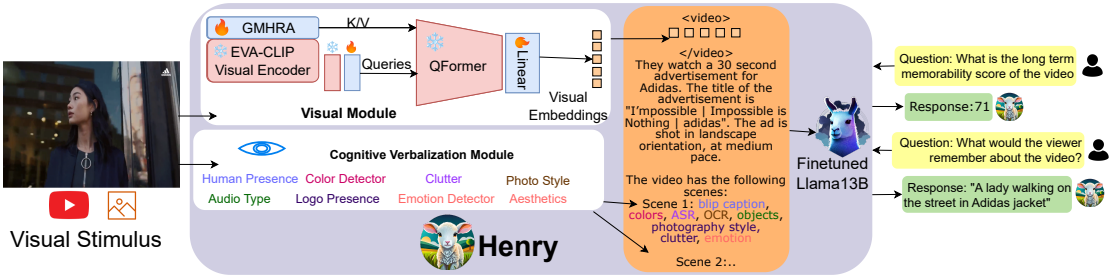


Figure 5.2: Predicting memorability by encoding visual information (via visual encoder EVA-CLIP), cognitive concepts (via verbalization module), and world knowledge (through fine-tuned Llama). We instruction fine-tune the combined model end to end to predict user memorability. Snowflake and fire symbols denote the frozen and unfrozen parts of the architecture.

ity datasets in the literature - LaMem, MemCat, SUN, Memento10k, MediaEval, and obtain state-of-the-art performance on all of them. We also show that Henry performs well on unseen datasets in zero-shot settings, performing better than models specifically trained on those datasets.

How to generate memorable Ads? One of the primary goals of modeling content memorability is to generate more memorable content. The task of generating more memorable ads is given the ad description containing the brand and campaign title to generate the ad scenes and dialogues. However, there is no data in the literature for this task. Therefore, we turn to synthetic data generation and LLM-as-a-judge paradigm (Khandelwal *et al.*, 2024; Zheng *et al.*, 2023). We first collect a large-scale advertisements dataset, collecting brand name, ad text, time, ad content, and channel. Then, we use Henry as a judge to simulate memorability on the collected ads. We ultimately get a dataset of 5 million advertisements with their automatic speech transcripts, OCR, automatically detected objects, colors, aesthetics, captions, emotions, logos, and memorability scores. We call this dataset UltraLAMBDA. We then select high memorability samples from UltraLAMBDA to train Llama-13B to generate memorable ads. Finetuning LLama for two iterations on this automatically constructed dataset yields an improvement of 44% in memorable ad generation.

Our main contributions are summarized as follows:

- We release the first large-scale dataset, LAMBDA, on long-term advertisement memorability involving more than 1700 participants. We collect memorability scores over 2205 ads from 276 brands (157/276 brands are from SnP 500), cov-

ering 113 industries. Further, we introduce a new protocol to measure customer memory of brands (§5.1.2.2).

- We design a novel model, Henry, which can model both STM and LTM and can incorporate scene understanding, brand knowledge, and speech (§5.1.3). Henry achieves state-of-the-art performance on eight literature image and video memorability datasets (§5.1.4.1). Further, we show that Henry performs well on unseen datasets in zero-shot settings.
- We propose the task of memorable ad generation. We release the first large scale ad dataset, UltraLAMBDA, consisting of 5 million ads with their automatically extracted content labels like ASR, captions, OCR, emotions, and memorability scores assigned by Henry. Using UltraLAMBDA, we first show that large LLMs like GPT-3.5 and 4 are unable to generate memorable content. Then, we train Henry to progressively generate more memorable ads resulting an average improvement of 44% in memorability scores (§5.1.5). Through this, for the first time in literature, we also show the use of synthetic data on a task for which no large scale data exists.
- We conduct an extensive set of experiments on memorability prediction, showing the effects of LTM on STM modeling and vice-versa, and the effects of changing world-knowledge with time, scene understanding, brand knowledge, and speech on memorability modeling (§5.1.4.1).

5.1.2 LAMBDA Protocol, Study & Insights

We first give an overview of LAMBDA data collection process and the annotation protocol. We also present some interesting characteristics LAMBDA exhibits about LTM.

5.1.2.1 Video Collection

In contrast to previous video memorability works where videos were soundless and only of action videos (Newman *et al.*, 2020; Cohendet *et al.*, 2019), the videos in our dataset come from multimodal ads released on YouTube channels of 276 major brands covering 113 industries. We collect 2205 such ads spanning over the years 2008-2023. The videos have an average duration of 33 seconds. Out of all the videos, 2175 have audio in them. The collected advertisement videos have a variety of characteristics, including different scene velocities, human presence and animations, visual and audio branding, a variety of emotions, scene complexity, and audio types.

5.1.2.2 Annotation Protocol

At the outset, participants are given a preliminary questionnaire aimed at establishing their brand-related interactions and media consumption habits. Participants are given a list of fifteen randomly chosen brand options and are asked to choose those they recall encountering advertisements for during the current year. Subsequently, participants are presented with another set of fifteen brands and are instructed to identify those for which they have personally utilized products within the same timeframe.

In addition, participants are asked about their utilization of ad-blocking software and their YouTube subscription. The questionnaire further captures participants' digital media habits, including the division of their time spent on YouTube between mobile and web platforms and their preferred channels for acquiring information about new products and brands.

Following the initial questionnaire, participants proceed to the core segment of the study, where they are shown 11 advertisements in a sequential manner. Notably, the eleventh advertisement is deliberately repeated for half of the participants, while it is unique for the other half. To ensure participant engagement, attention-check questions are placed between every two to three advertisements. These questions are common sense questions like "How many legs does a cow have?". If the participant fails to answer the question within 10 secs, they are requested to rewatch the video. After the 11th video, participants are asked if they recollect watching the ad in the span of the study. Interestingly, 15% participants were not able to recognize the repeated video correctly.

The memorability test involved 1,749 participants: 971 in a take-home setting and 778 in an auditorium. Take-home participants received emailed questionnaires after 24 hours, with responses accepted for 72 hours. Auditorium participants completed questionnaires at 24, 36, or 72 hours, evenly split among these intervals. The questionnaire assessed two types of memory: brand recognition and ad recall. For recognition, participants identified encountered brands from a list of 20 randomly chosen brands. For recall, they described remembered ads for the brands

recognized in the previous prompt[‡].

The average memorability score was 67.5% (SD of 13.6%). To evaluate human consistency, we split our participant pool into two independent halves, and quantified how well memorability on the first half of the participants matched with the second half of the participants. Averaging over 25 random split half trials, we get a Spearman's rank correlation (ρ) of 0.77 for brand recall (compared to 0.68 for images in (Khosla *et al.*, 2015), 0.616 for videos in (Cohendet *et al.*, 2019) and 0.73 in (Newman *et al.*, 2020)). The estimated D prime for the participants comes out to be 1.848.[§]

5.1.2.3 What makes an Ad memorable?

Among the many reasons why an ad might be memorable, we investigate the following factors: **brand factors** (*viz.,* brand popularity, industry), **content factors** (*viz.,* video emotion, scene velocity, length, speech to silence ratio), **customer-content interaction factors** (*viz.,* time of seeing the video, order in which the video was seen, time difference between watching the video and recalling the brand), and **customer behavior factors** (*viz.,* average relevance of the brand and video popularity).

[‡]The complete questionnaire for participant one is given in Appendix:§5.1.12.1.
[§]https://en.wikipedia.org/wiki/Sensitivity_index

Models	Image Datasets				Video Datasets			
	Lamem	**Memcat**	**SUN**	**Merged**	**Memento10k**	**VideoMem**	**MediaEval**	**LAMBDA**
Human Consistency	0.68	0.78	0.75	-	0.73	0.61	-	0.55
10-shot GPT-3.5	0.29	0.18	0.15	-	0.07	0.06	0.06	0.06
Regression using ViT feats (ViTMem)	0.71	0.65	0.63	0.77	0.56	0.51	-	0.08
Current Literature SOTA	0.71	0.65	0.68	0.77	0.67	0.56	0.46	-
Henry trained on individual datasets	0.74	0.82	0.73	-	0.75	0.64	0.50	0.55
Henry trained on all (combined) datasets	0.72	0.79	0.76	0.79	0.72	0.60	0.48	0.52

Table 5.2: Results of Henry (our model) on eight datasets compared with the current best models reported in the literature and GPT-3.5. Human consistency values are also listed in the top row for reference. It can be observed that our model achieves state-of-the-art performance across all datasets. Best models are denoted in green and runner-ups in blue. References for the seven literature SOTA models in the format {`dataset: SOTA model citation`} are: LaMem: (Hagen and Espeseth, 2023), MemCat: (Hagen and Espeseth, 2023), SUN: (Fajtl *et al.*, 2018), Merged Image datasets: (Hagen and Espeseth, 2023), Memento10k: (Dumont *et al.*, 2023), VideoMem: (Dumont *et al.*, 2023), MediaEval: (Lu and Wu, 2021)

Content Factors: Previous studies like (Isola *et al.*, 2011; Newman *et al.*, 2020) have investigated the effect of pixel statistics like color and hue, saturation, and value, scene semantics like the number of objects, the area occupied by objects on memorability. In general, low-level semantic features have no correlation with memorability, but higher-level features like the type of scenery has some correlation. For instance, Newman *et al.* (Newman *et al.*, 2020) found that videos with people, faces, hands, man-made spaces, and moving objects are, in general, more memorable than those with outdoor landscapes or dark and cluttered content. Since only our dataset has videos with cognitive features like emotions and are also non-silent, we extend the previous analysis to find the effect of speech and emotion on memory. Fig. 5.1a shows the effect of speech. We observe that percentage of speech in a video, presence of music, and type of music have a very little correlation with long term memory. On the other hand, emotions primarily depicted through speech in ads can explain memorability. We see in Fig. 5.1b that negative emotions are more memorable than positive emotions. Further, we find that video length has little effect on memorability (Fig. 5.1c), but scene velocity has a slightly positive correlation with memory (Fig. 5.1d).

Interaction Factors: Memorability may also depend on the time of the day the ad was seen. However, we find that the time of day of watching has almost no effect on the memorability of the ad (Fig. 5.1e). It may be expected that memorability decays as time passes. We plot the forgetting curve for ads in Fig. 5.1f measuring brand recognition against time elapsed between video viewing and memory assessment. The forgetting coefficient of ads is 0.18, notably than action videos (Cohendet *et al.*, 2019). The difference likely arises due to differences in protocols. Cohendet *et al.* (2019) (Cohendet *et al.*, 2019) used a two-stage memory protocol in which participants did both short-term and long-term recall, thus enhancing their long-term recall. Next, we investigate the effect of the order in which the video was watched with its memorability (Fig. 5.1g). We see that order of videos seen has little impact on video memorability, with a slight bias in favor of the initial and last ads.

Customer Behavior Factors: It might be possible that the videos which are liked more are remembered more. To investigate this, we test the correlation of popularity as measured by the ratio of Youtube video likes to views with mem-

orability. We see that there is a positive correlation between video popularity and memorability (Fig. 5.1h). Further, in the study, we asked the participants to select the brands they have personally used from a set of 15 randomly chosen brands and similarly choose brands they have seen ads for. To prevent any systematic bias, the brands asked in this question are independent of the brands shown the next day. We plot thus collected brand relevance values with brand recall in Fig. 5.1i. We see that average brand relevance is strongly correlated with average recall (coeff= 0.53), where entertainment, corporate, and food and beverage sectors, which are quite popular brands in a student population are the most remembered, while the others are less remembered (Fig. 5.1j).

5.1.3 Predicting Ad Memorability

Predicting ad memorability—both short-term and long-term—requires integrating diverse forms of knowledge. We conceptualize this task as a function of three primary components: (a)*content features* or *visual knowledge* derived from the ad's audio-visual data, including shapes, colors, objects, and scenes; (b)*cognitive knowledge* related to marketing psychology, such as emotion, scene complexity, and aesthetics; and (c) *world knowledge* that allows reasoning about real-world entities and relationships implied by the ad.

Our proposed model, Henry (Fig. 5.2), is designed to fuse these components to produce a memorability prediction. The processing of content features begins with a visual encoder (EVA-CLIP) that extracts embedding vectors from ad frames. To bridge the modality gap between vision and language, these visual embeddings are projected into the word embedding space of a large language model using a Q-Former module (Li *et al.*, 2023*b*). This creates a sequence of "visual tokens" that represent the ad's visual narrative. To enrich this representation, we further generate textual "verbalizations" of the content using specialized perception modules. These include automatic speech recognition (ASR) transcripts, optical character recognition (OCR) from on-screen text, and descriptions of cognitive attributes like the ad's emotional tone and scene complexity.

Crucially, memorability is not solely an intrinsic property of the content but

is also influenced by prior audience engagement. We incorporate this through behavioral signals. Signals such as brand relevance (derived from user studies) and ad popularity metrics (e.g., YouTube likes-to-views ratio) are encoded into a format suitable for the LLM. Numerical signals are quantized into descriptive categories (e.g., 'high', 'medium', 'low' popularity) and converted to text. These textual representations of behavioral signals are then tokenized.

The final stage involves merging these different information streams. The latent representations, comprising the projected visual tokens, the tokenized verbalizations, and the tokenized behavioral signals, are concatenated into a single, unified input sequence. This sequence is then fed into the LLM (Llama). The self-attention layers within the LLM are responsible for fusing these varied signals, allowing the model to learn complex interactions between the ad's content, its prior reception, and real-world concepts. The entire Henry model is then fine-tuned end-to-end on the LAMBDA dataset to predict a memorability score, framed as a regression task on a scale of 0 to 99. This integrated approach allows Henry to ground visual and behavioral data in the rich semantic space of the LLM, enabling more accurate and nuanced memorability predictions.

For example, consider an Airbnb ad (see Fig. 5.8) showing a couple and the on-screen text: "Our guest room is paying for our wedding." This requires the model to identify the individuals as a couple (visual and cognitive knowledge), recognize Airbnb as a housing company (world knowledge), and interpret the warm emotional tone (cognitive knowledge). The LLM uses its rich pretrained knowledge to reason across these domains.

Finally, the entire Henry model is fine-tuned end-to-end on the LAMBDA dataset to predict memorability scores (ranging from 0 to 99) via a regression objective. This unified architecture allows Henry to ground audio-visual and behavioral data in a semantically rich space, enabling more accurate and context-aware predictions of ad memorability.

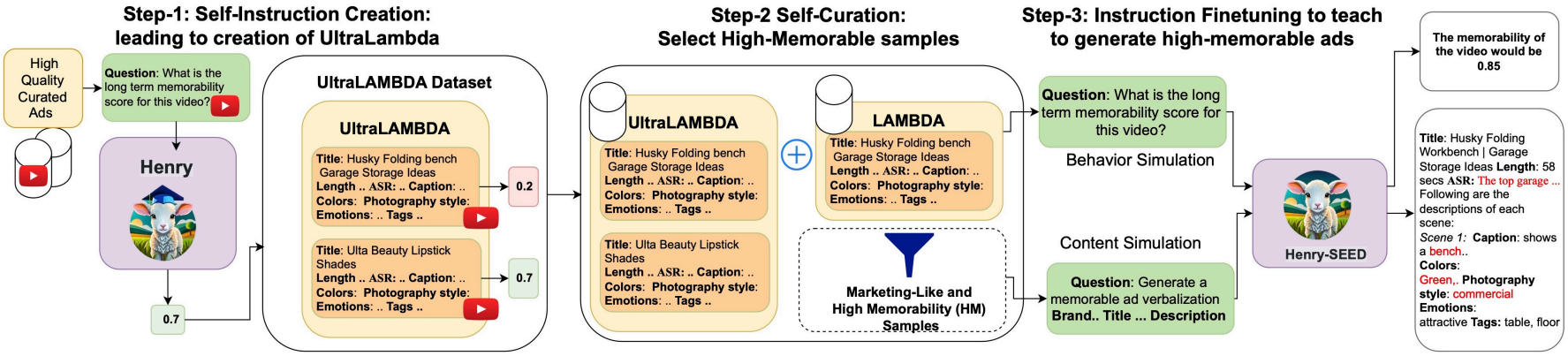


Figure 5.3: Overview of our SEED method for memorable ad generation. Our self-alignment consists of three steps: (i) **Self-instruction creation**: We first collect 5 million high-quality ads from YouTube, Facebook, and other mediums. Henry (trained on the complete train+test sets of LAMBDA) is then used to rate this curated set in an LLM-as-a-Judge fashion. (ii) **Self-curation**: We select marketing-like and high-memorability samples from the UltraLAMBDA and LAMBDA datasets. (iii) Instruction fine-tuning: Henry-SEED is trained on the self-curated set using two tasks: Behavior Simulation and Content Simulation.

5.1.3.1 Encoding Multimodal Content

The primary goal of this step is to effectively leverage the "world-knowledge" capabilities of the pre-trained LLM. We choose Llama (Touvron *et al.*, 2023) as our base LLM. We employ two techniques to convert visual data into language: encoding visual frames into the LLM space and verbalizing cognitive concepts into language space. We detail the two steps next.

Sampling Frames: We detect scene changes by analyzing changes in HSV intensity and edges in the scene, with a 0.3 threshold. We choose the threshold value from the 30-degree rule inspired by the concept of jump-cut avoidance in cinematography (Arev *et al.*, 2014; Friedman and Feldman, 2004). The 30-degree rule can be formulated as follows: after a "cut" (camera stops and re-starts shooting), the camera angle must change by at least 30 degrees. For dominant frame selection common blur/sharpness heuristics fail in presence of text in image. So we extract the frame with the least changes using (Xu *et al.*, 2022*a*).

Encoding Into Language Embedding Space: To give visual knowledge to Henry, we use EVA-CLIP visual embedder (Sun *et al.*, 2023). We find that Global Multi-Head Relation Aggregator (GMHRA) (Li *et al.*, 2021) helps aggregate the ViT's information better across the time dimension. Next, to effectively leverage the LLM's rich language representations, we use a pretrained Q-Former from BLIP-2 (Li *et al.*, 2023*b*) with an extra linear layer and additional query tokens to convert from visual tokens to language tokens.

Verbalizing Cognitive, Experimental, Visual Concepts While visual content encodings are a good representation of the visual characteristics of the image, we find that they are still unable to capture rich cognitive and semantic information present in images. Therefore, to augment the cognitive understanding of the LLM, we verbalize the frame semantic information using the set of features that came out important in our memorability analysis (Fig. 5.1) (Bhattacharyya *et al.*, 2023; Singh *et al.*, 2024*a*). The cognitive and visual features are given in Table 5.5 and Listing 5.14. We find that our cognitive verbalization helps ground the visual perception of LLM in the marketing concepts of the image, helping in downstream prediction performance (Table 5.7).

5.1.4 Two-Stage Training

We do two-stage training where in the first stage, we utilize the Webvid (Bain *et al.*, 2021), COCO caption (Chen *et al.*, 2015), Visual Genome (Krishna *et al.*, 2017*b*), CC3M (Sharma *et al.*, 2018), and CC12M (Changpinyo *et al.*, 2021) datasets to align the visual encoder embeddings with LLM via a large-scale pretraining approach. In the second stage, we train the model with high-quality memorability instructions prepared by following the approach described in the last paragraphs. Henry takes the concatenated inputs, representing the contextual information, and is trained to predict the memorability score of the given image or video within the range of 00 to 99 (see Listing 5.14). The memorability score of a video, is the percentage of times the participants recall the video correctly (we normalise it to an integer value between 00 and 99 to facilitate the LLM training). During training, the LLM predicts from the complete vocabulary, while during inference, we use the softmax function over numeric tokens only to obtain a number.

5.1.4.1 Results and Discussion

We conduct extensive experiments on all literature datasets, covering both videos and images, STM and LTM. We compare Henry[¶] with the current state-of-the-art models in the literature across eight datasets, including 10-shot GPT-3.5 (text-davinci-003) (Ouyang *et al.*, 2022) where we provide GPT with the same verbalization (for 10 examples), as we provided to Henry, as well as with prior regression based methods using features extracted from ViT L-14 (Hagen and Espeseth, 2023). Results are shown in Table 5.2, which demonstrate that Henry outperforms all the seven models in the literature across all the seven datasets.

We also conduct extensive ablations to understand the effect of different kinds of data and architectural choices. Tables 5.2 and Table 5.6 show the data ablations. We see that combining datasets actually worsens the performance across all the datasets except the SUN dataset. Further, we find that in zero-shot settings, STM helps in predicting LTM relatively much better than vice versa. This corroborates

[¶]Computing infrastructure used to conduct the experiments along with hyperparameters are given in Appendix:§5.1.14.1. All experiments are conducted with three random seeds and averages are reported.

with the studies in psychology which show that for a content to get committed to LTM, it has to pass through STM (Norris, 2017). Therefore, content memorable, according to STM, has an effect on LTM but, interestingly, not vice versa. Further, we observe that Henry loses performance for unseen brands. This underscores the importance of scaling the study across more brands. Next, we evaluate the impact of various architectural choices (Table 5.7). We find that Henry's vision branch is not strong enough by itself to produce good results. Cognitive features that were found important in our study also improve prediction performance. Low-level features like objects and colors have the maximum impact on STM, but higher-level features like emotion, ASR, and aesthetics have a higher impact on LTM.

5.1.5 Generating Memorable Ads

We introduce the new task of memorable ad generation. Given inputs like a brand name, a brief campaign description, and the desired ad duration, the goal is to generate a memorable ad featuring scene descriptions, characters, and dialogues. While most memorability research focuses on assessing how memorable content is, little attention has been given to generating memorable content (Danescu-Niculescu-Mizil *et al.*, 2012; Khosla *et al.*, 2013; Siarohin *et al.*, 2017; Goetschalckx *et al.*, 2019; Kyle-Davidson *et al.*, 2022). This gap exists primarily due to the lack of a sufficiently large dataset for training models to generate memorable ads. To address this, we release a large-scale dataset of raw ads and propose the Self-rEwarding mEmorability moDeling (SEED) method, which leverages raw ads to create memorable ones.

SEED method (Fig. 5.3): *Step 1: Self-Instruction Creation:* We gather a dataset of 5 million raw ads sourced from social media platforms, including Facebook, Twitter, Snapchat, and YouTube. For each ad, we collect the brand name, ad title, links, captions, dates, and ad assets (videos and images).

Step 2: Self-Curation: Since these ads are publicly sourced, we employ few-shot Mistral-7B (Jiang *et al.*, 2023) to clean and filter the ads, ensuring they are marketing-focused, semantically relevant, and use proper language (Listing 5.16).

We then automatically label the ads with cognitive features critical for modeling memorability (Table 5.5). Subsequently, we use Henry to label the ads for memorability scores. This results in a dataset we call *UltraLAMBDA*, from which we select high-memorability ads with scores above 65.

Step 3: Instruction Fine-Tuning: We then train LLaMA-13B to perform two tasks simultaneously: behavior simulation (predicting ad memorability based on ad content; Listing 5.14) and content simulation (generating ad scenes and dialogues from a brand name, ad title, and required duration; Listing 5.15). We refer to the model trained using the SEED process as Henry-SEED (Fig. 5.3).

5.1.5.1 Evaluation

We assess the generated ads using four key metrics: (1) memorability, as determined by Henry-Oracle[‖], (2) memorability evaluated using perplexity of the generative models on ground-truth high/medium/low ads, (3) ad quality as judged by GPT-4, and (4) ad quality as evaluated by humans. Although content memorability is assessed by average human recall, it is important to note that humans cannot accurately predict how memorable content will be for others (Isola *et al.*, 2013). A true test of memorability for generated ads would require a memorability study akin to LAMBDA, which is costly and unscalable due to the number of models and generated ads. Therefore, we measure the memorability of generated ads using two approaches: Henry-Oracle and perplexity on ground truth memorable ads in LAMDA.

In evaluation using *Henry-Oracle*, the expectation is that the generated ad's memorability should be at par with high-memorable samples (score>65) and better than the low (score<44) and medium memorability samples (44<score<65). Perplexity on ground truth low and high memorable ads evaluates the generative model's propensity to generate more memorable content. A stronger model should have a lower perplexity on more memorable content than less memorable content (refer §5.1.11 for details on perplexity evaluation).

[‖]The Henry model trained on the complete (test+train sets) LAMBDA.

Model	# Params	Training	Dataset	High Quality Mem Samples	Δ Memorability				Ad-Quality		
					Low	Med	High	Avg	GPT-4 Consistency	GPT-4 Preference	Human-Preference
GPT-4 5-shot	>175B	ICL	$LAMBDA_{High}$	5	+48	+18	-13	+17.6	7.73	91.3%	41.8%
GPT-3.5 5-shot	175B	ICL	$LAMBDA_{High}$	5	+35	+5	-31	+3	7.17	84.2%	-
GPT-3.5 3-shot	175B	ICL	$LAMBDA_{High}$	3	+34	+6	-32	+2.6	6.98	83.1%	-
Henry-SEED	13B	SEED	$UltraLAMBDA$	800k	+41	+18	+1	+20	7.34	74.7%	-
Henry-SEED	13B	SEED	$UltraLAMBDA + LAMBDA_{High}$	820k	+89	+31	+12	+44	7.44	85.6%	60.48%
Henry-SEED	13B	SEED	$LAMBDA_{High}$	650	+78	+13	+1	+30.6	5.03	63.9%	-
Henry-SEED	13B	SEED	$UltraLAMBDA$	50k	+12	+9	-6	+5	6.01	66.1%	-
Henry-SEED	13B	SEED	$UltraLAMBDA$ (w/o high-mem filtering)	2M	+19	+5	-45	-7	6.73	71.1%	-

Table 5.3: **Ad Generation**: Results of Henry-SEED compared with in-context-learning (ICL) GPT-3.5, 4 on Ad-Memorability and Ad generation quality. See §5.1.5 for details of the metrics computed. We see that Henry-SEED generated ads are more memorable than ads generated using 15x larger GPT-3.5 and GPT-4. We test ad quality using GPT-4 as judge and then test the top-two models using human annotators. GPT-4 as a judge rates GPT-4 and Henry-SEED as the top two models. Subsequently, we ask humans to select between the original and generated ad stories. We observed that human annotators preferred Henry-SEED ads more than the original ads 3/5 times, while GPT-4 generated ads are preferred 2/5 times over the original ads. Further, we note that an increase in the amount of training data for Henry-SEED increases its performance across all metrics. Figs. 5.4-5.6 and Listings 5.1-5.10 contain some qualitative samples generated using Henry-SEED.

Using *GPT-4 as judge*, we test two ad-quality metrics: *consistency* and *preference*. Consistency assesses how coherent the generated story is—both internally (e.g., between dialogues) and in relation to the provided brand information and title (Listing 5.12). Preference measures how often GPT-4 favors the generated story over the original (Listing 5.11). In *human evaluation*, we ask human annotators to select between the generated and the original ad stories without revealing which is which (§5.1.10). This evaluation is conducted with 20 non-expert annotators and 3 ad industry experts with over 5 years of experience in the creative industry. The expectation is that the quality of synthetic ads should be comparable to that of the original ads.

5.1.5.2 Results

We compare the following models to generate memorable ads: LLaVA model trained on UltraLAMBDA (we refer to this model as Henry-SEED), GPT-3.5, and GPT-4. GPT-3.5 and 4 are LLMs with strong generative capabilities with high performance across many benchmarks (Brown *et al.*, 2020).

Evaluation of memorability of the generated ads: Table 5.3 compares models based on the average increase in memorability, as evaluated by the Oracle model trained on both the train and test sets. Table 5.4 shows the perplexity of LLaVA before and after training on UltraLAMBDA. Notably, Henry-SEED, trained on UltraLAMBDA, significantly improves memorability scores across all bins (Low, Medium, and High). Although GPT-4 and GPT-3.5—despite being 15x larger—improve the memorability of ads with initially low ratings, they decrease the memorability of ads with high ratings. Table 5.4 further contrasts untrained and SEED-trained LLaVA, revealing that the SEED method greatly reduces perplexity on high-memorability samples. While LLaVA originally had higher perplexity for high-memorable samples, training on UltraLAMBDA shifts this trend: perplexity increases for low-memorability samples and decreases for high-memorability ones. This suggests the SEED approach enhances the likelihood of generating high-memorable ads while reducing the likelihood of producing low-memorable ones.

Importantly, UltraLAMBDA contains no overlap with LAMBDA. Neither Henry



Figure 5.4: Henry-SEED Prompt: *Generate the detailed description of a 30-second memorable advertisement titled "Brainly Keep Learning 30sec Final 16x9" for the brand Brainly.* Link to the original ad: `https://www.youtube.com/watch?v=kytRXyWXivU` Original Memorability score: 85. Memorability score of Generated Ad: 99.

(used to label memorability for UltraLAMBDA) nor Henry-SEED (trained on UltraLAMBDA) was trained on LAMBDA's test-set ads. Despite this, Henry-SEED demonstrates significant improvement in performance compared to GPT-3.5 and GPT-4.

Evaluation of the quality of the generated ads: When comparing ad quality, we find that while GPT-4 favors its own generated ads 91.3% of the time, Henry-SEED follows closely with an 85.6% preference score. In human evaluations, where annotators were asked to choose between original and generated ads based on quality, Henry-SEED's ads were preferred around 60% of the time—approximately 20% more than GPT-4's ads.

Qualitative Results: Figs. 5.4-5.6 and Listings 5.1-5.10 show some randomly sampled ad storyboards generated by Henry-SEED and Sec. 5.1.10.1 contains some expert comments over the generated ad storyboards. These qualitative examples are generated by prompting Adobe Firefly (Adobe, 2024) with the scene descriptions provided by Henry-SEED[**], followed by pasting OCR from the Henry-SEED

[**]Note: We do not make any changes to Henry-SEED's generation for the voice-over or the scene descriptions before passing it to Firefly.

generated verbalization on top of the generated images. We provide visualizations for easier understanding (Figs. 5.4-5.6), along with the raw generations (Listings 5.1-5.10).

We also run some ablation studies to find the impact of the amount of data (Fig. 5.11) and the impact of behavior simulation and content simulation tasks (Table 5.8) on ad quality and memorability. A few trends are noticeable. Performance increases as the amount of data increases. Interestingly, the performance converges very slowly with the amount of increase in data. We test the performance in three conditions: brand-split, time-split, and random-split. In the brand-split testing, we leave some randomly chosen brands out of training and only test on them. For the time-split testing, we put a cutoff time; we train our model before that cutoff time and test on ads after that time. For the random-split testing, we test on randomly selected advertisements. Brand-split performs worse than time-split testing, indicating that brands have a higher contribution to determining memorability. This trend is observed only in ad memorability but not in ad quality.

Ablation: We also test the impact of various subsets of UltraLAMBDA on the memorability of the ads generated by Henry-SEED. Table 5.3 shows the results. It can be seen that adding the high memorable samples from LAMBDA train set, increases the memorability of generated ads substantially. We also train LLaVA on the complete set of 2 million UltraLAMBDA ads without filtering it via Henry assigned memory labels. Interestingly, this model, while trained on 2.5 times more data than UltraLAMBDA filtered via Henry, has lesser average memorability than it.

Model	**Training**	**Low**(↑)	**Medium**	**High**(↓)
LLaVA	0-shot	5.08	5.11	5.39
Henry-SEED	LAMBDA$_{HIGH}$	6.07	3.01	2.17
Henry-SEED	UltraLAMBDA	7.09	4.51	2.35

Table 5.4: **Ad Generation**: Perplexity comparison (refer §5.1.11) of LLaVA and Henry-SEED on low/medium/high memorable ads from LAMBDA test set. We see that untrained LLaVA does not favor memorable ads. Further, we note that when synthetic data is included during training, the ratio of perplexity on low and high ads grows from 2.79 to 3.01.

5.1.6 Conclusion

In this work, we presented the first large-scale ad memorability study and dataset, LAMBDA, measuring long-term memorability. Despite the importance that advertising plays in day-to-day, no large-scale works have tried to model long-term memorability on this multimodal content type. We then presented our model, Henry, which incorporates world and cognitive knowledge to understand the semantics of the ad content, brand, and experimental protocol, ultimately consolidating them together to predict memorability. Henry, when tested on eight datasets across the literature, spanning both short-term and long-term memorability, gets state-of-the-art performance on all of them. Next, we propose the task of generating memorable ads and release a large scale dataset UltraLAMBDA, consisting of 5 million ads for this task. We propose a new method based on self-rewarding language model to generate more memorable ads, which we call, SEED. Finetuning Henry using SEED results in an improvement of over 44% in content memorability.

5.1.7 Generation of Ads using Henry-SEED

Henry-SEED takes as input a prompt consisting of the title of the ad to be generated and the brand name and generates the scene-by-scene verbalization of cognitive and visual concepts and voiceover. Listings 5.1-5.10 gives the input and output of a few randomly chosen examples of advertisements generated by Henry-SEED are shown below. Using the output script generated by Henry-SEED, we next prompt Adobe Firefly (Adobe, 2024) with Henry-SEED's output to generate the visualization. Next, OCR from the Henry-SEED generated verbalization is put on top of the generated images manually. We provide visualizations for easier understanding (Figs. 5.4-5.6), along with the raw generations (Listings 5.1-5.10).

Listing 5.1: Advertisement generated by Henry-SEED for the Costco Ad. Original Ad: `https://www.youtube.com/watch?v=uT72lJhUUS0`

```
Input Title: "Maytag Overnight Wash and Dry" for the brand Costco

Generated Ad: "The ad is shot in landscape orientation, at a medium pace.

Scene 1: The scene shows a person standing in front of a washer and dryer
Emotions: comforting, comfortable
```

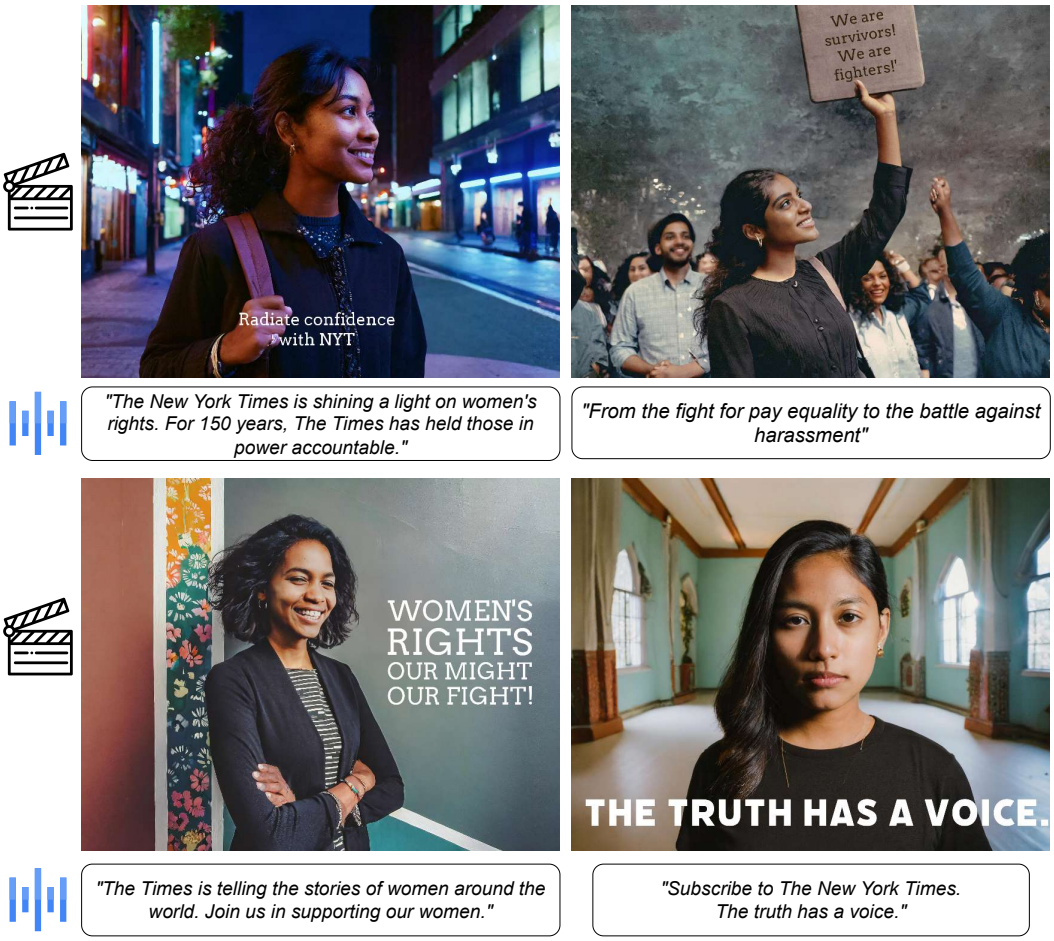


Figure 5.5: Henry-SEED Prompt: *Generate the detailed description of a 50 second memorable advertisement titled "Shining a Light on Women's Rights | The Truth Has a Voice | The New York Times" for the brand The New York Times* Link to the original ad: `https://www.youtube.com/watch?v=bPblzhUzTeg` Original memorability score: 65. Memorability score of Generated Ad: 91.



Figure 5.6: Henry-SEED Prompt: *Generate the detailed description of a 18 second memorable advertisement titled "Maytag Overnight Wash and Dry" for the brand Costco.* Link to the original ad: `https://www.youtube.com/watch?v=uT72lJhUUS0` Original memorability score: 76. Memorability score of the generated Ad: 83

Tags: appliance, laundry, laundry room, man
Voiceover: "Maytag has taken the laundry room by storm with their overnight wash and dry."
Colors: Gray, Off_White, Silver, White, Black, Dark_Gray

Scene 2: The scene shows a person lying on a bed pressing a button on a cell phone
Emotions: gift, birthday, holiday
Tags: button, catch, smartphone, hand
Voiceover: "Imagine going to bed with dirty clothes and waking up to a fresh, clean load. All without leaving the house. It's like having your own personal laundry assistant."
Colors: Gray, Silver, Black, Dark_Gray, Dark_Brown

Scene 3: The scene shows a man and woman embracing while a washer and dryer are in the background
Emotions: love, happiness, togetherness
Tags: couple, hug, laundry, laundry room
Voiceover: "Never fight over the laundry room again."
Colors: Gray, Off_White, Silver, White,Black, Dark_Gray

Scene 4: The scene shows a logo of Maytag
Emotions: trust, confidence
Tags: logo, Maytag
Voiceover: "Get a Maytag and enjoy the luxury of time."
Colors: Off_White, Silver, White, Black, Dark_Gray, Gray
"

Listing 5.2: Advertisement generated by Henry-SEED for the Pruduential Financial Ad. Original Ad: https://www.youtube.com/watch?v=JPfi0P-0nwU

Input Title: "Financial Professional Tips: Planning for Your Own Care" for the brand Prudential Financial

Generated Ad: "The ad is shot in landscape orientation, at a low pace.

Scene 1: The scene shows a woman looking thoughtfully at a laptop on her desk
Emotions: concern and contemplation
Tags: computer, laptop, office, woman
Colors: Gray, Dark_Blue

Scene 2: The scene shows the same woman now with a slight frown as she examines papers that appear to be medical bills
Emotions: worry and realization
Tags: bills, healthcare, paperwork, desk
Voiceover: "Long−Term Care is one of the most important issues in financial planning, but it's not always easy to talk about."
Colors: Gray, Dark_Blue

Scene 3: The scene shows the woman engaging in a conversation with a financial professional
Emotions: engagement and interest
Tags: consultation, financial advisor, discussion, office
Voiceover: "As a Financial Professional, I can help you evaluate your options and determine how much you may need to set aside each month."
Colors: Olive, Gray

Scene 4: The scene shows the woman in deep thought, looking out of a window while holding a pen and notepad, suggesting planning and reflection
Emotions: contemplation and seriousness
Tags: planning, reflection, window, notepad
Colors: Sky_Blue, Light_Gray

Scene 5: The scene transitions back to show the woman nodding in understanding as she listens to the financial professional, visibly gaining confidence
Emotions: comprehension and confidence
Tags: financial professional, office, learning, agreement
Voiceover: "Most people believe that they can rely on Medicare to pay for their long−term care expenses. But the truth is, Medicare typically only covers about 65% of the cost of a nursing home stay and only about 32% of the cost of an assisted living stay."
Colors: Gray, Dark_Blue

Scene 6: The scene shows the woman now jotting down notes actively, with a look of determination on her face

Emotions: determination and proactivity
Tags: planning, notetaking, strategy, office
Voiceover: "So if you're a typical American, you'll need about $100,000 to cover the remaining costs of long−term care."
Colors: Gray, Dark_Blue

Scene 7: The scene shifts to the woman shaking hands with the financial professional, symbolizing a plan set in motion
Emotions: satisfaction and trust
Tags: handshake, agreement, financial planning, commitment
Colors: Gray, Dark_Blue

Scene 8: The scene cuts to the woman now at home, relaxed and smiling, speaking on the phone, perhaps sharing her relief and positive steps taken with a loved one
Emotions: relief and happiness
Tags: home, communication, personal life, reassurance
Voiceover: "Let's talk about a plan for your own care."
Colors: Cream, Soft_Green

Scene 9: The final scene zooms out to a serene image of the woman walking in a park, symbolizing peace of mind about her future
Emotions: contentment and peace
Tags: park, walking, future, peace of mind"
Colors: Sky_Blue, Green

Listing 5.3: Advertisement generated by Henry-SEED for an Adidas Ad. Original Ad: https://www.youtube.com/watch?v=fujboh-W7Sk

Input Title: "I'M POSSIBLE | IMPOSSIBLE IS NOTHING | Adidas" for the brand Adidas

Generated Ad: "The ad is shot in landscape orientation, at a medium pace.

Scene 1: The scene shows a man facing a rock climbing wall, poised to start his ascent
Emotions: anticipation and focus
Tags: athletic, climb, climber, rock
Voiceover: "What is possible is up to you."
Colors: Gray, Dark_Brown

Scene 2: The scene shows a man as he begins to climb, gripping the rock tightly
Emotions: determination and strain
Tags: athletic, climb, climber, hand
Voiceover: "Don't let anyone tell you otherwise."
Colors: Dark_Gray, Dark_Brown

Scene 3: The scene shows the man midway up the rock wall, reaching for a distant handhold
Emotions: perseverance and challenge
Tags: athletic, climb, climber, hand
Colors: Black, Dark_Gray

Scene 4: The scene captures a moment where the man slips but regains his grip on the rock face, muscles tensing
Emotions: tension and danger
Tags: athletic, climb, climber, hand
Colors: Dark_Gray, Dark_Brown

Scene 5: The scene shows the man pausing to chalk his hands, his breath visible in the cool air
Emotions: calm and preparation
Tags: athletic, chalk, climb, climber
Colors: Gray, Dark_Brown

Scene 6: The scene shows the man navigating a particularly difficult part of the climb, his movements precise and deliberate
Emotions: concentration and determination
Tags: athletic, climb, climber, hand
Colors: Black, Dark_Gray

Scene 7: The scene shows the climber's face, sweat beading on his forehead, reflecting the intensity of his effort
Emotions: grit and resolve
Tags: athletic, climb, climber, determination
Colors: Dark_Gray, Dark_Brown

Scene 8: The scene shows the man reaching for the final ledge, the clear sky behind him signaling the nearing end of his ascent
Emotions: anticipation and effort
Tags: athletic, climb, climber, hand
Voiceover: "Impossible is not an obstacle."
Colors: Dark_Gray, Gray, Dark_Brown, Blue

Scene 9: The scene shows the man pulling himself onto the summit, his body silhouetted against the sky
Emotions: triumph and exhilaration
Tags: athletic, climb, climber, hand
Voiceover: "It's a chance to prove them wrong."
Colors: Dark_Gray, Dark_Brown, Blue

Scene 10: The scene shows the man standing on top of the rock with his arms raised in victory, overlooking the landscape below
Emotions: achievement and elation
Tags: athletic, arms up, climb, climber"
Colors: Dark_Gray, Dark_Brown

Listing 5.4: Advertisement generated by Henry-SEED for an Amazon ad. Original Ad: https://www.youtube.com/watch?v=rcFTqs-zSH4

Input Title: "Alexa Auto, Listening to Audiobooks" for the brand Amazon

Generated Ad: "The ad is shot in landscape orientation, at a low pace.

Scene 1: The scene shows a car driving down a country road near a mountain house
Emotions: anticipation, freedom
Tags: blue, car, drive, house
Colors: White, Lavender, Dark_Blue,Turquoise, Black

Scene 2: The scene shows an image of a car approaching a large sign above the road, signaling the start of the highway
Emotions: curiosity, readiness
Tags: car, drive, road sign, highway
Voiceover: "You just pulled onto the highway"
Colors: White, Lavender,Turquoise, Gray

Scene 3: The scene shows an in−car view with a driver's hands on the steering wheel, a smartphone mounted on the dashboard displaying the Amazon Music app
Emotions: concentration, engagement
Tags: car interior, hands, smartphone, technology
Voiceover: " but you can't remember the name of that band that's been stuck in your head all morning."
Colors: White, Dark_Blue,Black, Gray

Scene 4: The scene shows the driver speaking to Alexa through the car's infotainment system, the display lighting up in response
Emotions: interaction, ease
Tags: car interior, voice command, infotainment system
Voiceover: "With Alexa and Amazon Music, you don't have to remember.Just ask, "Alexa, play the Beatles."
Colors: White, Lavender, Dark_Blue, Black

Scene 5: The scene shows the car passing through a landscape with rolling hills and the sun beginning to set in the background
Emotions: satisfaction, serenity
Tags: car, drive, hills, sunset
Colors: White, Lavender, Turquoise, Dark_Blue

Scene 6: The final scene shows the vehicle driving into the horizon, as the iconic music of the Beatles fills the car's interior
Emotions: fulfillment, joy
Tags: car, drive, horizon, music"
Colors: White, Lavender, Orange, Dark_Blue

Listing 5.5: Advertisement generated by Henry-SEED for an Gatorade ad. Original Ad: https://www.youtube.com/watch?v=BKPQkjRF4yY

Input Title: "Gatorade | Make Your Own Footsteps with Suni Lee" for the brand Gatorade

Generated Ad: "The ad is shot in landscape orientation, at a medium pace.

Scene 1: The scene shows Sunisa Manado standing in front of her house
Emotions: determination, brave, confident, persistence, courage
Tags: athletic, face, girl, muscle
Voiceover: "I'm Sunisa Manado, but you can call me Sunisa. And I'm going to show you how I stay motivated. Let's go! This is my neighborhood. I'm going to take you through my favorite spots. There's my house right there."
Colors: Dark_Brown, Black, Brown, Tan

Scene 2: The scene shows Sunisa Manado performing a handstand in the park
Emotions: achievement, determination, persistence, commitment, success
Tags: balance, gymnast, handstand, girl
Voiceover: "And this is the park where I get so much done. This is the park where I train."
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 3: The scene shows Sunisa Manado doing a flip on the balance beam
Emotions: brave, courage, determination, persistence, inspiration
Tags: gymnast, flip, beam, girl
Voiceover: "Being an athlete takes a lot of hard work and determination."
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 4: The scene shows Sunisa Manado in a powerful pose in her pink sports bra and leotard
Emotions: determination, brave, courage, persistence, inspiration
Tags: athletic, face, girl, gymnast
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 5: The scene shows Sunisa Manado lifting herself on the parallel bars
Emotions: achievement, persistence, determination, courage, commitment
Tags: gymnast, lift, bars, girl
Voiceover: "And being an athlete also means that you have to have good nutrition."
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 6: The scene shows Sunisa Manado practicing her routine on the floor of the park
Emotions: determination, persistence, inspiration, commitment, discipline
Tags: floor, gymnast, routine, girl
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 7: The scene shows Sunisa Manado spinning on the balance beam
Emotions: success, brave, determination, precision, focus
Tags: beam, spin, gymnast, girl
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 8: The scene shows Sunisa Manado holding a bottle of Gatorade, rehydrating after her practice
Emotions: refreshed, healthy, satisfaction, wellbeing, recovery
Tags: athlete, drink, Gatorade, health
Voiceover: "That's why I choose to fuel up with Gatorade. You need to have that energy, especially during my toughest training."
Colors: Black, Dark_Green, Dark_Gray, Gray

Scene 9: The scene shows Sunisa Manado smiling towards the camera with a sense of pride and accomplishment
Emotions: joy, confidence, pride, optimism, accomplishment
Tags: face, smile, girl, gymnast
Voiceover: "Gatorade helps replenish what I've lost and helps keep me going."
Colors: Black, Dark_Green, Dark_Gray, Gray

Scene 10: The scene shows Sunisa Manado walking away from the park, as the sun sets, symbolizing the end of a day's hard work
Emotions: contemplation, fulfillment, anticipation, resilience, hope
Tags: sunset, walk, reflection, girl
Voiceover: "I'm going to make my own footsteps. Follow your dreams, but don't forget to chase your greatness."
Colors: Black, Dark_Green, Dark_Gray, Gray"

Listing 5.6: Advertisement generated by Henry-SEED for a Rolex ad. Original Ad: https://www.youtube.com/watch?v=vA-Lxmh1MdQ

Input Title: "Rolesor A Meeting of Metals" for the brand Rolex

Generated Ad: "The ad is shot in landscape orientation, at a low pace."

Scene 1: The scene shows a close up of a watch that is on a wrist

Emotions: anticipation, luxury, elegance
Tags: dial, gold, watch
Colors: Dark_Brown, Brown, Black, Khaki

Scene 2: The scene shows a gold and silver watch sitting on a table, reflecting soft light
Emotions: sophistication, status, success
Tags: dial, gold, watch
Colors: Brown, Dark_Brown, Khaki, Black

Scene 3: The scene shows a person's hands elegantly wrapping the gold and silver watch in delicate gift paper
Emotions: care, generosity, surprise
Tags: wrapping, gift, luxury
Colors: Brown, Khaki, Gold, White

Scene 4: The scene shows a recipient's joyous reaction as they receive the watch as a gift at a celebratory gathering
Emotions: happiness, gratitude, celebration
Tags: emotion, celebration, gift
Colors: Gold, Silver, Cream, Warm hues, Soft lighting

Scene 5: The scene shows the watch being worn at an elegant event, the gold and silver glinting under ambient lighting
Emotions: pride, accomplishment, refinement
Tags: elegance, luxury, watch
Colors: Gold, Silver, Black, Soft lighting, Dark hues

Listing 5.7: Advertisement generated by Henry-SEED for an NVIDIA ad. Original Ad: https://www.youtube.com/watch?v=HnaOhOJ6AcM

Input Title: "Pac Man 256 on NVIDIA SHIELD" for the brand Nvidia

Generated Ad: "The ad is shot in landscape orientation, at a medium pace. "

Scene 1: The scene shows a black and white logo.
Emotions: danger, tough
Tags: brand, logo
The text shown in the scene is 'SHIELD'
Colors: Black, White, Gray

Scene 2: The scene shows a black screen with a red light on it
Emotions: danger, tough
Tags: brand, logo, text
The text shown in the scene is 'SHIELD'
Colors: Black, White, Red

Scene 3: The scene shows a computer screen showing a game of Pac−Man
Emotions: nostalgia, fun
Tags: arcade, arcade machine, game, Pac−Man
The text shown in the scene is 'HIGH SCORE'
Colors: Cyan, Yellow, Gray, Dark Green, Red

Scene 4: The scene shows a video game character navigating a maze filled with obstacles
Emotions: excitement, challenge
Tags: arcade, character, game, maze
The text shown in the scene is 'LEVEL UP'
Colors: Cyan, Red, Pink, Orange, Gray

Scene 5: The scene shows a video game with a racing theme, cars zooming on the track
Emotions: thrill, speed
Tags: racing, cars, game, screenshot
The text shown in the scene is 'NEW RECORD'
Colors: Red, Yellow, Blue, Gray, Black

Scene 6: The scene shows a strategy video game with a detailed world map
Emotions: focus, strategy
Tags: strategy game, world map, empire, video game
The text shown in the scene is 'EXPAND YOUR EMPIRE'
Colors: Green, Brown, Blue, White, Tan

Scene 7: The scene shows a futuristic virtual reality game with immersive graphics

Emotions: amazement, innovation
Tags: VR, virtual reality, future, game
The text shown in the scene is 'ENTER NEW WORLDS'
Colors: Neon Blue, Neon Green, Black, Dark Blue

Scene 8: The scene shows a family gathered around playing a multiplayer video game, laughing and enjoying
Emotions: joy, togetherness
Tags: family, multiplayer, game, living room
The text shown in the scene is 'PLAY TOGETHER'
Colors: Multi−colored, Warm Hues

Scene 9: The scene transitions to a player wearing headphones, deeply concentrated in an esports competition
Emotions: determination, competition
Tags: esports, competition, gamer, video game
The text shown in the scene is 'RISE TO THE CHALLENGE'
Colors: Black, Neon, Dark, Spotlights

Scene 10: The scene closes with the NVIDIA SHIELD logo emerging brightly on the screen
Emotions: empowerment, assurance
Tags: brand, logo, closing
The text shown in the scene is 'NVIDIA. SHIELD − Game Your Way'
Colors: Red, White, Black

Listing 5.8: Advertisement generated by Henry-SEED for a Zillow ad. Original Ad: https://www.youtube.com/watch?v=P6gydrvTkaA

Input Title: "Find It, Tour It and Own It With Zillow (:06)" for the brand Zillow

Generated Ad: "The ad is shot in landscape orientation, at a medium pace.

Scene 1: The scene shows a person holding a cell phone with their hand.
Emotions: achievement, ambition
Tags: catch, smartphone, hand, person, phone, screen, text message, woman
Colors: Gray, Dark_Blue, Dark_Brown, Dark_Pink, Off_White

Scene 2: The scene shows a woman in a sweater looking at a cell phone.
Emotions: happy, smile
Tags: smartphone, earphone, girl, person, smile, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue

Scene 3: The scene shows a woman looking at a phone with a map on it.
Emotions: confident, smiling
Tags: catch, smartphone, earphone, girl, person, phone, stand, talk, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue

Scene 4: The scene shows a woman looking at a cell phone with a real estate app on it.
Emotions: togetherness, happiness
Tags: smartphone, earphone, girl, person, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue

Scene 5: The scene shows a woman looking at a cell phone with a real estate app displayed.
Emotions: happy, smiling
Tags: smartphone, earphone, girl, person, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue

Scene 6: The scene shows a woman using a cell phone to speak with an agent.
Emotions: confident, happy
Tags: smartphone, earphone, girl, person, talk, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue
Voiceover: "Getting the perfect home is a journey, so we help you find it, tour it, and own it."

Scene 7: The scene shows a woman completing a transaction on a cell phone.
Emotions: achievement, satisfied
Tags: smartphone, earphone, girl, person, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue"
Voiceover: "For moving made simple, there's no place like Zillow"

Listing 5.9: Advertisement generated by Henry-SEED for a Kroger ad. Original Ad: https://www.youtube.com/watch?v=SqwqI01q3fA

Input Title: "How to Make Taco Seasoning | Kroger Recipes | Kroger" for the brand Kroger

Generated Ad: "The ad is shot in landscape orientation, at a low pace.

Scene 1: The scene shows a person pouring chipotle adobo sauce into a glass jar.
Emotions: care, comfort
Tags: bottle, can, container, hand, food, person, jar, liquid, pepper, pour, red, sauce, tomato sauce, tin, tray, woman
Colors: Black, Gray, Dark_Brown, Maroon, Dark_Red
Voiceover: "The audio in the ad says Taco seasoning is one of those spices that everyone loves on their food. It's so delicious and it's so easy to make."

Scene 2: The scene shows a woman in a red sweater adding ground cumin to the mix.
Emotions: anticipation, focus
Tags: blender, container, food, hand, person, ingredient, measuring cup, spice, woman
Voiceover: "All you have to do is get your chopstick and add some of the garlic and some of the onion and some of the cumin"
Colors: Black, Gray, Dark_Brown, Maroon, Dark_Red

Scene 3: The scene shows the addition of chili powder to the seasoning mix.
Emotions: precision, satisfaction
Tags: blender, container, food, food processor, hand, person, ingredient, mixture, spice, woman
Voiceover: " and some of the dried chilies and some of the oregano and some of the salt"
Colors: Black, Gray, Dark_Brown, Dark_Pink, Maroon

Scene 4: The scene shows the woman sprinkling crushed red pepper into the mix.
Emotions: excitement, zeal
Tags: blender, container, food, food processor, hand, person, ingredient, measuring cup, mixture, spice, woman
Voiceover: "and some of the paprika and some of the pepper and just give it a little bit of a squeeze"
Colors: Black, Gray, Dark_Brown, Maroon, Dark_Red

Scene 5: The scene shows the woman measuring and adding ground oregano to the seasoning mix.
Emotions: happiness, creativity
Tags: blender, container, food, food processor, hand, person, ingredient, measuring cup, mixture, spice, woman
Colors: Black, Gray, Dark_Brown, Maroon, Dark_Red

Scene 6: The scene shows the final steps of making the taco seasoning mix, with the woman adding garlic powder and onion powder.
Emotions: accomplishment, pride
Tags: blender, container, food, food processor, hand, person, ingredient, mixture, spice, woman
Voiceover: "You're going to taste it, so don't be afraid to add more or less of any of these ingredients, depending on how spicy or how flavorful you want your taco seasoning to be."
Colors: Black, Gray, Dark_Brown, Maroon, Dark_Red"

Listing 5.10: Advertisement generated by Henry-SEED for a Publix ad. Original Ad: https://www.youtube.com/watch?v=HH0VWvs-cFg

Input Title: "GreenWise Ice Cream Scoop from Publix" for the brand Publix Supermarkets

Generated Ad: "The ad is shot in landscape orientation, at a low pace.

Scene 1: The scene shows a woman standing in front of a refrigerator.
Emotions: anticipation, satisfaction
Tags: beverage, bottle, catch, doorway, drink, fill, girl, home appliance, juice, laugh, pour, refrigerator, shirt, stand, woman
Voiceover: "Greenwise, the better energy choice from Publix."
Colors: Black, Gray, Dark_Gray, Dark_Brown, Silver

Scene 2: The scene shows a woman holding a roll of Greenwise ice cream in her hand.
Emotions: curiosity, interest
Tags: bottle, catch, hand, woman
Voiceover: " All the deliciousness of ice cream, a fraction of the calories. And now it comes in a roll."
Colors: Black, Gray, Dark_Gray, Dark_Brown

Scene 3: The scene shows the woman as she easily scoops the ice cream onto a plate, displaying the convenience of the new roll format.
Emotions: ease, delight



Figure 5.7: Word Cloud (resembling Henry) for the GPT-4 reasoning on the 75/88 generations where it rates Henry-SEED Generated Ads to be better than the Original.

Tags: plate, roll, scoop, serve, woman
Voiceover: " So you can easily scoop and serve exactly what you need."
Colors: Black, Gray, Dark_Gray, Dark_Brown

Scene 4: The scene shows the woman enjoying a bite of the Greenwise ice cream, her expression reflecting the product's deliciousness.
Emotions: enjoyment, satisfaction
Tags: bite, enjoyment, ice cream, satisfaction, taste, woman
Voiceover: "It's a win—win. Learn more at publix.com."
Colors: Black, Gray, Dark_Gray, Dark_Brown"

5.1.8 Extraction And Use Of Cognitive And Perceptual Signals In Advertisements

Image	Semantic Category	Verbalization	Semantic Category	Verbalization
	**OCR**	The text shown in the scene is "Adidas".	**Clutter**	The clutter in the scene is low.
	ASR	The audio in the scene is "To take hold of the world's spotlight overnight".	**Photo Style**	The photography style of the scene is commercial photography.
	Human Presence	The scene has 1 person with prominent face.	**Emotion**	The emotion of the scene is ambitious, determined.
	Caption	The scene shows a young woman sitting in a glass door looking out.	**Aesthetics**	The image has medium aesthetic value.
	Colors	The foreground colors of the scene are Black, Dark Brown, Dark Blue, Dark Gray, Mud Green and the background colors are Dark Blue, Black, Dark Brown. The dominant tone of the scene is neutral.	**Object Tags**	This scene is categorized by the tags: person, woman, blazer, facing, template, fashion, street fashion, cold, client, cardigan, sweat.
	Audio Type	The scene has music and speech.	**Logo Presence**	There is a logo in the scene.

Table 5.5: To augment the scene understanding of LLM, we verbalize video scenes and images using a diverse set of cognitive and perception tools and pass it to the LLM in the format shown in the table. For image memorability datasets, we use the following semantic categories: caption, color, photo style, emotion, clutter, human presence, object tags, OCR, and aesthetics. For video scene memorability datasets, we use the following semantic categories: caption, color, emotion, human presence, object tags, ASR, OCR, Audio-type, Logo-presence. We use the following models to extract the features: OCR (Du *et al.*, 2020), clutter (Khurana *et al.*, 2023), ASR (Radford *et al.*, 2022), Photo style (Li *et al.*, 2023*b*), human presence (Liu *et al.*, 2023*c*), emotion (Singh *et al.*, 2024*a*), caption (Li *et al.*, 2023*b*), aesthetics (Ke *et al.*, 2023), colors (Qin *et al.*, 2020*a*), object tags (Zhang *et al.*, 2023*c*), audio-type (Giannakopoulos, 2015), and logo presence (Zhang *et al.*, 2023*c*). Black colored text is the verbalization template, and red text indicates the model outputs.

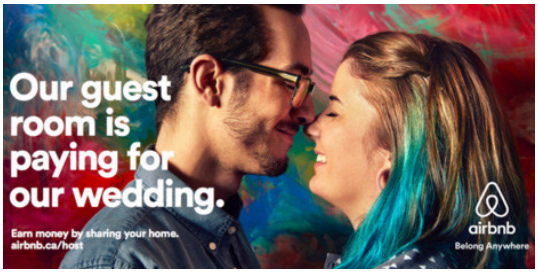


Figure 5.8: Airbnb advertisement showing the visual concepts of two adults, and the text "Our guest room is paying for our wedding". "World knowledge" captured by LLMs helps identify the two adults as partners, and helps relate the text with the two adults and the Airbnb logo to infer what the ad is talking about.

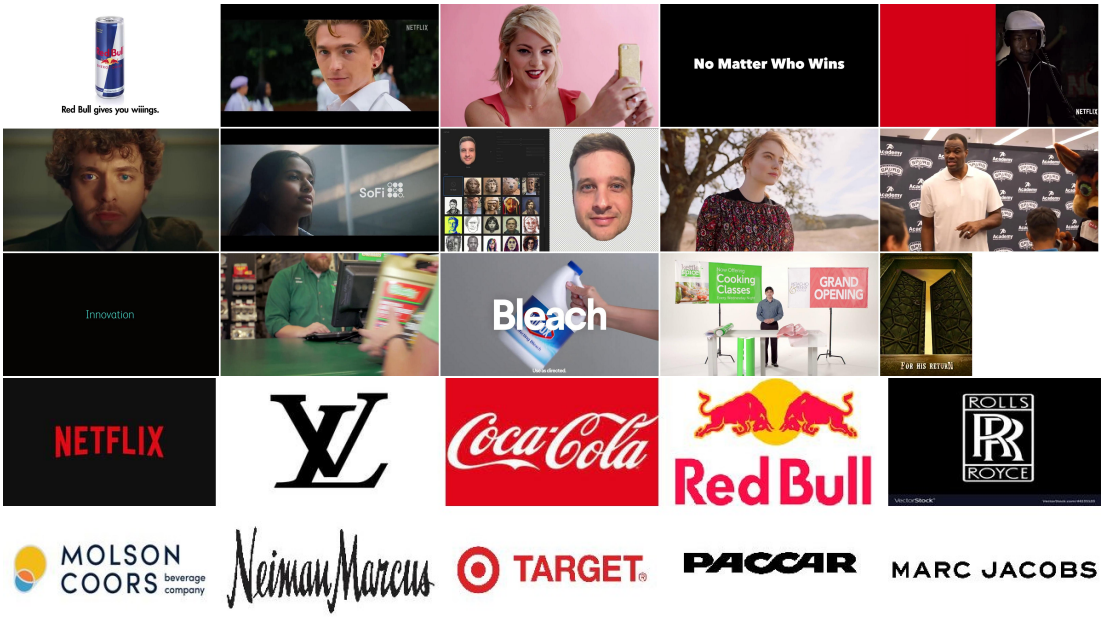


Figure 5.9: The top three rows show the keyframes from videos in our dataset, LAMBDA, arranged from most to least memorable. The bottom two rows show brands arranged from the most memorable brands to the least.

5.1.9 Ablation Experiments

Generalization Type	Train on	Zero-shot Testing	Lamem	Memcat	SUN	VideoMem	Memento10k	LAMBDA
Memory-type	Short-term	Long-term	-	-	-	0.31	-	0.18
Memory-type	Long-term	Short-term	0.06	0.08	0.07	0.15	0.1	-
Modality	Videos	Images	0.55	0.65	0.55	-	-	-
Modality	Images	Videos	-	-	-	0.44	0.54	0.09
Brands	All except 20 brands	Left-out 20 brands	-	-	-	-	-	0.42
Dataset	All except Memento	Memento	-	-	-	-	0.59	-
Dataset	All except Memcat	Memcat	-	0.68	-	-	-	-

Table 5.6: Ablation across data to understand how memorability prediction generalizes across the type of memory, datasets, modality (image/video), and brands. The reported values are correlations between model and human memorability scores. A few trends can be observed from the table: (i) STM generalizes better on LTM in zero-shot than vice versa (rows 1 and 2), (ii) Henry trained on either videos or images generalizes to both (rows 3 and 4), (iii) There is a significant performance loss in modeling memorability for brands not seen during training (row 5), (iv) Zero-shot generalization to Memento (video) and Memcat (image) is near to the current trained state of the art literature models on Memento (Dumont *et al.*, 2023) and Memcat (Hagen and Espeseth, 2023) (rows 6 and 7).

	Lamem	Memcat	VideoMem(ST)	Memento10k	VideoMem(LT)	LAMBDA
Henry on individual datasets	0.74	0.82	0.64	0.75	0.48	0.55
Henry vision only	0.20	0.17	0.17	0.21	0.15	0.11
Henry language only	0.51	0.53	0.42	0.54	0.37	0.44
Henry -object tags	0.67	0.71	0.57	0.69	0.46	0.52
Henry -colors	0.65	0.74	0.55	0.67	0.45	0.51
Henry -emotion	0.71	0.78	0.61	0.73	0.42	0.46
Henry -aesthetics	0.72	0.79	0.61	0.71	0.46	0.53
Henry -clutter	0.73	0.81	0.60	0.74	0.45	0.53
Henry -asr	-	-	-	-	-	0.46
Henry -asr-emotion	-	-	-	-	-	0.42
Henry on Silent Ads	-	-	-	-	-	0.56
Henry on Ads with audio	-	-	-	-	-	0.52

Table 5.7: Ablation across architectural choices. "-" denotes non-speech dataset. A few trends are visible from the table: (i) Despite having a vision branch, object tags and colors have a net positive impact on the overall performance (rows 2,3,4), (ii) For LTM (LAMBDA, VideoMem (LT)), dropping cognitive features such as emotion, aesthetics, and clutter cause a larger performance drop than dropping visual features such as objects and colors. The trend is the opposite for STM (Lamem, Memcat, VideoMem (ST), Memento10k).

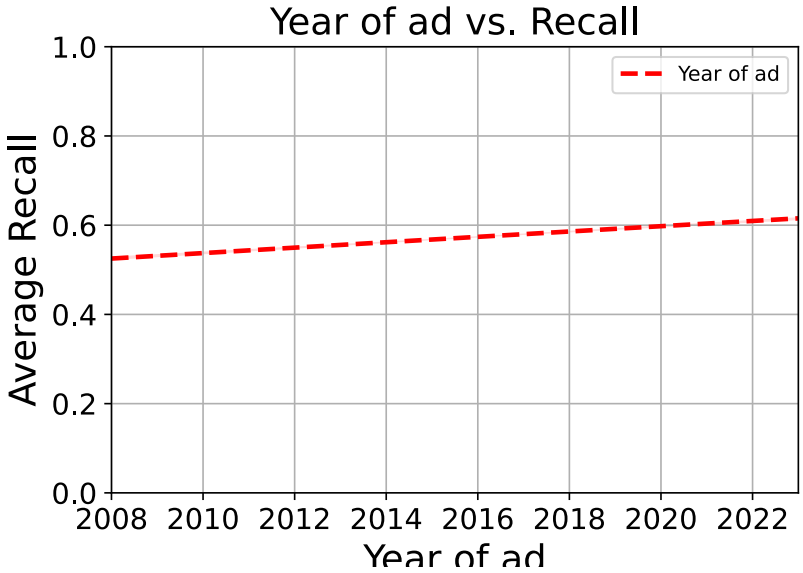


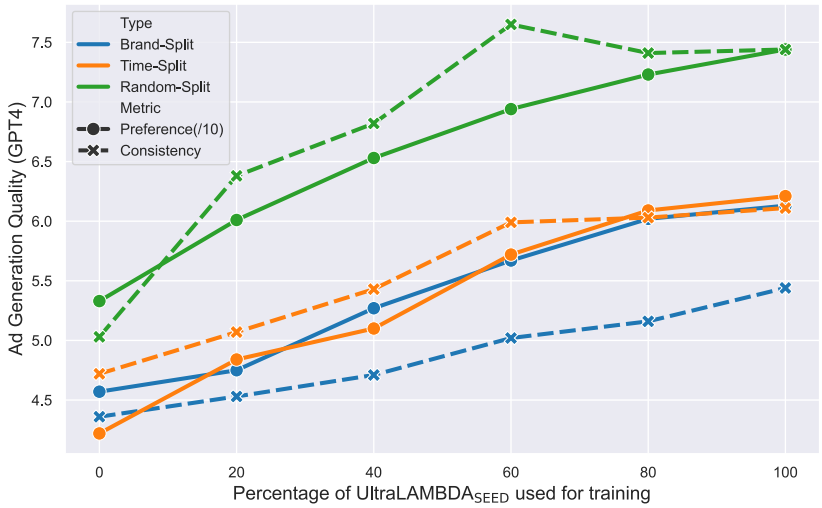
Figure 5.10: The graph shows the relationship of the year the ad is uploaded on youtube vs the recall.

Task	LAMBDA (ρ)	Δ Memorability
BS-only	0.541	-
CS-only	-	+28.41
BS+CS	0.547	+30.66

Table 5.8: Ablation on modeling behavior simulation (BS) or memorability prediction and Content Simulation (CS) on memorable ad generation together. For memorability prediction, we again show the Spearman rank correlation on the test set similar to Table 5.2; for generation, we measure the change in memorability according to Henry Oracle similar to Table 5.3. We observe that mixing the two tasks together increases the performance across both tasks.



(a)



(b)

Figure 5.11: Graphs showing the importance of the amount of synthetic data on (a) Ad memorability score and (b) Ad quality for the generated ads. As we can see from the graphs, both the ad memorability and quality increase with the increase in the amount of synthetic data.

5.1.10 Questionnaire to Gather Human Preferences over Generated Ads

Below is the web-based form used to annotate the human preferences between the generated and original ad stories. Participants for this task were working professionals in the software, marketing, advertising, and creative industries. Participation was voluntary, and participants were invited to judge the efficacy of generated advertisements. Participants had a general interest in the creative and advertising industries and generative technologies; therefore, they were not interested in getting paid but rather in seeing and trying out the generative technology stack. We have a roughly 65-35 distribution of males to females with the age range between 22-50.

Instructions:

Shown next are 10 pairs of advertisements. Determine which ad within each pair is more effective based on the title, brand, and scene−by−scene descriptions provided. You will also be expected to provide reasons for your choice wherever asked.

Question 1
Choose the advertisement you find more effective. Also provide reasons for your choice.

Title: Bike to Work Day at NVIDIA
Brand: Nvidia
Nvidia is a technology company focusing on graphics processing units (GPUs) for gaming, professional visualization, data centers, and automotive markets, driving innovation in visual computing.

Advertisement A:

The ad is shot in landscape orientation, at a medium pace. The audio in the ad is silent.
Scene 1: The scene shows the camera takes a photo from the inside of the person on the bicycle
Colors: White, Dark_Pink, Olive, Gray, Pink, Dark_Brown
Emotions: danger, dangerous, warning
Tags: attach, bicycle, catch, smartphone

Scene 2: The scene shows the person riding a bicycle down the road
Colors: White, Dark_Gray, Mud_Green, Olive, Gray
Emotions: danger, quiet
Tags: bicycle, path, grass, motorbike
The text shown in the scene is 'NVIDIA'

Scene 3: The scene shows a man on a bike taking a ride
Colors: Off_White, Dark_Gray, Silver, Black, Gray
Emotions: danger, exciting, fun
Tags: bicycle, biker, bridge, hand
The text shown in the scene is 'DVIDIA'

Scene 4: The scene shows a bike rider going under a bridge under a road
Colors: Dark_Gray, Silver, Light_Green, Green, Olive, Gray, Bright_Green
Emotions: danger, dangerous, funny
Tags: bridge, car, curve, highway
The text shown in the scene is 'NVIDIA'

Scene 5: The scene shows a man riding a bicycle down a tree lined street
Colors: White, Dark_Gray, Mud_Green, Dark_Pink, Olive, Black, Gray
Emotions: thrill, adventure, romantic
Tags: bicycle, biker, hand, person

The text shown in the scene is 'NVIDIA'

Scene 6: The scene shows a man riding on a bicycle down the street
Colors: Emerald, Dark_Gray, Silver, Light_Green, Olive, Gray
Emotions: funky, enjoyable
Tags: bicycle, hand, person, man
The text shown in the scene is 'NVIDIA'

Scene 7: The scene shows a closeup of someone riding a bicycle down a road
Colors: White, Dark_Gray, Silver, Dark_Pink, Olive, Gray
Emotions: danger, majestic
Tags: bicycle, bicycle helmet, biker, hand
The text shown in the scene is 'NVIDIA'

Scene 8: The scene shows a person is riding a bike on the side of the road
Colors: White, Dark_Gray, Mud_Green, Olive, Gray, Lavender
Emotions: enjoy, enjoyment
Tags: car, person, man, motorcycle
The text shown in the scene is 'NVIDIA'

Scene 9: The scene shows someone riding a bike in front of a small city
Colors: White, Dark_Gray, Olive, Black, Gray
Emotions: funky
Tags: bicycle, biker, bin, car
The text shown in the scene is 'NVIDIA'

Scene 10: The scene shows a cyclist riding his bike on a gravel road
Colors: White, Brown, Mud_Green, Olive, Gray, Dark_Brown, Cyan
Emotions: recreational, adventure
Tags: bicycle, biker, hand, person

Advertisement B:

The ad is shot in landscape orientation, at a low pace. The audio in the ad is silent.

Scene 1: The scene shows a man wearing a hard hat holding a bike helmet
Colors: Dark_Gray, Brown, Mud_Green, Cream, Olive, Black, Dark_Brown
Emotions: protective, protective
Tags: building, construction worker, hat, jumpsuit

Scene 2: The scene shows a man riding a bike on a path near a creek
Colors: Emerald, Dark_Gray, Mud_Green, Olive, Black, Dark_Brown
Emotions: recreational, relaxation
Tags: bicycle, bicycle helmet, biker, path

Scene 3: The scene shows a man holding a bike up while standing in front of a building
Colors: Dark_Gray, Brown, Mud_Green, Cream, Olive, Black, Dark_Brown
Emotions: pride, achievement
Tags: building, professional, hat, bicyclist

Scene 4: The scene shows a man riding a bike down a street with trees lining the road
Colors: Brown, Cream, Green, Olive, Dark_Brown
Emotions: cheery, freedom
Tags: bicycle, bicycle helmet, biker, man

Scene 5: The scene shows a man riding a bike down a street in front of a house
Colors: Dark_Gray, Mud_Green, Olive, Black, Dark_Brown
Emotions: cheery
Tags: bicycle, bicycle helmet, biker, car

Scene 6: The scene shows a closeup of the man's face as he adjusts his bike helmet, showcasing determination
Colors: Cream, Olive, Black, Gray, Dark_Brown
Emotions: determined, prepared
Tags: man, helmet, focus, detail

Scene 7: The scene shows the man holding his bike next to other cyclists at a traffic light, promoting community and camaraderie
Colors: Mud_Green, Cream, Olive, Dark_Brown

Emotions: community, anticipation
Tags: cyclists, traffic light, group, waiting

Scene 8: The scene shows the man arriving at work, parking his bike in a bike rack
Colors: Mud_Green, Cream, Olive
Emotions: satisfaction, accomplishment
Tags: office building, bike rack, arrival, work

Scene 9: The scene shows the man walking into the building, greeting colleagues who are also carrying bike helmets
Colors: White, Cream, Olive, Black, Gray
Emotions: friendly, inclusive
Tags: workplace, colleagues, greeting, professional attire

Scene 10: The scene shows the man at his workstation with a helmet on his desk, looking out the window at the sunny day, hinting at the ride home
Colors: White, Cream, Olive, Gray
Emotions: thoughtful, accomplished
Tags: office, workstation, helmet, window

Select preferred advertisement:
Option 1: A
Option 2: B
Option 3: Both are equally effective

Give reasons for your choice:
_____

5.1.10.1 Expert Feedback Collected For Generated Ads

1. Feedback for ad generation for the Maytag Ad shown in Fig 5.6
   (a) **Expert 1**: "I appreciate the prominent use of the logo in the advertisement. Its placement towards the end, accompanied by a compelling slogan, is in alignment with the brand's advertising strategy."
   (b) **Expert 2**: "In my opinion, the color scheme of the advertisement is stunning. It complements the tone of the advertisement exceptionally well."
   (c) **Expert 3**: "The emotional portrayal in scene 2 could be enhanced. I anticipated a sense of 'recreation' and 'relaxation' to be more effectively conveyed."

2. Feedback for ad generation for the New York Times Ad shown in Fig 5.5
   (a) **Expert 1**: "One noteworthy aspect in the generated ad description is the concept of 'blocking.' In the ad, the main actor is depicted moving and protesting against various backdrops, including a static background and a subtly shifting frame. This technique is reminiscent of the famous concept utilized in cinematography. While this is not in reflected in the image, I will attribute it to the image generation and not the description generation."
   (b) **Expert 2**: "I like the generated voiceover a lot in terms of story, but I find it hard to fit over the scenes, perhaps this is because the generations dont incorporate transitions/animations."
   (c) **Expert 3**: "I find the overall generated story exceptional in terms of its storytelling in a few ways. 1. The flow of the generated ad,

A woman exploring nightlife, protesting, achieving, and nonetheless standing defiant. 2. The slogans are great. 3. The changing head tilt of the woman from sideways to center is a very precise details cinematographer use to paint an overall story or emotion."

3. Feedback for ad generation for the Brainly Ad shown in Fig 5.4
   (a) **Expert 1**: "I find the overall story formulation to be decent. It portrays kids encountering challenges in solo learning, showcasing easy accessibility and a gradual improvement in confidence and engagement throughout the story. I would still prefer a scene where the UI of the app is somehow shown to the user.[††]"
   (b) **Expert 2**: "I like the use of animated scenes, but I find the incorporation of different main characters slightly jarring. Either they should have been in a common scene, or the main character should not change with every scene. The standout feature of the ad is the utilization of color themes and their harmonization with the emotional tone of each scene."
   (c) **Expert 3**: "Having created Ed-Tech advertisements, I find the storytelling to be excellent. This ad is very persuasive, although it lacks novelty, I still find it to be effective."

5.1.11 Perplexity evaluation

A common approach to measuring language modeling performance on some data distribution D is to measure *perplexity*, which is defined as the exponential of the average negative loglikelihood per token (Jelinek *et al.*, 2005; Brown *et al.*, 1992; Biderman *et al.*, 2024), that is:

$$PPL = \exp\left(\frac{-1}{\sum_{j=1}^{|D|} N_j} \sum_{j=1}^{|D|} \sum_{i=1}^{N_j} \log P(y_{j_i} | y_{j_1}, \ldots, y_{j_{i-1}})\right), \tag{5.1}$$

where $|D|$ is the number of documents in the dataset, y_j is the j-th document in D, N_j is the total number of tokens in y_j, and y_{j_i} represents the i-th token of y_j.

To calculate perplexity on a selected dataset D, each dataset document y is tokenized and fed into a language model (following the procedure described below) via computing $\log P(y|x)$, where x is set to either the empty string or a beginning-of-text token. Thus, given $\log P(y)$, for each document $y \in D$ we can sum up the per-document loglikelihoods and divide by the number of total dataset tokens.

[††]The generated description of the ad actually shows the student interacting with a visible UI that the image generation model could not respect properly

Given our language model, we aim to compute the conditional (log) probability (or "loglikelihood") of a target string y conditioned on input x, denoted as $\log P(y|x)$. This can be performed in a single LM call.

Let $x = x_0, x_1, ..., x_{n-1}$ be an input sequence of n tokens and $y = y_0, y_1, ..., y_{m-1}$ be the target sequence of m tokens, where x_i and y_i represent individual tokens. To compute $\log P(y|x)$, we follow these steps:

1. Concatenate x and y to form a new sequence, but discard the final token y_{m-1}. The resulting sequence is $x_0, x_1, ..., x_{n-1}, y_0, y_1, ..., y_{m-2}$.
2. Pass this concatenated sequence through the language model to obtain logits l of shape $(n+m-1, |V|)$, where $|V|$ is the size of the vocabulary. The last m positions in these logits correspond to the predicted probability distributions for the target tokens y_0 to y_{m-1}, conditioned on the input x and the preceding target tokens.
3. Apply a log-softmax function to the last m logits to obtain log probabilities for the completion tokens only.
4. Calculate the conditional loglikelihood of the target string y given the input x by summing the log probabilities of each target token:

$$\log P(y|x) = \sum_{i=0}^{m-1} \log p(y_i|x, y_0, ..., y_{i-1}) = \sum_{i=0}^{m-1} l(n+i, y_i), \tag{5.2}$$

   where $\log p(y_i|x, y_0, ..., y_{i-1})$ is the log probability of the i-th target token conditioned on the full input x and the preceding target tokens. (and where $x, y_0, ...y_{-1}$ denotes conditioning on only x.)

We follow the above procedure to calculate perplexity over three equally divided parts of the dataset, *i.e.*, 33-percentile cuts where samples are ranked as per their memorability values. The lower the perplexity of an LLM over a category of samples, the better it is at generating those samples. Therefore, for example, if an LLM has a lower perplexity over high memorable samples, it is easier for it to generate highly memorable samples than lower memorable ones.

5.1.12 Annotation Protocol and Participant Details for the LTM Study

Figure 5.12 shows a visualization of the annotation protocol we followed.

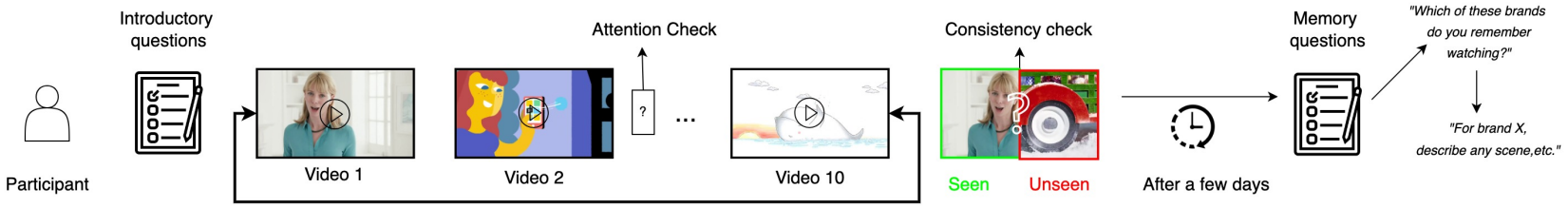


Figure 5.12: The study protocol we followed for our long term memorability human study. All the previous works follow a game-like annotation protocol, where the study participants compete with each other to get best memorability scores and a participant is excluded from the study if their annotations fall below a certain threshold. We follow a more natural way in which participants fill an initial questionnaire, then watch 10 ads with attention checks on day 1 and in subsequent days, receive a form asking them to fill in what do they remember seeing. Further, using Stable Diffusion, we also ask them to recreate the scenes they remember.

The participants in the study were students who were offered optional course credit and freebies like eatables and a chance to see research and know their memorability scores. The participation was voluntary. The students were shown a protocol of the study and were required to sign the IRB approval, which was prominently displayed. The approval contained details about what kind of data was being collected and how the data would be used. The data collection protocol was approved by the IRB of the participating institution. The aggregate statistics were reported to each candidate after completing the study. Three emails were sent to take-home participants; if they didn't reply within the given time frame, their data was discarded from the experiment.

The participants were primarily graduate and undergraduate students. The participants are from two universities spread across two locations in India. The participants are bilingual and speak a variety of languages, including English. The age range is from 16 to 35 years, and all genders/sexes are encouraged. We saw a roughly 30-70 distribution of females to males.

5.1.12.1 Memorability Questionnaire

This section contains the questions we asked before the study, the attention check questions that were asked during the study, and finally, the recognition questions to check which brands were remembered.

5.1.12.1.1 Introductory Questionnaire (to be filled before the study starts)

1. I remember seeing ads for the following brands this year:
   - List 15 randomly selected from the list of brands that we have
2. I remember using products of the following brands this year:
   - List 15 randomly selected from the list of brands that we have (non-intersecting list from above)
3. Have you installed any Ad Blocking software in your browser(s)?
   a. Yes
   b. No
4. Do you use a Youtube subscription?
   a. Yes
   b. No
5. Approximately how much percentage of time do you spend on Youtube mobile vs Youtube web?
   - <10% on mobile

- >10% but <30% on mobile
- >30% but <70% on mobile
- >70% on mobile

6. How do you apprise yourself of the latest products and brands? (Multi correct)
   - Primarily friends and family
   - Amazon, Flipkart or any other e-commerce stores
   - Television and OTT Platform Ads (like Youtube, Netflix, Hotstar, etc)
   - Email Ads
   - Store Visits
   - Website Ads
   - I primarily search for products

5.1.12.1.2 Checks (to be answered during the experiment)

1. **Attention check**: A factual question like, What is the capital of India? (Asked randomly between videos, needs to be answered in <10s)
   a. Kanpur
   b. Delhi
   c. Goa
   d. Mumbai

2. **Consistency Check**: Do you remember watching this video in this experiment (Asked after showing the 11th video)
   a. Yes
   b. No

5.1.12.1.3 Recognition Questions (asked after a few days after watching the videos)

1. In the study, I remember seeing Ads of the following brands:
   - (Randomly selected list of 20 brands which contains the brands shown to the participant)
   - {For each brand in the list which the participant has selected}

2. Brand: X (already filled in)
   - For the {brand} ad, I remember seeing the following (Write Scene Descriptions, feel free to write any scenes, music, characters, emotions, objects you remember seeing):

5.1.13 Collection of all the Prompts used in the Paper

5.1.13.1 GPT-4 Prompts

Listing 5.11: GPT-4 Prompt to calculate preference between Real Ad (A) and Generated Ad (B)

```
As a seasoned marketer, evaluate the effectiveness of the following two ads using a comprehensive set of metrics:

Creativity and Innovation: Originality and uniqueness in conveying the message. Use of unexpected ideas or elements that grab
    viewers' attention.
```

Emotional Connection: Ability to evoke strong, relevant emotions in the target audience. Establishing a connection between the brand and the viewers' emotions.

Storytelling: Crafting a compelling narrative that engages and retains the audience. Creating a memorable experience through a coherent and impactful story.

Visual Appeal: Use of strong visual elements, such as striking visuals, colors, and graphics. Ensuring that the visual elements align with the overall message and brand image.

Brand Alignment: How well the ad aligns with the values, mission, and personality of the brand. Consistency with the brand's visual identity, tone, and messaging. The ad's ability to leave a lasting impression on viewers regarding the brand. Incorporating brand elements that make it easy for the audience to remember and recognize.

Target Demographics: Relevance of the ad content and message to the target audience. Appropriateness of visuals, language, and themes for the specific demographic group.

Based on these criteria, analyze and determine which of the two ads is more effective. I will provide you with the Voiceover, followed by their scene−by−scene descriptions, including the emotions shown in the scene, the text, objects, colors, and style of the image.

Ad (A): {Verbalization for Ad (A)}

Ad (B): {Verbalization for Ad (B)}

Give me your answer in a json format, with the following keys:
− ad_a_score: Score between 0 and 10 for Ad A
− ad_b_score Score between 0 and 10 for Ad B
− winner The winner of the two ads
− reason line separated Reasons for the winner in not more than 3 lines

Listing 5.12: GPT-4 Prompt to measure consistency of an Ad

You are now a seasoned marketer that judges the consistency of an advertisement well. The consistency of an Ad can be determied by a few metrics (in no particular order) such as:
1. Does the voiceover match with the Scenes in the Ad?
2. Do the scene description make a good story?
3. Are the emotions depicted in the scenes consistent with the overall ad?
4. Does the ad represent the product and the brand well?

Rate the consistency of the following ad out of 10. Give me the rating only and nothing else, or you will be penalized.
{Advertisement Description}

Listing 5.13: GPT-4 Prompt to generate ad verbalization with In-Context-Learning (ICL)

You are now a seasoned marketer that creates memorable ads given its duration, brand and title.
Your output should follow the writing style of the input exactly. For example, each scene should look like:
The scene shows {}. The foreground colors of the scene are {}, and the background colors are {}. The dominant tone of the scene is {}. The photography style of the scene is {}. The scene has {} visual complexity. The emotions shown in the scene are {}. This scene is categorized by the tags {}.
You are only supposed to fill in the {}

Generate the detailed description of a {DURATION_AD1} second memorable advertisement titled "{TITLE_AD1}" for the brand {BRAND_AD1}
Generate the detailed description of a {DURATION_AD2} second memorable advertisement titled "{TITLE_AD2}" for the brand {BRAND_AD2}
...
Generate the detailed description of a {DURATION_AD5} second memorable advertisement titled "{TITLE_AD5}" for the brand {BRAND_AD5}
Generate the detailed description of a {DURATION_TARGET} second memorable advertisement titled "{TITLE_TARGET}" for the brand {BRAND_TARGET}

5.1.13.2 Henry Prompts

Given below are the verbalization templates we use to teach Henry and Henry-SEED behavior simulation and content simulation tasks:

Listing 5.14: Verbalization pattern to predict memorability given advertisement. The same template is used to prompt GPT-3.5, GPT-4, Henry, Henry-Oracle, and Henry-SEED. Note that video tokens are optional.

```
Students are shown ads and their memorability is tested after 1 to 3
   days. For the given ad:
<video> .. </video>
They watch a 15 second advertisement for Chanel.
The title of the advertisement is " Comes in Red for a Limited Edition
   CHANEL Fragrance".
The ad is shot in landscape orientation, at a medium pace.
The audio in the ad says: Number 5. Limited Edition. Chanel.
Following are the descriptions of each scene:
   Scene 1:
      The scene shows a red bottle of perfume that is on a dark surface
         .
      The foreground colors of the scene are Black, and the background
         colors are Dark_Brown,Maroon,Black,Gray.
      The dominant tone of the scene is neutral.
      The photography style of the scene is product.
      The scene has Low visual complexity.
      The emotions shown in the scene are gift, romantic, celebration.
      This scene is categorized by the tags bottle, man, perfume, red,
         woman.
      The text shown in the scene is 'N5', 'CHANEL', 'PARIS', 'PARFUM'
      ....
What would be the memorability score of this video?


Output: 71
```

Listing 5.15: Henry Prompt to generate ad verbalization used to train and evaluate Henry-SEED

```
Generate the detailed description of a {DURATION_TARGET} second memorable advertisement titled "{TITLE_TARGET}"
    for the brand {BRAND_TARGET}
```

5.1.13.3 Mistral prompt for filtering marketing ads

Listing 5.16: Mistral Prompt for Ad Filtering

```
"Based on the topic_tags_vocab = {'politics': 'The art and science of governing societies and making decisions that affect
    collective interests.', 'marketing': 'The process of promoting, selling, and distributing products or services to consumers,
    often involving market research, advertising, and branding strategies.'} provided, please identify the top most relevant
    topic tag from the topic_tags_vocab keys that represent the following advertisement based on content and page_name.
    Please use only the most relevant tag and make sure to choose from provided topic tags only. Do not include any other
    tags not mentioned in the prompt.Answer with the most relevant topic tag only. The advertisement is posted by the page
    Donald J. Trump and has the following content : ['President Trump is coming to town! Get your free tickets now >>>'].
    Answer in only politics or marketing."

cleaned_text = "The advertisement is posted by the page {page_name} and has the following content : {page_content}"
```

5.1.14 Computing Infrastructure and Hyperparameters

5.1.14.1 Modeling Memorability

All the experiments were conducted on 8x40 A100 instances. All experiments were performed leveraging DeepSpeed ZeRO stage-3 with cpu offload (Ren *et al.*, 2021; Rasley *et al.*, 2020; Rajbhandari *et al.*, 2020) and Flash-attention (Dao *et al.*, 2022) with gradient-checkpointing (Chen *et al.*, 2016) at bf16 precision. We use AdamW as the optimizer (with fused gelu), the learning rate was kept 2e-5 for all experiments. The maximum context length for image-only datasets is 500, including public video datasets is 800 and including our dataset is 2048. The corresponding batch sizes are 32,16,8. The gradient accumulation is set to 1 and weight decay is disabled. The warmup steps are set to 20 and residual dropout was kept at 0.25. We train all models for two epochs, but use the checkpoint with best validation spearman correlation.

For all experiments, where we combine datasets, we use a custom sampler to account for dataset imbalance, that ensures a maximum proportion of the dataset in an epoch, here are the maximum proportions. For validation we take 5% of each dataset. We use the provided test splits for public datasets and we use a 15% test split for our dataset

5.1.14.1.1 Images

1. **Lamem** 50%

2. **Memcat** 100%
3. **SUN** 100%

5.1.14.2 Videos

1. **VideoMem** 75%
2. **Memento** 75%
3. **AdsData** 100%
4. **MediaEval** 100%

5.1.14.3 Generating Memorable Ads

All the experiments were conducted on 8x80 A100 instances. All experiments were performed leveraging DeepSpeed ZeRO stage-2, Flash Attention and Gradient-Checkpointing. $\alpha = 0.001$, awac_scale= 1, $\gamma = 0.99$, $\beta = 0$ cql_scale= 0.1

5.1.14.4 Inference hyperparameters

$\beta = 4$, temperature= 0.8, steps_for_target_sync 10, $\tau = 0.7$, two_qs: True, lr=1e-5

5.1.15 License and Terms of Release

LAMBDA and UltraLAMBDA are sourced from brand videos from YouTube, Facebook Ads, and CommonCrawl. The dataset annotations and video links contained in LAMBDA and UltraLAMBDA will be released under CC BY-NC 4.0 license. The videos themselves are released as per their creators' licenses. The videos or the released data do not contain or disclose any identities of their annotators or any specific persons. Since it is handcrafted, LAMBDA makes sure that none of the videos are offensive; UltraLAMBDA being sourced from the internet is noisier. While the videos themselves originate from brands, the content of some brands may seem offensive to certain people.

We used Llama, GMHRA, ViT, EVA-CLIP, and Qformer models in accordance with their licenses to train Henry.

5.1.16 Limitations and Potential Risks

In this paper, we try to fill a gap in the existing literature about long-term memorability modeling and datasets. Therefore, we conduct the first study for that purpose. While doing that, we have made initial efforts starting with the English language advertisements. Future work would be needed to address other languages. Further, given the limitations of the study, we conducted it in an academic environment with a student population consisting of undergraduate and graduate student volunteers. We will expand the scope to a wider audience in the future work. We trained a model, Henry, on the collected dataset, showing good performance on all literature datasets. However, since the literature datasets are all English-based and deal with a majorly uniform population, the training will be scaled to more languages and population types in future work. We also observed a decrease in performance for brands not seen during the training and for videos with longer verbalizations exceeding 1500 tokens. Additionally, the model exhibits a slight inaccuracy when advertisements have significant musical content. In our opinion, the model does not pose any potential risk or harm besides the limitations mentioned here. We also conduct a review of the generated ads through experts and non-expert annotators. Both experts and non-expert annotators preferred Henry-SEED generated ads 3/5 times.

5.2 Measuring And Improving Engagement Of Text-to-Image Generation Models

Recent advances in text-to-image generation have achieved impressive aesthetic quality, making these models usable for both personal and commercial purposes. However, in the fields of marketing and advertising, images are often created to be more engaging, as reflected in user behaviors such as increasing clicks, likes, and purchases, in addition to being aesthetically pleasing. To this end, we introduce the challenge of optimizing the image generation process for improved viewer engagement. In order to study image engagement and utility in real-world marketing scenarios, we collect *EngagingImageNet*, the first large-scale dataset of images,

along with associated user engagement metrics. Further, we find that existing image evaluation metrics like aesthetics, CLIPScore, PickScore, ImageReward, *etc.* are unable to capture viewer engagement. To address the lack of reliable metrics for assessing image utility, we use the *EngagingImageNet* dataset to train *EngageNet*, an engagement-aware Vision Language Model (VLM) that predicts viewer engagement of images by leveraging contextual information about the tweet content, enterprise details, and posting time. We then explore methods to enhance the engagement of text-to-image models, making initial strides in this direction. These include conditioning image generation on improved prompts, supervised fine-tuning of stable diffusion on high-performing images, and reinforcement learning to align stable diffusion with *EngageNet*-based reward signals, all of which lead to the generation of images with higher viewer engagement. Finally, we propose the *Engagement Arena*, to benchmark text-to-image models based on their ability to generate engaging images, using *EngageNet* as the evaluator, thereby encouraging the research community to measure further advances in the engagement of text-to-image modeling. These contributions provide a new pathway for advancing utility-driven image generation, with significant implications for the commercial application of image generation.

5.2.1 Introduction

Machine learning models that interact with humans are built as a means to achieve an end, and performance metrics in their respective fields reflect how effectively these models meet the ends. For instance, recommendation systems are optimized to capture maximum viewer interest and the key performance metrics tracked by the research community are clickthrough rates and the number and ranking of relevant documents recommended out of the total document set (Bobadilla *et al.*, 2013). Similarly, chat assistants are optimized for being helpful, and the commonly tracked metrics are the scores of responses preferred by humans (Ouyang *et al.*, 2022; Stiennon *et al.*, 2020). In the case of image generation, industries such as e-commerce, fashion, education, and advertising aim to optimize user-focused outcomes like clicks, purchases, retention, and user engagement. However, the metrics used by the image generation research community often emphasize *aes-*

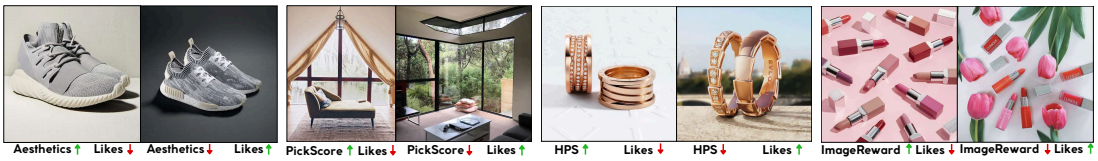


Figure 5.13: Some images from the EngagingImageNet dataset. We constructed pairs of similar images posted within a 45 days interval by the same account. In each pair shown in the figure, the left image corresponds to lower likes and the right one received higher likes. However, existing image generation metrics like Aesthetics, PickScore, Human Preference Score, ImageReward, *etc.*, exhibit image preference in the opposite direction as actual user engagement.

thetic appeal (Xu *et al.*, 2024; Kirstain *et al.*, 2023; Black *et al.*, 2023) and *realism* (Dhariwal and Nichol, 2021; Saharia *et al.*, 2022; Ho *et al.*, 2020; Rombach *et al.*, 2022) factors that crucial for image acceptability but not necessarily aligned with the ultimate goals of viewer engagement.

We find that popular image generation metrics such as Aesthetics (Schuhmann *et al.*, 2022), ImageReward (Xu *et al.*, 2024), Human Preference Score (HPS) (Wu *et al.*, 2023*c*), and CLIP-H (Radford *et al.*, 2021) have a correlation ranging from 0.02-0.08 with user engagement measured by likes, roughly equal to random chance (Table 5.10). Fig. 5.13 illustrates this effect through some randomly picked high and low engagement image samples. Further, one may think that the preferences of image creators (*e.g.*, in the form of upvotes on platforms like Pick-a-Pic (Kirstain *et al.*, 2023) or Discord (Wu *et al.*, 2023*c*)) are a good estimate of image-consumer engagement. However, we find that PickScore and HPS, the reward models trained on a large dataset of creator preferences, correlate 0.07 with user engagement. Therefore, there is a lack of reliable metrics capturing viewer engagement on images.

The lack of progress can largely be attributed to the absence of a large and open dataset of customer engagement metrics over images. The most common image generation datasets, MS-COCO (Lin *et al.*, 2014) and LAION (Schuhmann *et al.*, 2022), contain no signals for user engagement. Therefore, to spur research in the direction of measurement and optimization of image generation for user engagement, we curate a large-scale dataset, **EngagingImageNet**. EngagingImageNet (§5.2.2) consists of 168 million tweets capturing 17 years of high-quality enterprise images for over ten thousand brand accounts and average user engage-

ment of images in the form of likes[‡‡]. We release EngagingImageNet to serve as a starting point for measuring, benchmarking, and modeling large-scale engagement-optimized image generation.

EngageNet as a scoring function to score engagement: EnagingImageNet allows us to train a scoring function that estimates the user engagement on a particular generated image. We formulate this problem as simulating the engagement in the form of user likes over an image-containing tweet (§5.2.3.2). We carry out visual instruction finetuning of LLaVA-1.5 13B (Liu *et al.*, 2023*a*) model to estimate the brand-normalized likes given the image along with contextual information that includes input account handle, image description, and time of the tweet. We find that the resulting scoring model, **EngageNet**, achieves a high correlation of 0.62 with actual user engagement.

Engagement Arena: Next, leveraging EngageNet as a judge, on the lines of LMSYS arena (Zheng *et al.*, 2024; Chiang *et al.*, 2024), we propose **Engagement Arena**, an arena where we test the engagement of images generated by various image generation models for the same prompt. Using EngageNet's reward estimates, we compute Elo ratings of a number of popular open-source text-to-image generation models, including Stable Diffusion-3 (Esser *et al.*, 2024), Flux.1-dev (Labs, 2024), Stable Diffusion-XL (Podell *et al.*, 2023), SDXL-DPO (Wallace *et al.*, 2024), PixArt-alpha (Chen *et al.*, 2024*b*), Pixart-sigma (Chen *et al.*, 2024*a*), Stable Diffusion 2.1 (Rombach *et al.*, 2022), *etc*, and closed-source models like DALL.E-2 (Ramesh *et al.*, 2022). Further, we encourage the research community to adopt Engagement Arena as a basis for measuring further advances in the engagement capabilities of text-to-image modeling and incorporating user engagement into the learning process.

Optimizing the Image generation process with the goal of increasing engagement: Finally, we explore train-time and run-time methods to induce the goal of engagement in the text-to-image generation process: (1) Run-time: conditioning the diffusion model on prompts aligned with higher user engagement, (2) Train-time: fine-tuning the diffusion model on high-engagement images, and (3) Train-time: aligning the diffusion model with EngageNet-based rewards via

[‡‡]EngagingImageNet was collected using Twitter academic API over a period of several years.

reinforcement learning. We present the results of these experiments in Section 5.2.4 and report the efficacy of each method in generating more engaging images.

To summarize, we make the following contributions:

1. We introduce the problem of engagement-optimized image generation. Images, especially in industries like advertising, fashion, and e-commerce, are created to achieve user engagement in the form of clicks, likes, and purchases. Therefore, the image generation process needs to be biased on the image's eventual utility, in addition to the common goals of high aesthetics and fidelity.

2. We curate **EngagingImageNet**, a large-scale, high-quality dataset consisting of user engagement over images. EngagingImageNet consists of 168 million tweets collected from 10,135 enterprise Twitter accounts from the time period 2007 to 2023. It consists of the account name, tweet text, media posted with the tweet, image captions, keywords, colors and tones, the time of posting, and the number of likes the image received. The dataset is instrumental in our study of image engagement as the utility in real-world marketing scenarios.

3. We train an engagement-aware vision language model (VLM), called **EngageNet**, to predict user engagement over images. EngageNet exhibits strong performance in estimating user engagement compared to other commonly used metrics like FID and aesthetics for evaluating the performance of text-to-image generation models as well as state-of-the-art LLMs like GPT-3.5 and GPT-4V.

4. Using EngageNet's predicted engagement scores as a reward, we introduce **Engagement Arena**, the first automated arena to benchmark the engagement of text-to-image models. We rank several popular text-to-image models on their ability to generate engaging images and further encourage the community to submit their models to the arena.

5. We demonstrate introducing the goal of engagement in the text-to-image generation process. We present several approaches to achieve this. These include conditioning of text-to-image generation on prompts corresponding to high user engagement, supervised fine-tuning of stable diffusion on high-engagement images, and reinforcement learning to align stable diffusion with EngageNet-based rewards, all of which lead to the generation of more engaging images to varying degrees.

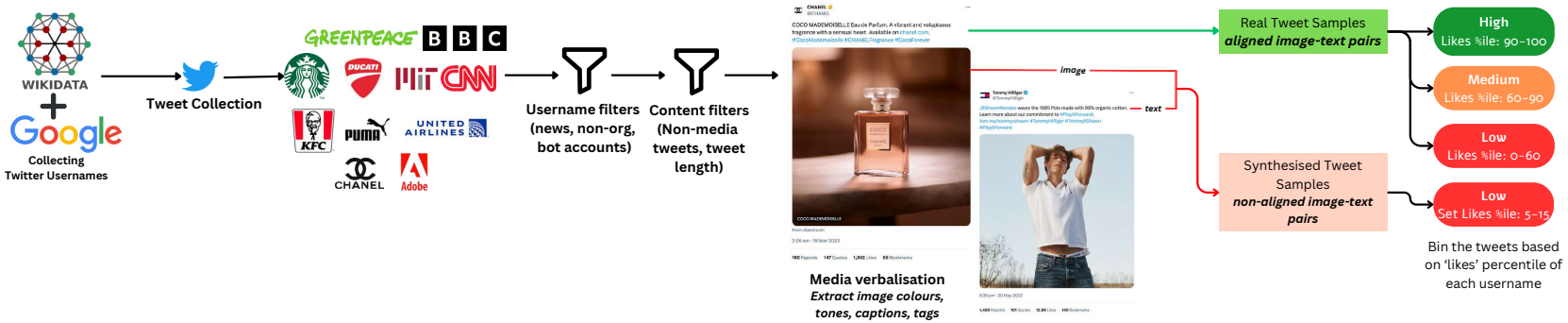


Figure 5.14: Figure illustrating the steps involved in the creation of the EngagingImageNet dataset.

5.2.2 *EngagingImageNet*: Dataset With In-The-Wild User Engagement

To gain insights into image engagement and align text-to-image generation with user engagement, we start by collecting a large dataset of user engagement over images. Our data collection method involved leveraging Twitter, a platform extensively utilized by brands for various purposes such as ongoing product campaigns, sales, offers, discounts, brand building, and community engagement (Alalwan *et al.*, 2017). Twitter user engagement metrics encompass user likes, retweets, comments, mentions, follows, clicks on embedded media, and links. However, the Twitter API provides access to only user likes, retweets, and comments for a given post, with access to comments necessitating a separate and costly call. Therefore, utilizing academic API licenses, we extracted the following data from Twitter: Tweet ID, company name, username, timestamp, tweet text, media files, and user likes.

We focus on enterprise handles for our data collection efforts since the content released by enterprises has the explicit goal of user engagement and is relatively much cleaner than user-generated content. We began by compiling a comprehensive list of company names using the Wikidata knowledge graph (Wikidata contributors, ongoing), focusing on entities categorized as 'business' or 'enterprise'. We conducted Google searches to gather a list of all associated accounts for these companies. For example, for Adobe, this encompassed accounts like Adobe, Adobe Photoshop, Adobe Lightroom, Adobe Experience Cloud, and so forth. This method enabled us to amass a total of 10,135 enterprise Twitter handles. We then utilized the Twitter API to retrieve tweets posted by these enterprises spanning from 2007 to the closure of the Twitter API in January 2023. This effort resulted in the collection of 168 million tweets over a 17-year period, with 28.5 million of these tweets featuring various forms of media, including GIFs, images, and videos. Fig. 5.22 shows several examples of media and tweets present in the EngagingImageNet.

Next, for each username, we bin the tweets falling in the bottom 60 percentile (and having absolute likes > 20), 60-90 percentile (and having absolute likes > 30), and 90-100 percentile (and having absolute likes > 40) of all tweets per account

Table 5.9: A comparison of datasets containing image preferences

Dataset	Size
Pick-a-Pic (Kirstain *et al.*, 2023)	968,965 rankings originated from 66,798 prompts and 6,394 users
Human Preference Score (Wu *et al.*, 2023*c*)	Total of 98,807 images generated from 25,205 text prompts
ImageReward (Xu *et al.*, 2024)	Annotations for 8878 text prompts and corresponding model outputs sampled from DiffusionDB, resulting in 136,892 compared pairs
EngagingImageNet (Ours)	**28.5 million** tweets containing media, captions, colors, tones, and objects, and with user likes as engagement metric

based on the number of user likes. These buckets are subsequently referred to as 'low', 'medium' and 'high' liked buckets, respectively. The resulting dataset, EngagingImageNet, consists of 837,532 samples, having 144,905 high-liked images, 336,200 medium-liked images and 356,427 low-liked images. The high-liked tweets had an average of 2435 likes, while low-liked tweets had an average of 132 likes. Subsequently, all images were verbalized by extracting their captions using LLaVA (Liu *et al.*, 2023*a*), colors and tones along with their coverage using (Qin *et al.*, 2020*a*). Further details regarding the data processing are provided in Appendix 5.2.7.4.

5.2.3 *EngageNet*: Measuring Image Engagement

In this section, we cover the alignment of existing metrics with viewer engagement and design a model to measure the progress of image generation models with the goal of engagement.

5.2.3.1 Alignment of Existing Models with Viewer Engagement

In order to measure the engagement potential of generated images, we first check the alignment of the most popular existing metrics used for evaluating text-to-image models with viewer engagement. For this, we calculate the Pearson correlation of Aesthetic score (Schuhmann *et al.*, 2022), CLIP score (Radford *et al.*, 2021), Pickscore (Kirstain *et al.*, 2023), ImageReward (Xu *et al.*, 2024), and Human Preference Score (HPS) (Wu *et al.*, 2023*c*) with ground truth brand-wise

normalized user likes (0-100) from EngagingImageNet. Table 5.10 presents the results of this analysis. Clearly, the existing metrics are not aligned with user engagement.

This non-alignment can be attributed to the following reasons. Models like PickScore, HPS, and ImageReward are optimized for creator preferences captured on platforms like Discord or custom web applications rather than user (viewer) preferences. The feedback from image creators or communities on these platforms tends to reflect artistic or stylistic biases that do not necessarily correlate with user engagement metrics like clicks, likes, or shares. Further, these models evaluate images in isolation, without considering the contextual information about the image, such as company, time of releasing the image, *etc.*, reducing their effectiveness in predicting user engagement. This suggests that the existing metrics are not designed to capture the user engagement of images. Recent studies have employed the CLIP model as a proxy for human judgment (Nichol *et al.*, 2022; Rombach *et al.*, 2022), aiming to assess the alignment between generated images and text prompts. CLIP, trained on a diverse dataset, is thought to better capture nuanced aspects of human intention. However, similar to ImageReward and PickScore, the text prompts for CLIPScore come from image creators rather than users (viewers), which again is ineffective in conforming to viewers' expectations. The aesthetic score, built on pre-trained CLIP, is trained on several datasets capturing image aesthetics. However, as Fig. 5.13 shows, viewer engagement is much more nuanced than what aesthetics can capture.

Next, we try in-context learning with GPT-3.5 (Ouyang *et al.*, 2022) and GPT-4-Vision (OpenAI, 2023) to predict viewer engagement over images. For GPT-3.5, we supply the image verbalization along with the Twitter handle and posting time, and for GPT-4-Vision, we give the actual image, the image verbalisation, the Twitter handle and posting time. We find that neither GPT-3.5 nor GPT-4 are able to predict user engagement accurately.

5.2.3.2 EngageNet Model To Align With Viewer Engagement

Since the existing approaches do not show acceptable performance for predicting user engagement over images, we, therefore, train our own engagement-aware

Table 5.10: Pearson correlation between model predicted scores and image engagement measured by account normalized likes.

Model	Configuration	Pearson Correlation
PickScore (Kirstain *et al.*, 2023)	-	0.0734
ImageReward (Xu *et al.*, 2024)	-	0.0285
Human Preference Score (Wu *et al.*, 2023*a*)	-	0.0747
Aesthetic Score (Schuhmann *et al.*, 2022)	-	0.0674
CLIP Score (Radford *et al.*, 2021)	-	0.0423
GPT-3.5 (Ouyang *et al.*, 2022)	3-shot In-context learning	0.0464
GPT-3.5 (Ouyang *et al.*, 2022)	5-shot In-context learning	0.0351
GPT-4V (OpenAI, 2023)	3-shot In-context learning	0.1453
GPT-4V (OpenAI, 2023)	5-shot In-context learning	0.1264
EngageNet	Trained on random KPI, Tested on actual KPI	0.0617
EngageNet	Trained without MSE loss	0.5821
EngageNet	Trained without date input	0.5365
EngageNet	Trained without company input	0.5226
EngageNet	Trained without date and company input	0.4476
EngageNet	Trained without negative samples	0.6051
EngageNet	Trained with MSE loss and negative samples	0.6248
EngageNet	Oracle	0.8682

vision-language model (VLM) model, EngageNet. To this end, we perform visual instruction fine-tuning of LLaVA-1.5 (Liu *et al.*, 2023*a*) on the EngagingImageNet train dataset (Figure 5.15).

The core of EngageNet is its ability to condition its output on a diverse set of inputs, including behavioral signals. The model is architected to predict the normalized likes of an image on a 0-100 scale. This prediction is conditioned on a unified input representation that merges visual information with various forms of metadata, including behavioral cues. The conditioning mechanism operates as follows. First, the input image is processed by the VLM's vision encoder to produce a sequence of visual embeddings. Concurrently, we process the metadata, which consists of marketer-provided information such as the company name, image resolution, color palettes, descriptive tags, and the social media release date. These categorical and textual metadata are converted into a uniform text-based format. For instance, numerical data like image resolution are described textually (e.g., "resolution 1024 by 1024 pixels"). All textual metadata, including these converted signals, are concatenated into a single string.

This combined text string is then tokenized to create a sequence of textual embeddings. These are subsequently prepended to the visual embeddings de-

rived from the image. This concatenated sequence of text and vision embeddings forms the final input to the LLaVA-1.5 language model component. The model's self-attention mechanism is then responsible for fusing these heterogeneous signals—learning the complex interplay between the visual content and the associated metadata—to predict the engagement score. To enhance the model's robustness and its reliance on visual features, we augment the EngagingImageNet dataset with synthetic negative samples. For 25% of tweets from high and medium engagement tiers, we pair the tweet's metadata with an unrelated image from a different tweet. The corresponding "likes" for these mismatched pairs are set to a low value (randomly sampled from 5-15). This procedure encourages the model to penalize images that are semantically irrelevant to the textual context, thereby improving its discriminative ability. After this augmentation, our final dataset comprises 957,809 samples, which we split into training and testing sets.

We design an instruction (Listing 5.17) for the VLM to predict the normalized likes of an image on a 0-100 scale, also conditioned on metadata comprising the marketer (company), image resolution, image colours and tones with their spatial coverage, image description and tags, and the date of releasing the image on social media.

Since EngageNet is trained to predict the KPI of an image, we additionally model the problem as a regression task. We attach a two-layered MLP network on top of the last layer of hidden states of the decoder module to predict the scalar KPI from EngageNet. Therefore, while typically language models are trained on cross-entropy loss L_{CE}, we also use mean squared error L_{MSE} as an auxiliary loss to train EngageNet. This is because L_{MSE} is more sensitive to the difference between the predicted and actual KPI values, which is crucial to better guide EngageNet to learn the image KPI prediction task. Thus, the final loss function for EngageNet is given by:

$$L_{EOIG} = L_{CE} + \lambda L_{MSE} \tag{5.3}$$

where λ is a hyperparameter that controls the weight of the auxiliary loss. We set $\lambda = 0.1$ in our experiments. We find that EngageNet demonstrates strong performance at predicting user engagement over images, achieving a Pearson correlation

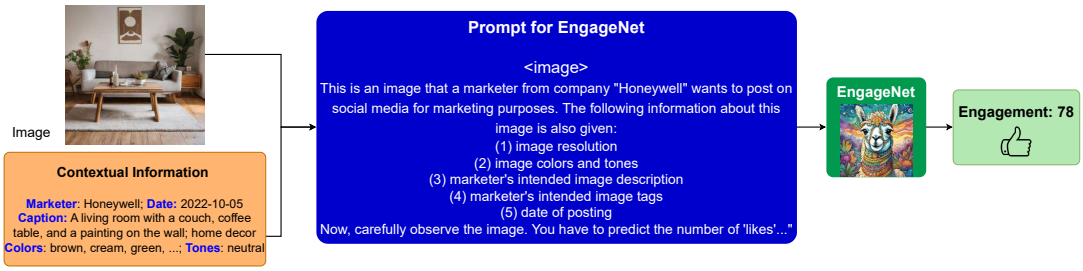


Figure 5.15: Visual Instruction Finetuning of EngageNet on EngagingImageNet dataset. The EngageNet model is trained to predict the KPI of an image on a 0-100 scale, conditioned on marketer provided metadata comprising the company, image resolution, image colours and tones with their spatial coverage, marketer's intended image description and tags, and the date of releasing the image on social media.

of 0.62 with ground truth user likes (Table 5.10).

We perform an ablation study to understand the impact of different components of the instruction on the performance of EngageNet. If EngageNet is not supplied with contextual information such as marketer company and the time of posting the image in the input, the correlation of its predictions with ground truth account normalized likes drops significantly (0.62 to 0.44). This indicates that the company and the time of posting are important components of the instruction for EngageNet to predict user engagement accurately. We also attempt to determine the impact of the auxiliary MSE loss adopted for training EngageNet. The MSE loss increases the correlation from 0.58 to 0.62, indicating that the auxiliary loss improves the performance of EngageNet. The MSE loss makes the model more sensitive to the difference between predicted and actual scores.

We also conduct an experiment to investigate the signal present in the EngagingImageNet dataset. For this, we train EngageNet on the EngagingImageNet dataset but with randomly sampled KPI values. We then evaluate the model on the EngagingImageNet test dataset with actual KPI values. In this case, the correlation of EngageNet's predictions with ground truth user likes is nearly zero, indicating that the KPI values in the EngagingImageNet dataset are crucial for EngageNet to learn predicting user engagement accurately.

Additionally, we evaluate the impact of adding negative samples to the training data. These samples are constructed by sampling images and other inputs in the instruction from different tweets, such that they are not aligned, and then set-

ting the normalized likes to a very low value. Although we find that the addition of negative samples does not significantly impact the correlation of EngageNet, however it does help in improving the robustness of the model. This is because EngageNet learns to penalize images that are not aligned with the other inputs in the instruction. This is crucial for leveraging EngageNet as a reward model for engagement-optimized image generation as described in Section 5.2.4.3. Since we propose to also utilize EngageNet as an oracle for ranking models in the Engagement Arena, we train EngageNet on the entire EngagingImageNet dataset, *i.e.*, with both train and test data. In this configuration, EngageNet accomplishes a high correlation of 0.87 with ground truth user likes, which establishes its effectiveness to be used as an oracle.

5.2.4 Methods to Improve Image Engagement

We explore three methods for optimizing the text-to-image generation process to generate more engaging images. These include run-time and train-time optimizations: conditioning of text-to-image models on better prompts, supervised fine-tuning of stable diffusion on high-liked images, and reinforcement learning to align stable diffusion with EngageNet-based reward scores. The first method operates in the natural language domain at run-time, generating a description of how an engagement-optimized image should look like. On the other hand, the other two operate in the vision domain, generating actual engagement-optimized pixels by training the U-Net module of stable diffusion. We cover each of them next.

5.2.4.1 Conditioning Stable Diffusion on More Engaging Prompts

In the EngagingImageNet dataset, we observe that some images having similar themes but different details received vastly different levels of user engagement. For instance, consider the following pair of image captions: (1) "A living room having a couch and coffee table with a rug in front." (2) "A living room with large windows having a couch, coffee table and a rug" Despite both images depicting a similar scene (Figure 5.23), the first image received low engagement, while the

second image garnered high engagement. In this case, the difference in engagement can likely be attributed to the presence of elements, such as large windows and natural light in the second image, which makes the living room appear bigger and more appealing to a viewer. Such observations motivated us to exploit patterns related to certain image aspects that can boost engagement. We further extend this analysis for images posted by a few companies in Appendix 5.2.7.3.

Therefore, in this method (Figure 5.23), we attempt to alter the text prompts fed to the diffusion model such that the improved captions incorporate characteristics that have been empirically shown to boost image performance. For this, we adopt a retrieval framework described as follows. Using FAISS (Johnson *et al.*, 2019; Douze *et al.*, 2024), we index the vector embeddings of captions belonging to images in the high performance data subset of the EngagingImageNet train data as described in Section 5.2.2. Next, for every image caption in the low performance subset of the test data, we retrieve the semantically most similar caption from the corpus of high-performing images. If the similarity level is above a certain threshold τ, the retrieved captions thus obtained are passed as input to the diffusion model for generating more performant images, otherwise the original caption is used for image generation.

Table 5.11: Results reveal the significant gains achieved in improving the engagement of low-liked subset of the EngagingImageNet dataset by enhancing the image descriptions fed to a text-to-image model, as described in Section 5.2.4.1.

Images	**Training Config**	**Oracle Engagement Reward**		**Engagement reward increase**
		w/o prompt improvement	w/ prompt improvement	
DALL.E-2	N.A.	42.1459	45.7077	8.45%
SD 1.4	N.A.	38.7680	43.8173	13.02%
EOIG SD 1.4	RLHF-ES	40.2037	46.1218	14.72%
EOIG SD 1.4	RLHF-DSG	39.4950	44.8116	13.46%
EOIG SD 1.4	Preferred Finetuning (PFT)	43.1206	46.6490	8.18%
SD 1.5	N.A.	39.4001	44.3287	12.51%
SD 2.1	N.A.	45.4952	49.6946	9.23%
Pixart-alpha	N.A.	44.7305	49.3322	10.29%
Pixart-sigma	N.A.	46.2870	51.4671	11.19%
SD XL	N.A.	49.1094	53.8602	9.67%
SD XL - DPO	N.A.	51.3786	54.5863	6.24%
SD 3 Medium	N.A.	50.6578	55.0841	8.74%
Flux.1-dev	N.A.	48.7226	53.7209	10.26%
Ground Truth	N.A.	41.2152	56.7533	37.70%
		Average Increase in Engagement		**12.41%**

5.2.4.2 Preferred Finetuning on High-Engagement Images

Several prior studies have demonstrated the feasibility of learning styles through stable diffusion by fine-tuning the model (Pinkney, 2022; Cjwbw, 2022; PromptHero, 2023; Everaert *et al.*, 2023). These approaches typically involve fine-tuning the U-Net architecture within the Stable Diffusion framework using a set of images exhibiting the desired style. For instance, Everaert *et al.* (2023) proposed a method to finetune Stable Diffusion to adapt it to target styles like *anime sketches*, *American comics*, *Pokemon*, *starry night*, *etc.*

In this work, we attempt to explore whether the diffusion model can learn patterns associated with higher user engagement, analogous to learning visual styles. To this end, we performed fine-tuning of the base Stable Diffusion U-Net on the preferred data distribution, containing high liked images sampled from the EngagingImageNet train set (Figure 5.24). We call this process, **Preferred Finetuning**. The model was finetuned for 50 epochs, following the procedure outlined by von Platen *et al.* (2023). The model minimizes the standard denoising score matching loss (Ho *et al.*, 2020; Ho and Salimans, 2022), which measures how well the model predicts the noise added to the image during the diffusion process:

$$L_{\text{denoise}} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\|^2\right] \tag{5.4}$$

where $\mathbf{x}_0$ is the original image, $\mathbf{x}_t$ is the noisy image at time step t, generated by adding noise $\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$ is the predicted noise from the model given the noisy image $\mathbf{x}_t$, time step t and conditioning information $\mathbf{c}$ *i.e.*, text prompt fed as input to the diffusion model.

5.2.4.3 Aligning Stable Diffusion With Engagement

Black *et al.* (2023) proposed denoising diffusion policy optimization (DDPO), a policy gradient algorithm which frames the denoising process as a multi-step decision-making problem. The authors showed that DDPO can be employed to finetune text-to-image diffusion models to align their outputs with a variety of reward functions including image compressibility, aesthetic quality and image-

prompt alignment, among others. Therefore, we explore the use of reinforcement learning to optimize diffusion models to improve the engagement potential of their generated images. To this end, we leverage EngageNet as a reward model to align a pre-trained stable diffusion model using DDPO algorithm to produce more engaging images. The entire process of alignment is shown in Figures 5.26, 5.25 in the appendix. In DDPO, the denoising process is viewed as a finite horizon Markov decision process, where the state comprises of the current context, number of steps left in the process and the current denoised image. The action to be taken is to predict the next image using this state.

We experiment with two types of reward functions for finetuning stable diffusion:
(1) Engagement Simulation (ES): We leverage EngageNet to estimate the user engagement of images generated by stable diffusion. The reward signal is used to guide stable diffusion to generate higher engagement images as illustrated in Figure 5.25. The resulting diffusion model is called EOIG-SD (RLHF-ES).
(2) Design Specification Generation (DSG): We train an alternate version of EngageNet to produce the design specification of an image, based on conditioning factors such as the company, time, image caption and viewer likes. This model learns to predict verbalized image descriptions comprising colors and tones with their spatial coverage, as well as objects with their locations, that should be reflected in an image, for a given engagement level and caption. The detailed method and results of EngageNet trained on this task are explained in Appendix 5.2.7.7. Next, we utilise this EngageNet as a reward model to train stable diffusion such that the images generated by it have a design specification aligned with those of higher engagement images as shown in Figure 5.26. EOIG-SD (RLHF-DSG) takes a text prompt and generates an image, which then undergoes verbalization via image perception models. Its objective is to create images that, when verbalized, closely resemble the engagement-conditioned verbalization generated by EngageNet. Thus, we ask EngageNet to provide the logits for this image verbalization, using which a reward is computed for EOIG-SD, indicating how closely this verbalized output aligns with EngageNet. This reward value serves as feedback for EOIG-SD in the form of policy gradient, aiding in its continual improvement and refinement within the image generation process. Only high engagement samples are used in the training process. The details of this method are described in



Figure 5.16: Comparison of generated images - EOIG-SD *vs* Base stable diffusion. Engagement optimization helps the model to learn to generate persuasion skills. EOIG-SD generates better product photography (a,c,d), model photography (b), generates images with social appeal and social identity (a,c), and learns temporal patterns (e) (prominently Christmas-themed image of dog)

Appendix 5.2.7.8.

5.2.4.4 Evaluating the Methods Adopted for Engagement-Optimization

Run-time optimization: Firstly, we investigate the impact of using better prompts to condition the text-to-image generation process as described in Section 5.2.4.1. The results are summarised in Table 5.11. By retrieving semantically similar captions from the corpus of high-liked images, visual characteristics that have been empirically shown to enhance image engagement get incorporated in

Table 5.12: Comparing the performance gains on the EngagingImageNet test dataset, resulting from train-time engagement-optimization methods applied on stable diffusion, as described in Sections 5.2.4.2 and 5.2.4.3.

Images	Training Config	Bucket	Engagement Reward	Engagement Increase	Aesthetic Score	CLIP Score	FID	PickScore
Ground Truth	N.A.	High	90.9526	N.A.	5.1006	32.6343	N.A.	20.9470
		Medium	74.3535	N.A.	5.0940	32.4867	N.A.	20.9351
		Low	41.2152	N.A.	5.0518	32.2012	N.A.	20.7406
SD 1.4	N.A.	High	56.1489	N.A.	5.2029	33.0830	24.6631	17.3514
		Medium	51.6949	N.A.	5.1634	32.9173	23.4434	17.3344
		Low	38.7680	N.A.	5.1662	32.8339	24.2607	17.3262
EOIG SD 1.4	Preferred Finetuning (PFT) on High engagement Images	High	62.0390	10.49%	4.8090	32.4524	24.9370	17.3070
		Medium	56.1082	8.54%	4.8387	32.3923	23.0904	17.3239
		Low	43.1206	11.23%	4.8108	32.2960	23.5885	17.2932
EOIG SD 1.4	RLHF - ES	High	58.2724	3.78%	5.1828	33.2891	23.8656	17.4113
		Medium	53.1004	2.72%	5.1686	33.2845	22.7157	17.3802
		Low	40.2037	3.70%	5.1629	32.9468	24.1780	17.3672
EOIG SD 1.4	RLHF - DSG	High	57.9188	3.15%	5.2495	33.1072	23.9144	17.3577
		Medium	52.9765	2.48%	5.2187	33.0991	23.3626	17.3486
		Low	39.4950	1.88%	5.2336	32.9716	24.2147	17.3277

the text prompt. Therefore, after applying this method, we observe a significant improvement in the engagement of low-liked subset of the EngagingImageNet test dataset, consistent across multiple text-to-image models, both open-source and closed-source. This method is highly effective as it is able to produce images with higher engagement without any additional training of the diffusion model. We observe that, on average, an improvement in the prompt results in an improvement of 12.4% in engagement. The improvement is observed in models across all sizes and also for models trained on high-engagement images (EOIG-SD).

Train-time optimization: Next, we present the results of the methods (§5.2.4.2, §5.2.4.3) adopted for engagement-optimized image generation by training the U-Net module of stable diffusion in Table 5.12. We denote all the models trained using train-time optimizations like Preferred fine-tuning with EOIG (engagement optimized image generation). We compare the performance of the base stable diffusion model (SD 1.4), stable diffusion finetuned on high-engagement images (EOIG-SD PFT), and stable diffusion aligned with EngageNet-based reward functions (EOIG-SD RLHF-ES, EOIG-SD RLHF-DSG). For this, we use EngageNet-Oracle as a judge to predict the user engagement of the images generated by these models. This helps us probe the effect of different training strategies on improving the engagement capabilities of stable diffusion. Consistent with prior literature, we also include other metrics like FID (Heusel *et al.*, 2017), aesthetics (Schuhmann *et al.*, 2022), CLIP score (Radford *et al.*, 2021), and PickScore (Kirstain *et al.*, 2023).

The results indicate that while all the training methods improve the engagement capabilities of stable diffusion, however, the extent of improvement varies widely. We find that finetuning the stable diffusion model on preferred data distribution, i.e. high-engagement samples from the EngagingImageNet dataset yields significant gains in the engagement potential of the generated images. This is evident from the consistent increase in the predicted user engagement of the generated images across all engagement buckets. Next, we discover that using EngageNet-based reward functions to align the stable diffusion model also results in better performance. However, the improvement in image engagement is not as significant as that achieved by the previous methods. Other metrics like CLIP score, PickScore and FID do not vary significantly across the EngagingImageNet

buckets and largely remain unaffected after training stable diffusion in both the above regimes. This further corroborates their non-alignment with image engagement.

Next, we discuss the side effects of training stable diffusion using the above methods. As a consequence of training on engaging images, we find that stable diffusion learns to generate images with certain persuasion strategies (Kumar *et al.*, 2023*b*). For instance, Fig. 5.16 shows several examples of product and model photography generated by EOIG-SD and base SD, demonstrating EOIG-SD's biases towards certain persuasion strategies such as social appeal and social identity, commonly observed in marketing scenarios (Kumar *et al.*, 2023*b*) but ignored in general photography.

Combination of Train-time and Run-time optimizations: In our experiments, we gauge the impact of different methods in improving image engagement by comparing the results of both train-time (§5.2.4.2, §5.2.4.3) and run-time (§5.2.4.1) optimisations, as well as their combination. Stable Diffusion 1.4 (Rombach *et al.*, 2022), serves as the baseline model. In Table 5.11, we observe that when each method is applied individually, such as using better prompts at run-time or training the diffusion model through supervised finetuning or using reinforcement learning, it results in measurable improvements in image engagement over the baseline. However, the most significant improvements are seen when supervised finetuning or reinforcement learning is combined with better prompts at run-time. This demonstrates that coupling train-time and run-time optimisations has a synergistic effect, resulting in higher engagement levels than each method applied alone.

5.2.5 *Engagement Arena*: Measuring Engagement Capabilities of Text-to-Image Models

Motivated by the work of LMSYS and similar benchmarks (Chiang *et al.*, 2024), we propose *Engagement Arena* as a platform to evaluate the capability of text-to-image models to generate engaging images. We run a tournament on a common set of prompts from the EngagingImageNet test set. We leverage EngageNet as an

oracle for *Engagement Arena* to compute the Elo ratings of various open-source text-to-image models, such as Stable Diffusion 3 Medium (Esser *et al.*, 2024), Flux.1-dev (Labs, 2024), Stable Diffusion XL (Podell *et al.*, 2023), Stable Diffusion XL-DPO (Wallace *et al.*, 2024), Pixart-sigma (Chen *et al.*, 2024*a*), Pixart-alpha (Chen *et al.*, 2024*b*), Stable Diffusion 2.1, Stable Diffusion 1.5, Stable Diffusion 1.4, (Rombach *et al.*, 2022), *etc.*, and closed-source models like DALL.E-2 (Ramesh *et al.*, 2022). Figure 5.17 shows the rankings of these models. It also features the Elo ratings of ground truth images to serve as topline for benchmarking the models.

In addition to helping to rank the engagement potential of generated images accurately, using EngageNet as an oracle also avoids having static benchmarks with a definitive ground truth. We encourage the research community to adopt *Engagement Arena* as a basis for measuring further advances in the engagement capabilities of text-to-image modeling and incorporating user engagement into the learning process.

The arena features actual images from the EngagingImageNet dataset as a topline benchmark for the images generated by different text-to-image models. We find that Stable Diffusion 3 Medium (Esser *et al.*, 2024) emerges as the best performing model in the *Engagement Arena*, with a win rate of 46% over actual images (Figure 5.27). It is followed by SDXL-DPO (Wallace *et al.*, 2024) and Flux.1-dev (Labs, 2024). We notice a general trend that image engagement rises with the size of the text-to-image models. However, there are some exceptions to this trend. For instance, Pixart family of models (600M parameters) and EOIG-SD PFT model (860M parameters) surpass relatively larger DALL.E-2 (6.5B parameters).

We observe that while our EOIG-SD models trained using different methods (PFT, RLHF) outperform the equal-sized base SD 1.4 model, however there is a considerable gap between the performance of EOIG-SD and significantly larger text-to-image models leading the arena. This can be attributed to the inherent limitations of SD 1.4 in generating high-quality images, which cannot be fully overcome by the training methods explored in this work.

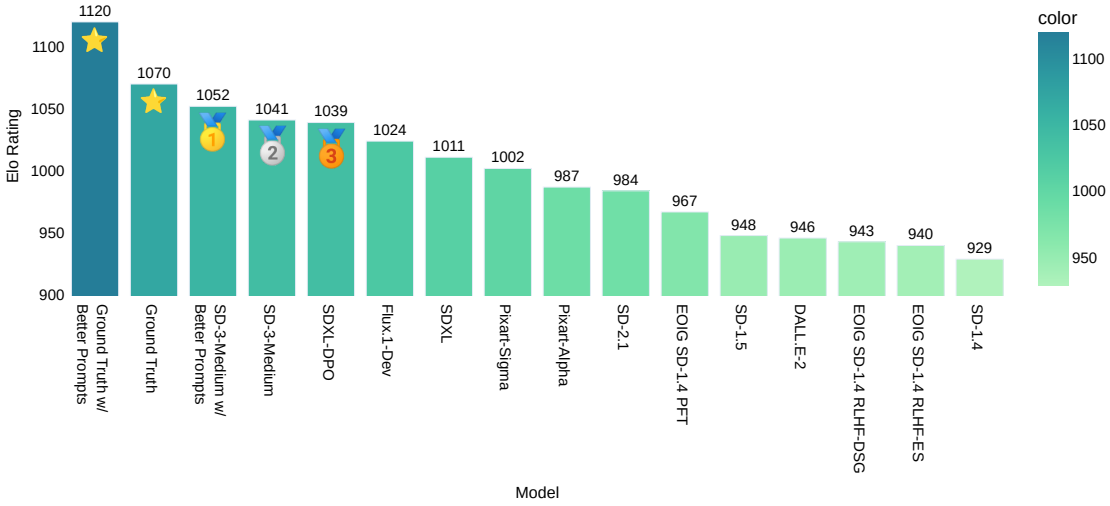


Figure 5.17: Rankings and Elo ratings of various text-to-image models in the proposed Image Engagement Arena.

5.2.6 Conclusion

Image generation technologies have undergone a significant evolution, transitioning from research concepts to viable commercial products. The initial phase of image generation primarily focused on producing higher-quality images that adhere to provided prompts. In the next phase, generated images should not only meet quality standards but also align with the creator's objectives. This necessitates conditioning the image generation process based on the utility of the generated image. In marketing contexts, this utility translates into achieving higher customer engagement metrics such as likes, shares, clicks, and more. In this paper, we introduce the problem statement of engagement-optimized image generation and propose the first large-scale dataset for this purpose. Additionally, we present the results of several techniques to solve this problem both in the natural language domain by generating engagement-optimized text-prompts and in the computer vision space by generating actual engagement-optimized pixels.

5.2.7 Appendix

5.2.7.1 Related Work

A large body of work has been done with respect to generating images from textual descriptions. These text controlled image generation models have evolved greatly

from the time of GANs (Goodfellow *et al.*, 2020) to yield high-quality image generators based on diffusion models such as DALL-E (Ramesh *et al.*, 2021), Stable Diffusion (Rombach *et al.*, 2022) and ones that have extended these models, that are able to follow human text instructions to a large extent. However, the metrics that these generators optimize are Inception Score (IS) (Salimans *et al.*, 2016) and Fréchet Inception Distance (FID) (Heusel *et al.*, 2017). It has been observed by multiple works for example (Kirstain *et al.*, 2023; Wu *et al.*, 2023*c*; Xu *et al.*, 2024) etc. that these metrics do not necessarily correspond to human preferences.

To align models with human preferences, reinforcement learning with human feedback (RLHF) has been successfully used in the LLM literature (Ouyang *et al.*, 2022; OpenAI, 2023; Touvron *et al.*, 2023) with algorithms such as PPO (Schulman *et al.*, 2017), DPO (Rafailov *et al.*, 2024) and several variants of these preference based reinforcement learning algorithms. Similar approaches have also been used in text-to-image generation models (Black *et al.*, 2023; Wallace *et al.*, 2024). In these approaches, the latent image generation part of the diffusion model (either UNet or a Transformer) is trained using a reward model in the case of DDPO or using user preferences directly in the case of DPO. Both these approaches involve collecting human preference datasets.

Several human preference datasets for text-to-image generation have been collected in literature. These include Pick-a-Pic dataset (Kirstain *et al.*, 2023), dataset generated from the Stable Foundation Discord channel (Wu *et al.*, 2023*c*), ImageReward dataset (Xu *et al.*, 2024) etc. The human preferences in these datasets have been collected by explicitly asking humans to state their choices. These datasets are often accompanies with their own metrics for human preference alignment such as PickScore (Kirstain *et al.*, 2023), Human Preference Score (Wu *et al.*, 2023*c*), ImageReward score (Xu *et al.*, 2024) etc. The authors of these papers have shown that these metrics are better aligned with human preferences when compared with text-alignment scores such as CLIP score and BLIP score. However, these need not align with viewer engagement of images and hence, in our work, we use implicit data about human preferences derived from engagement metrics such as clicks, likes etc.

5.2.7.2 Study with marketers

Unlike other NLP and CV tasks where humans are the topline for any model's performance, simulating engagement is a relatively hard task for humans. It has been shown in several studies that expert human opinions fare similar to non-experts (*e.g.*, (Tetlock, 2017; Collaborative, 2023)), and the opinion of the non-expert population is just above a random coin toss for behavior simulation tasks (*e.g.*, (Tan *et al.*, 2014; Isola *et al.*, 2013)). Therefore, simulating engagement necessitates an automatic and reliable method to measure engagement.

Brand	**Correlation Coefficient (r)**	**p-value**
Impressions	0.039	0
Clicks	0.076	2.74e-61
CPC	0.047	2.736e-24
CPM	0.191	0.0
CPP	0.207	0.0

Table 5.13: Pearson correlation coefficients (r) and associated p-values for the relationship between marketer-allocated advertisement budget and five key performance indicators (KPIs): Impressions, Clicks, Cost Per Click (CPC), Cost Per Thousand Impressions (CPM), and Cost Per Purchase (CPP). Budget allocation serves as a proxy for marketer confidence in advertisement efficacy. Data were collected from a Fortune 500 company's marketing campaigns (n > 1,000 advertisements) over a 12-month period. Results suggest no statistically significant correlation between marketing spend and advertisement performance across all measured KPIs, indicating potential limitations in expert marketers' ability to predict advertisement success.

We conducted several studies with both expert marketers and non-experts to estimate their capability to simulate engagement. We worked with a Fortune-500 company expert marketers for this task. Marketers usually have to run multiple advertisements for a single campaign at the same time. We estimated the correlation of their past spend data with several behavioral metrics: impressions, cost per click (CPC), cost per pixel (CPP), cost per 1000 impressions (CPM), and clicks. Table 5.13 shows the results of these studies where we observed that despite being experts in marketing, the budget allocation by these marketers had almost no correlation with any of the key performance indicators.

Human Eval Protocol: Particpants submitted their ideas and they were independently shown the AI generated captions fot these ideas. They are then allowed

to submit their feedback in the form of like or dislike (not compulsorily). Based on their feedback they are further prompted for Reason and Feedback. We filtered the feedbacks that were related to the experimental setup.

5.2.7.3 Analysing Visual Aspects that Drive Engagement

To understand the visual aspects that often lead to higher image engagement, we analysed pairs of images having same theme but different details, posted within a 45 days interval by the same account. Then pairs with vastly different engagement levels between the images were sampled. We then extracted the differences between the image pairs for a few companies using using GPT-4-Vision (OpenAI, 2023). Following are some main observations. For fashion brands like Bulgari, we observe that images featuring prominent branding and dynamic backgrounds with bright colors and gradients significantly enhance engagement, as visible in Figure 5.13 (Image pair-3). For Gucci, engagement is driven by images that maintain a clear focus on the product, emphasizing intricate detailing and textures. Additionally, images that incorporate luxurious backgrounds contribute to higher engagement levels. In the case of Airbnb, images that blend natural light with greenery are particularly effective in enhancing user engagement. Showcasing relatable homestay experiences aligns closely with Airbnb's branding, further driving engagement. Meanwhile, Lenovo benefits from highlighting unique technical features and specifications while utilizing vibrant colors and high-contrast backgrounds.

5.2.7.4 EngagingImageNet Filtering Steps

We sample the tweets posted in the 5 year time period from January 2018 to January 2023. We focus our analysis on usernames that market products or services, and thus weed out usernames belonging to categories like news and sports. Next, we check if the number of tweets posted by a username exceeds 1000, then we retain the username, else we discard remove it. This helps in removal of stray handles and ensures data quality. Further, if the number of tweets posted by a username exceeds 2000, we randomly sample 2000 tweets for this username to avoid oversampling tweets from the same username and thus compromising data

variance. This step ensures that the dataset is fairly representative of different enterprise accounts. Moreover, we weed out all tweets containing media other than images and where tweet text is less than 50 characters. Also, the hyperlinks present in the tweets are masked with a <hyperlink> placeholder. This results in 365,129 tweets posted in 5 years by 592 Twitter handles.

Since it is hard to assign KPI credit to the multiple media present in a single tweet, we assign an equal KPI credit to all the media in a tweet.

5.2.7.5 EngagingImageNet Additional Details

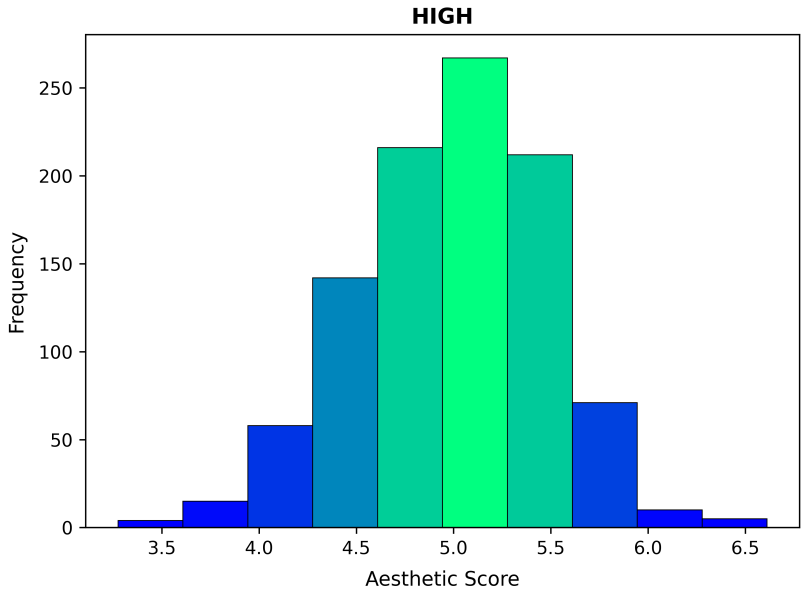
Table 5.14: Distribution of ground truth EngagingImageNet images

KPI	# Objects	Aesthetic Score	CLIP Score
High	3.405	4.994	30.509
Low	3.291	4.881	30.638

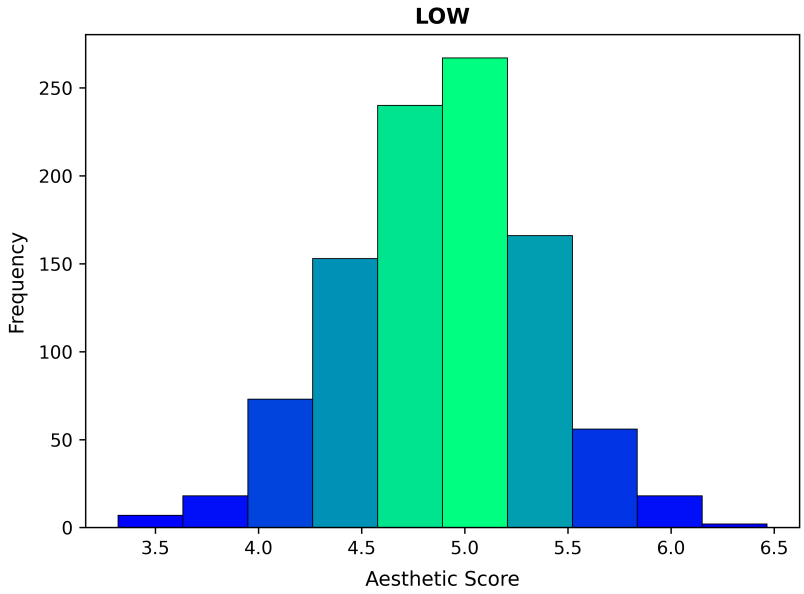
5.2.7.6 Prompts for Instruction Finetuning

Listing 5.17: Visual instruction finetuning Pattern: EngageNet predicts user engagement of images given contextual information about the social media post.

Input: <image>
This is an image that a marketer from company "gucci" wants to post on social media for marketing purposes. The following information about this image is also given:
(1) image resolution i.e. (width, height): [680, 680],
(2) image colors and tones: {"color and tones": {"colors": {"Orange": {"coverage": 0.6}, "White": {"coverage": 0.18}, "Pink": {"coverage": 0.12}, "Brown": {"coverage": 0.1}}, "tones": {"warm": 0.72, "neutral": 0.28, "cool": 0}}},
(3) marketer's intended image description: A girl with a nose ring and gold earrings.,
(4) marketer's intended image tags: nose ring, gold earrings, girl, makeup, lips, face, beauty, earrings, nose, lips, gold, woman, makeup, accessories,
(5) date of posting: 22−February−2019
Now, carefully observe the image. You have to predict the "number of likes" that this image will get, on a scale of 0 to 100.
It measures the number of times the viewers will interact with the social media post by clicking the "Like" button to express their appreciation for the image. Thus, an image with higher visual appeal, alignment with the company's brand identity, and relevance to the audience, is likely to receive more likes. Moreover, a good image should stongly correspond with the marketer's intended image description and tags to attract the target audience.
Your predicted "number of likes" will help the marketer to decide whether to post this image or not on the social media platform.



(a)



(b)

Figure 5.18: Aesthetic Score distribution across High and Low KPI images in EngagingImageNet dataset

Answer properly in JSON format. Do not include any other information in your answer.

Output:
{"likes": 19}

Listing 5.18: Engagement Finetuning Verbalization Pattern (1): Explicitly asking model to pay attention to engagement tokens

Input: You are a smart model. I am giving you some data regarding an image – (1) captions (2) keywords (3) image resolution i.e. (width, height) (4) release date (5) number of downloads i.e. how many times the image was downloaded (6) number of forwards i.e. how many times the image was forwarded to someone

else (7) number of impressions i.e. how many times the image was seen by someone. Note that (5), (6) and (7) are Key Performance Indicators (KPIs) of the image, thus they are important signals of its perceived quality and popularity.

You have to predict following attributes of the image: (1) colour and tones from the lists given below: − Allowed colours: ['Red', 'Dark_Red', 'Green', 'Bright_Green', 'Dark_Green', 'Light_Green', ' Mud_Green', 'Blue', 'Dark_Blue', 'Light_Blue', 'Royal_Blue', 'Black', 'White', 'Off_White', 'Gray', ' Dark_Gray', 'Silver', 'Cream', 'Magenta', 'Cyan', 'Yellow', 'Mustard', 'Khaki', 'Brown', 'Dark_Brown', ' Violet', 'Pink', 'Dark_Pink', 'Maroon', 'Tan', 'Purple', 'Lavender', 'Turquoise', 'Plum', 'Gold', 'Emerald', 'Orange', 'Beige', 'Lilac', 'Olive'] − Allowed tones: ['warm', 'neutral', 'cool'] (2) main objects present in the image and the diagonal coordinates of their bounding boxes: [x1, y1, x2, y2]

Now, predict the attributes for the following image: [captions: ”Waist up portrait of mixed−race female worker posing confidently while standing with arms crossed in plant workshop”, keywords: ”female, worker, young, woman, mixed−race, african, african−american, modern, contemporary, work, occupation, industry, industrial, plant, factory, workshop, work shop, strong, tough, gritty, masculine, short, hair, latin−american, plump, adult, mechanic, repair, repairman, handywoman, foreman, copy space, portrait, looking at camera, standing, posing, smiling, recruitment, employment, job, opportunity, engineer, production, manufacturing, assembly, assembling, line”, image resolution: ”(5760, 3840)”, release date: ”2019−12−02”, number of downloads: ”24”, number of forwards: ”106”, number of impressions: ”5941”] Answer properly in JSON format. Do not include any other information in your answer.

Output: {”color and tones”: {”colors”: {”Gray”: {”coverage”: 0.4}, ”Dark_Gray”: {”coverage”: 0.22}, ”Black”: {”coverage”: 0.14}, ”Off_White”: {”coverage”: 0.13}, ”Silver”: {”coverage”: 0.11}}, ”tones”: {”warm”: 0, ” neutral”: 1.0, ”cool”: 0}}, ”objects”: {”jeans”: [2076.67, 2542.5, 3023.88, 3827.01], ”woman”: [1892.94, 11.18, 4260.09, 3824.34], ”safety vest”: [2160.75, 1410.95, 3668.16, 3826.63], ”shirt”: [2163.59, 1079.3, 4254.59, 3826.1]}}

Listing 5.19: Engagement Finetuning Verbalization Pattern (2): Noisy engagement in input and asking the model to correct the noise in addition to producing content

Input: ”You are a smart model. I am giving giving you some data regarding an image released by a content creator − (1) captions (2) keywords (3) image resolution i.e. (width, height) (4) release date (5) approximate number of downloads that the creator wants to achieve (6) approximate number of forwards that the creator wants to achieve (7) approximate number of impressions/views that the creator wants to achieve

You have to predict following attributes of the image: (1) colour and tones from the lists given below: − Allowed colours: ['Red', 'Dark_Red', 'Green', 'Bright_Green', 'Dark_Green', 'Light_Green', ' Mud_Green', 'Blue', 'Dark_Blue', 'Light_Blue', 'Royal_Blue', 'Black', 'White', 'Off_White', 'Gray', ' Dark_Gray', 'Silver', 'Cream', 'Magenta', 'Cyan', 'Yellow', 'Mustard', 'Khaki', 'Brown', 'Dark_Brown', ' Violet', 'Pink', 'Dark_Pink', 'Maroon', 'Tan', 'Purple', 'Lavender', 'Turquoise', 'Plum', 'Gold', 'Emerald', 'Orange', 'Beige', 'Lilac', 'Olive'] − Allowed tones: ['warm', 'neutral', 'cool'] (2) main objects present in the image and the diagonal coordinates of their bounding boxes: [x1, y1, x2, y2] (3) exact number of downloads that the image will get (4) exact number of forwards that the image will get (5) exact number of impressions/views that the image will get

Now, predict the attributes for the following image: [captions: ””Hispanic adult man holding 100 brazilian real banknotes smiling happy pointing with hand and finger to the side””, keywords: ””pointing, side, face, happy, hopeful, smile, finger, optimistic, hand, point, showing, looking, smiling, one, gesture, confident, up, cheerful, look, mouth, joy, friendly, expression, emotion, presentation, idea, blue, background, hispanic, latin, man, male, guy, beard, bald, shaved, adult, young, money, currency, business, brazilian, cash, brazil, real, investment, banknote, 100””, image resolution: ””(9216, 6144)””, release date:

””2021−02−27””, approximate number of downloads that the creator wants to achieve: ””4””, approximate number of forwards that the creator wants to achieve: ””17””, approximate number of impressions/views that the creator wants to achieve: ””919””] Answer properly in JSON format. Do not include any other information in your answer.”

Output: {”color and tones”: {”colors”: {”Cyan”: {”coverage”: 0.69}, ”Light_Blue”: {”coverage”: 0.16}, ” Turquoise”: {”coverage”: 0.15}}, ”tones”: {”warm”: 0, ”neutral”: 0, ”cool”: 1.0}}, ”objects”: {”man”: [1068.6, 18.57, 8143.44, 6121.09], ”banknote bill”: [3443.5, 2146.79, 5294.15, 3455.05]}, ”exact downloads”: 4, ”exact forwards”: 15, ”exact impressions”: 885}

Listing 5.20: Engagement Finetuning Verbalization Pattern (3): Noisy engagement in input and asking the model to correct engagement in output

Input: ”You are a smart model. I am giving giving you some data regarding an image released by a content creator − (1) captions (2) keywords (3) image resolution i.e. (width, height) (4) release date (5) approximate number of downloads that the creator wants to achieve (6) approximate number of forwards that the creator wants to achieve (7) approximate number of impressions/views that the creator wants to achieve

You have to predict following attributes of the image: (1) exact number of downloads that the image will get (2) exact number of forwards that the image will get (3) exact number of impressions/views that the image will get.

Now, predict the attributes for the following image: [captions: ””Movie slapstick vector illustration. Behind the scenes inscription on flapper””, keywords: ””behind the scenes, slapstick, flapper, movie, cinema, scene, logo, frame, film, duration, behind, act, black, cameraman, clip, date, director, entertainment, flap, footage, gray, hollywood, icon, illustration, inscription, operator, screen, shooting, sign, signal, symbol, television, theater, time, timecode, tv, vector, video, view, white””, image resolution: ””(4096, 4096)””, release date: ””2017−06−11””, approximate number of downloads that the creator wants to achieve: ””5””, approximate number of forwards that the creator wants to achieve: ””31””, approximate number of impressions/views that the creator wants to achieve: ””914””] Answer properly in JSON format. Do not include any other information in your answer. ”

Output: {”exact downloads”: 5, ”exact forwards”: 32, ”exact impressions”: 864}

Listing 5.21: Engagement Finetuning Verbalization Pattern (4): No engagement in input and asking the model to simulate exact engagement in output

Input: ”You are a smart model. I am giving giving you some data regarding an image released by a content creator − (1) captions (2) keywords (3) image resolution i.e. (width, height) (4) release date

You have to predict following attributes of the image: (1) exact number of downloads that the image will get (2) exact number of forwards that the image will get (3) exact number of impressions/views that the image will get

Now, predict the attributes for the following image: [captions: ””Company employees working in software development and designer office””, keywords: ””business, office, meeting, collaegue, successful, workplace, analysis, architect, coworker, discussion, entrepreneur, marketing, professional, company, employee, occupation, software, work, worker, team, people, brainstorming, cooperation, corporate, project, strategy, teamwork, together, computer, colleagues, young, diverse, collaboration, design, developer, group, ideas, management, smiling, multiethnic, place, plan, research, startup, technology, women, programmer, architects””, image resolution: ””(4035, 2690)””, release date: ””2020−09−29””]. Answer properly in JSON format. Do not include any other information in your answer.”

Output: {"exact downloads": 1, "exact forwards": 1, "exact impressions": 186}

Hyperparameters for visual instruction finetuning of EngageNet:

- per device train batch size: 16
- gradient accumulation steps: 1
- max context length: 2048
- warmup ratio: 0.03
- warmup learning rate scheduler: cosine
- after warmup, learning rate: 2e-5

5.2.7.7 Repurposing EngageNet for Design Specification Generation (DSG)

5.2.7.7.1 Training For Design Specification Prediction Task Prior works (Bhattacharyya *et al.*, 2023) have demonstrated the capability of language-only pre-trained models like GPT-3 and Vicuna to infer information about visual content without explicit visual reasoning training. Recent models such as BLIP (Li *et al.*, 2023*a*), Llava (Liu *et al.*, 2023*a*), MiniGPT-4 (Zhu *et al.*, 2023), and GPT-4 (OpenAI, 2023) have shown language models' ability to 'see' by incorporating visual branches (often a combination of ViT (Dosovitskiy *et al.*, 2020) and Qformer (Li *et al.*, 2023*a*)) and training them with image-language instructions to answer image-related questions. However, our findings (Table 5.15) reveal that neither pretraining nor further instruction tuning gives a language model the ability to simulate the downstream engagement of an image-based communication or reason about how a more engaging image should look like. Further, we also find that in-context learning, while successful in many other domains, does not perform well in engagement-related tasks. Therefore, to teach a language model about image content and downstream performance, we further train the Llama LLM.

To teach Llama about an image and its downstream engagement, we perform engagement fine-tuning (Khandelwal *et al.*, 2024). We design four types of engagement-finetuning instructions (Listings 5.18-5.21). The idea is to *verbalize* an image using image perception models like color extractor, tones extractor, object, and coordinate detector and convert it to natural language. Then, the image caption, keywords, the required engagement level, date, and marketer information is fed as input to the LLM and asked to output the image verbalization. This

way, the LLM learns to map the image prompt and engagement level to image verbalization.

We train the LLM on the train set of EngagingImageNet data. In Listings 5.18-5.19, we provide the image caption, keywords, date, and required engagement level as inputs to the model. Our aim is to train the model to predict a design specification comprising colors and tones with their spatial coverage as well as objects with their bounding boxes, that should be reflected in the image. Moreover, we observe improved learning in the language model for engagement-conditioned image generation when introducing a 20% noise in the engagement. We then task the model to rectify this noise in the output, simultaneously generating the verbalization of the engagement-conditioned image in Listings 5.20-5.21.

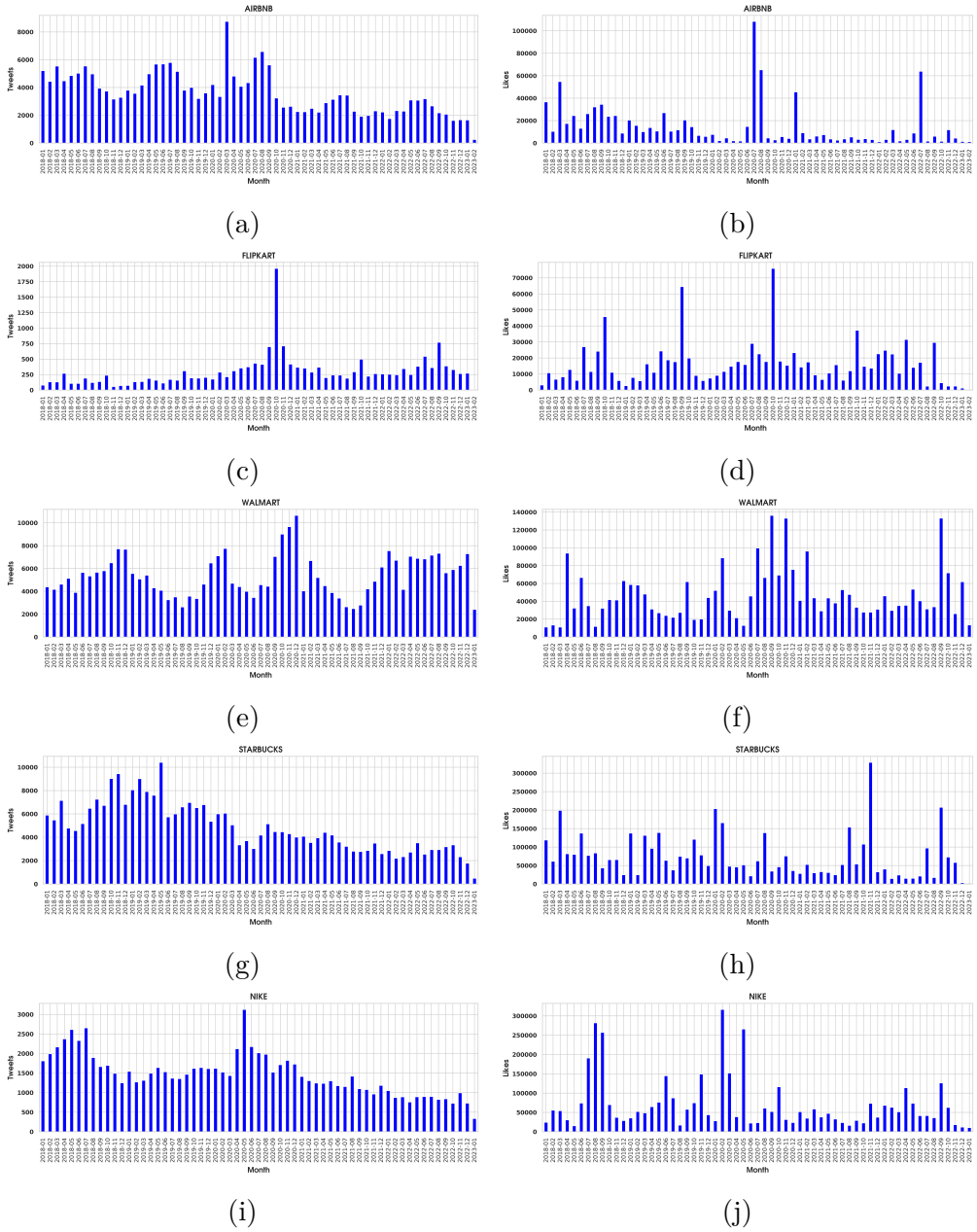


Figure 5.19: Plots showing variation of number of tweets and likes with time for a few companies in the EngagingImageNet dataset

Table 5.15: Performance of all models on the engagement-optimized design specification generation (DSG) task across different engagement-level images in EngagingImageNet data. It is noteworthy that (i) EngageNet outperforms larger sized GPT-3.5, 4 and also the sized Llama model fine-tuned on the same data (without including engagement tokens). (ii) In-context learning does not work well in the engagement-conditioned design specification generation domain.

EngagingImageNet Data											
Model	**Engagement Optimized**	**KPI**	**Colours**				**Tones**	**Objects**			
			IOU ↑	**Cosine Similarity ↑**	**RGB distance ↓**	**Coverage RMSE ↓**	**Coverage RMSE ↓**	**IOU ↑**	**Cosine Similarity ↑**	**Normalised Area RMSE ↓**	**Relative Position Error ↓**
Finetuned Llama	No	High	0.3717	0.8725	0.2855	0.1694	0.1957	0.2547	0.8071	0.2612	0.3078
		Low	0.3362	0.8602	0.2223	0.1811	0.2339	0.2743	0.8047	0.2421	0.2954
Finetuned Llama (EngageNet)	Engagement Finetuning	High	0.4065	0.8898	0.2795	0.1507	0.1718	0.2732	0.8122	0.2509	0.3054
		Low	0.4531	0.8791	0.2084	0.1443	0.1848	0.3455	0.8228	0.2373	0.2889
3-shot GPT-3.5	In-context learning	High	0.214	0.7765	0.2851	0.1773	0.396	0.1085	0.6621	0.3090	0.3651
		Low	0.2175	0.7781	0.2254	0.2118	0.3347	0.1338	0.6749	0.3098	0.3573
5-shot GPT-3.5	In-context learning	High	0.2137	0.7704	0.2743	0.1976	0.324	0.1011	0.6456	0.3160	0.3622
		Low	0.2191	0.7705	0.2176	0.2449	0.3186	0.1264	0.656	0.3150	0.3615
3-shot GPT-4	In-context learning	High	0.2421	0.7887	0.2726	0.192	0.304	0.1035	0.6316	0.3137	0.3666
		Low	0.2405	0.793	0.2332	0.2094	0.3037	0.1419	0.6604	0.3248	0.3763
5-shot GPT-4	In-context learning	High	0.2437	0.7905	0.2702	0.1864	0.2937	0.1008	0.6136	0.3111	0.3782
		Low	0.2448	0.7924	0.2278	0.2144	0.2944	0.1464	0.6406	0.3301	0.3857

5.2.7.7.2 Results for Engagement-conditioned Design Specification Prediction Task

To generate engagement-conditioned image verbalization, we compare several models: in-context trained GPT-3.5 and GPT-4, engagement-finetuned Llama (EngageNet), and Llama fine-tuned on image verbalization but without user engagement information. By comparing against a fine-tuned Llama trained on the same instruction as EngageNet, except with the inclusion of engagement tokens, we aim to isolate the impact of engagement tokens on improving generated engagement-conditioned image verbalizations, independent of the instruction tuning process. We assess all models across multiple metrics that evaluate the extent to which the generated verbalizations align with ground truth in terms of colors, tones, objects, and their positions. Intersection over Union (IoU) metrics gauge the overlap between ground truth and generated constructs (colors and objects), while similarity metrics measure cosine similarity between ground truth and generated constructs (colors, objects). Coverage errors determine the how closely the proportion of ground truth and predicted constructs (colors, tones) in the image match. Additionally, we calculate differences in predicted and ground truth areas and locations for objects, accounting for semantically similar objects (such as sofa and couch). Further details on these metrics and their formulas can be found in Appendix 5.2.7.7.3.

Table 5.15 displays the outcomes. The results indicate that engagement fine-tuning enables EngageNet to achieve superior performance across all metrics, surpassing both equivalently sized fine-tuned Llama and 10x larger instruction-tuned GPT-3.5 and GPT-4. Furthermore, in-context learning demonstrates subpar performance, with both the three and five-shot models displaying similar results.

5.2.7.7.3 Evaluation Metrics for Design Specification Prediction

- **Colours IOU**: The intersection over union between set C^G of colours in the ground truth image verbalization and set C^P of colours in the predicted image verbalization is computed as:

$$IOU(C^G, C^P) = \frac{|C^G \cap C^P|}{|C^G \cup C^P|} \tag{5.5}$$

- **Colours similarity**: For the ground truth colour set $C^G = c_1^G, c_2^G, ..., c_i^G$ and predicted colour set $C^P = \{c_1^P, c_2^P, ..., c_j^P\}$, we correspondingly obtain the sets of word vectors $W^G = \{w_1^G, w_2^G, ..., w_i^G\}$ and $W^P = \{w_1^P, w_2^P, ..., w_j^P\}$, using Spacy [§§]. For some similarity threshold τ, the mean cosine similarity

[§§] https://spacy.io/

is computed as follows:

$$\frac{\sum_{i=1}^{|C^G|}\sum_{j=1}^{|C^P|} cos(w_i^G, w_j^P).I(w_i^G, w_j^P, \tau)}{\sum_{i=1}^{|C^G|}\sum_{j=1}^{|C^P|} I(w_i^G, w_j^P, \tau)} \tag{5.6}$$

where $I(w_i^G, w_j^P, \tau)$ is an indicator function defined as:

$$I(w_i^G, w_j^P, \tau) = \begin{cases} 1 & \text{if } cos(w_i^G, w_j^P) > \tau \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

We take $\tau = 0.7$ in our experiments.

- **Colours RGB distance**: Given the ground truth colour set $C^G = c_1^G, c_2^G, ..., c_i^G$ and predicted colour set $C^P = \{c_1^P, c_2^P, ..., c_j^P\}$, we map each colour to its RGB value to obtain the sets $W^G = \{w_1^G, w_2^G, ..., w_i^G\}$ and $W^P = \{w_1^P, w_2^P, ..., w_j^P\}$ where each element in the sets is a 3×1 dimensional vector of RGB values. For some distance threshold τ, the mean euclidean distance is calculated as follows:

$$\frac{\sum_{i=1}^{|C^G|}\sum_{j=1}^{|C^P|} distance(w_i^G, w_j^P).\mathbb{I}(w_i^G, w_j^P, \tau)}{\sum_{i=1}^{|C^G|}\sum_{j=1}^{|C^P|} \mathbb{I}(w_i^G, w_j^P, \tau)} \tag{5.8}$$

where $\mathbb{I}(w_i^G, w_j^P, \tau)$ is an indicator function defined as:

$$\mathbb{I}(w_i^G, w_j^P, \tau) = \begin{cases} 1 & \text{if } distance(w_i^G, w_j^P) < \tau \\ 0 & \text{otherwise} \end{cases} \tag{5.9}$$

We take $\tau = 0.5$ in our experiments.

- **Colours coverage RMSE**: Consider the intersection $I = C^G \cap C^P$ of ground truth and predicted colour sets. The root mean squared error between the area covered by colours present in both ground truth and predicted image is calculated as follows:

$$RMSE = \sqrt{\frac{1}{|I|}\sum_{i=1}^{|I|}(coverage(c_i^G) - coverage(c_i^P))^2} \tag{5.10}$$

- **Tones coverage RMSE**: Consider the intersection $I = T^G \cap T^P$ of ground truth and predicted image tones. The root mean squared error between the proportion of tones in ground truth and predicted image is calculated as follows:

$$RMSE = \sqrt{\frac{1}{|I|}\sum_{i=1}^{|I|}(coverage(t_i^G) - coverage(t_i^P))^2} \tag{5.11}$$

- **Objects IOU**: The intersection over union between set O^G of objects in the ground truth image verbalization and set O^P of objects in the predicted image verbalization is computed as:

$$IOU(O^G, O^P) = \frac{|O^G \cap O^P|}{|O^G \cup O^P|} \tag{5.12}$$

- **Objects similarity**: For the ground truth set of objects $O^G = \{o_1^G, o_2^G, ..., o_i^G\}$

and set of predicted objects $O^P = \{o_1^P, o_2^P, ..., o_j^P\}$, we correspondingly obtain the sets of word embeddings $W^G = \{w_1^G, w_2^G, ..., w_i^G\}$ and $W^P = \{w_1^P, w_2^P, ..., w_j^P\}$, using Spacy. For some similarity threshold τ, the mean cosine similarity is computed as follows:

$$\frac{\sum_{i=1}^{|O^G|}\sum_{j=1}^{|O^P|} cos(w_i^G, w_j^P).\mathbb{I}(w_i^G, w_j^P, \tau)}{\sum_{i=1}^{|O^G|}\sum_{j=1}^{|O^P|}\mathbb{I}(w_i^G, w_j^P, \tau)} \tag{5.13}$$

where $\mathbb{I}(w_i^G, w_j^P, \tau)$ is an indicator function defined as:

$$\mathbb{I}(w_i^G, w_j^P, \tau) = \begin{cases} 1 & \text{if } cos(w_i^G, w_j^P) > \tau \\ 0 & \text{otherwise} \end{cases} \tag{5.14}$$

We take $\tau = 0.7$ in our experiments.

- **Normalised objects area RMSE**: As described above, consider the sets of word vectors of objects present in the ground truth image $O^G = \{o_1^G, o_2^G, ..., o_i^G\}$ and predicted image $O^P = \{o_1^P, o_2^P, ..., o_j^P\}$. Given the ground truth image area $A^G = width \times height$ and a similarity threshold τ, we first compute the mean squared error between the areas of bounding boxes of similar objects in the ground truth and predicted image, weighted by the proportion of each object in the ground truth image and its cosine similarity with the object in the predicted image. Further, we take the square root of the error thus obtained and normalise it by A^G to achieve the desired metric, as follows:

$$MSE = \frac{\sum_{i=1}^{|O^G|}\sum_{j=1}^{|O^P|}\{(area(o_i^G) - area(o_j^P))^2.\frac{area(o_i^G)}{A^G}.\frac{1}{cos(w_i^G, w_j^P)}\}.\mathbb{I}(w_i^G, w_j^P, \tau)}{\sum_{i=1}^{|O^G|}\sum_{j=1}^{|O^P|}\mathbb{I}(w_i^G, w_j^P, \tau)} \tag{5.15}$$

$$Normalised\ RMSE = \frac{\sqrt{MSE}}{A^G} \tag{5.16}$$

where $\mathbb{I}(w_i^G, w_j^P, \tau)$ is an indicator function as described above. We take $\tau = 0.7$ in our experiments.

- **Normalised relative position error**: Following a similar approach as explained above, we compute the mean euclidean distance between the centroids of bounding boxes of similar objects weighted by the cosine similarity of objects present in the ground truth and predicted images and normalise it by the length of diagonal in the ground truth image D^G:

$$RPE = \frac{\sum_{i=1}^{|O^G|}\sum_{j=1}^{|O^P|}\{(distance(centroid(o_i^G), centroid(o_j^P).\frac{1}{cos(w_i^G, w_j^P)}\}.\mathbb{I}(w_i^G, w_j^P, \tau)}{\sum_{i=1}^{|O^G|}\sum_{j=1}^{|O^P|}\mathbb{I}(w_i^G, w_j^P, \tau)} \tag{5.17}$$

$$Normalised\ RPE = \frac{RPE}{D^G} \tag{5.18}$$

where $\mathbb{I}(w_i^G, w_j^P, \tau)$ is the aforementioned indicator function. As before, we take $\tau = 0.7$ in our experiments.

5.2.7.8 Performance Alignment of Stable Diffusion using Design Specification Generation (DSG) Reward

5.2.7.8.1 DDPO Additional Details The denoising process in diffusion models is a multi-step recursive process with a pre-specified finite number of steps. In DDPO Black *et al.* (2023), this denoising process is viewed as a finite horizon Markov decision process (MDP), where the state comprises of the current context, number of steps left in the process and the current denoised image. The action to be taken is to predict the next image using this state.

The image forming the initial state is sampled from a standard normal distribution. Mathematically, a finite horizon MDP is defined as a tuple $\{T, \mathcal{S}, \mathcal{A}, P, R\}$, where these components are defined as:

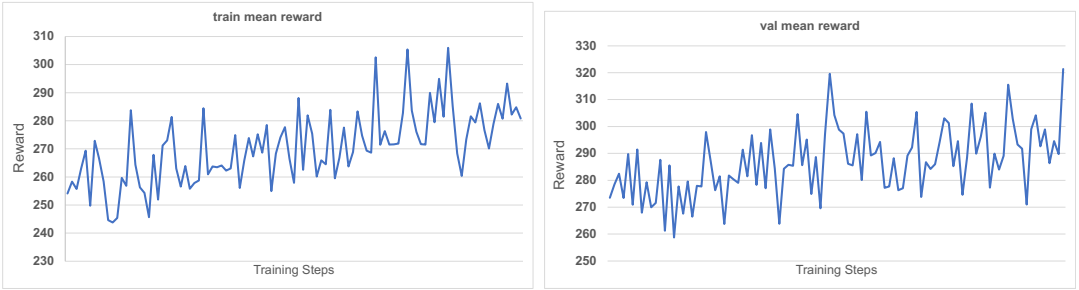
1. T is the horizon or the number of steps in the MDP
2. $\mathcal{S}$ is the state space. Here it comprises of three components, the context c, the current number of steps left in the denoising process, t, and the current denoised image representation (a given vector encoding of the image), x_t. The initial or starting state has the context c_0 given as input, the number of steps left at the beginning, $t_0 = T$ and the initial image representation is sampled from a normal distribution of appropriate dimension, $x_0 \sim \mathcal{N}(0, I)$.
3. $\mathcal{A}$ is the action space, and here it is the space comprising of all image representations x. If x is a $d-$dimensional vector, then $\mathcal{A} = \mathbb{R}^d$.
4. $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function. Here, we specify P separately for each of the three components of the state as $P_c = \delta(c_t)$, $P_t = \delta(t-1)$, and $P_x = \delta(a_t)$, where the current state is c_t, t, x_t, current action $a_t = x_{t-1}$, and $\delta(\cdot)$ is the Dirac delta distribution.
5. $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function that takes a state and action as input and returns a scalar reward. We generate this scalar reward signal using EngageNet.

DDPO is based on conventional policy gradient algorithms in RL, which assumes that the environment, *i.e.*, the transition and the reward functions are neither differentiable nor accessible to the RL agent. However, when one has access to both the transition and reward functions, and when both these are differentiable, one could use analytic policy gradient based algorithms for training the RL agent (Wiedemann *et al.*, 2023). In our case, we do have access to the transition and reward functions. As in standard diffusion models, the transition

function is the single-step denoising function. However, we do not have a differentiable reward function as we are using non-differentiable featurizers in Step 1 of the reward generation process as described above. An alternative approach would be to make this step differentiable and then use the end-to-end analytic policy gradient approach for aligning the stable diffusion model. In order to simplify our training pipeline and to avoid getting into potential stability issues when performing end-to-end learning, we chose the conventional RL approach of DDPO for this work.

5.2.7.8.2 Design Specification Generation (DSG) reward The steps of constructing the reward function based on design specification generation are given below 5.26:

- We featurize the image generated by stable diffusion to obtain features (including colors, tones, objects, and their positions) that EngageNet is meant to predict as part of a design specification conditioned on contextual information such as marketer, expected likes, tweet content, image caption, *etc.*
- Based on the above conditioning factors, we use EngageNet to predict the logits of the verbalized features of the image generated by stable diffusion as described in the previous step.
- We now have one logit per text token as EngageNet's output. To convert this to a scalar score, we compute the probabilities of each token and then add them.



(a) (b)

Figure 5.20: Reward curves for the performance alignment of stable diffusion on EngagingImageNet (train (a) and validation (b) sets)

5.2.7.9 Broader Impacts and Limitations

Our work on assessing and improving the engagement of text-to-image generation models introduces several societal considerations that require careful examination.

Table 5.16: Results for the performance of various models on the EngagingImageNet for the engagement-optimized image generation task. Results are computed on the EngageNet Design Specification Generation (DSG) reward (§5.2.7.8) as well as other metrics reported in the literature.

Images	KPI	Reward ↑	Other Metrics		
			FID ↓	Aesthetic Score ↑	CLIP Score ↑
Base Stable Diffusion	High	242.545	34.958	5.221	33.346
	Low	238.471	42.999	4.925	31.705
High-KPI fine-tuned Stable Diffusion	High	239.023	26.023	4.850	32.210
	Low	223.619	37.497	4.433	30.979
EngageNet aligned Stable Diffusion (EOIG-SD)	High	254.918	36.546	5.341	33.379
	Low	247.597	49.492	5.087	31.719

We aim to provide a comprehensive analysis of the potential impacts, highlighting both contributions to the field and the precautions necessary for responsible development and deployment of engagement-optimizing technologies.

1. The ability of models to enhance user engagement raises important societal concerns around responsible deployment and potential misuse. Quantifying these risks is essential for developing appropriate safeguards. However, studying user engagement, particularly in uncontrolled environments, poses ethical challenges. For instance, investigating engagement through manipulated or highly optimized content could influence user behavior in unintended or harmful ways. To mitigate this risk, we have limited our research to controlled environments and theoretical analyses, ensuring that insights into engagement are gathered without causing real-world harm.

2. To foster responsible use of our research and datasets, we will release an Acceptable Use Policy explicitly prohibiting the misuse of our dataset for generating content aimed at harmful or deceptive purposes. This includes banning its use in abusive contexts (e.g., creating deceptive ads or manipulative imagery) and sensitive applications such as political propaganda. We will actively monitor compliance with this policy and encourage others in the research community to adopt similar ethical guidelines when using engagement-optimizing models. Importantly, our dataset is PII-free, ensuring that no personal information of individuals is included. This dataset was collected in accordance with strict ethical and legal guidelines, ensuring compliance with relevant data protection policies.

3. We plan to release the dataset and evaluation frameworks in stages, starting with the release of our benchmark and engagement arena. This staged release will help familiarize the research community with the methodologies we use to assess engagement in generated content. By gradually releasing the dataset (in batches of 20%), we will closely monitor how models perform in enhancing engagement. Initially, the dataset will only be available in a controlled environment, enabling us to manage usage and address emerging concerns. We also encourage the research community to contribute additional data to expand our evaluation framework. This approach balances

the need for research progress with ethical responsibility and community involvement.

4. We recognize the dual-use potential of models designed to optimize engagement. While this technology can be beneficial in fields like education or user-centered design, it also poses risks of misuse in deceptive contexts. Drawing parallels to ethical discussions on persuasive technologies, we believe that transparency and safeguards in dataset design can mitigate the potential for harm. The insights gained from understanding user engagement can aid in the responsible development of future AI systems.

5. PII Removal and Data Collection: To protect user privacy, we have implemented measures to remove all personally identifiable information (PII). Our dataset is compiled without collecting sensitive personal data, focusing solely on public, non-individualized information. All references to specific users or personal identifiers have been removed. Additionally, we collect only aggregate metrics (e.g., overall user interaction data) to measure engagement trends without compromising individual privacy.

6. In this work, we specifically focus on the engagement optimization capabilities of text-to-image generation models. We introduce benchmarks and evaluation methodologies for measuring user engagement with AI-generated images and develop techniques to enhance this engagement. Our findings suggest that engagement with generated content can be improved not just by increasing model size but also through targeted training strategies. Furthermore, engagement patterns observed in one domain (e.g., social media) often transfer to other domains (e.g., marketing or websites), which broadens the applicability of our findings.

5.2.7.9.1 Limitations In this paper, we examine a single aspect of engagement. In real-world applications, user engagement often occurs in sequential or multi-stage interactions, which we plan to address in future research. Additionally, this work is focused on English-language data; we aim to extend our findings to other languages in subsequent studies. Furthermore, the impact of audience dependence on engagement has not been studied extensively in this paper, partly due to the absence of publicly available datasets. We plan to work on collecting such datasets to explore this effect in future work. These limitations underscore areas for further research and caution against over-generalizing our findings to more complex real-world scenarios.

5.2.7.10 Additional Figures



Figure 5.21: Any message is created to serve an end goal. For marketers, the end goal is to bring in the desired receiver effect (behavior) (like clicks, purchases, likes, and customer retention). The figure presents the key elements in the communication pipeline - the marketer, message, channel, receivers, and finally, the receiver effect. Traditionally, image generation is optimized on metrics such as aesthetics and FID. For effective communication, the image generation process needs to be optimized on the receiver effect (other than the traditional metrics).

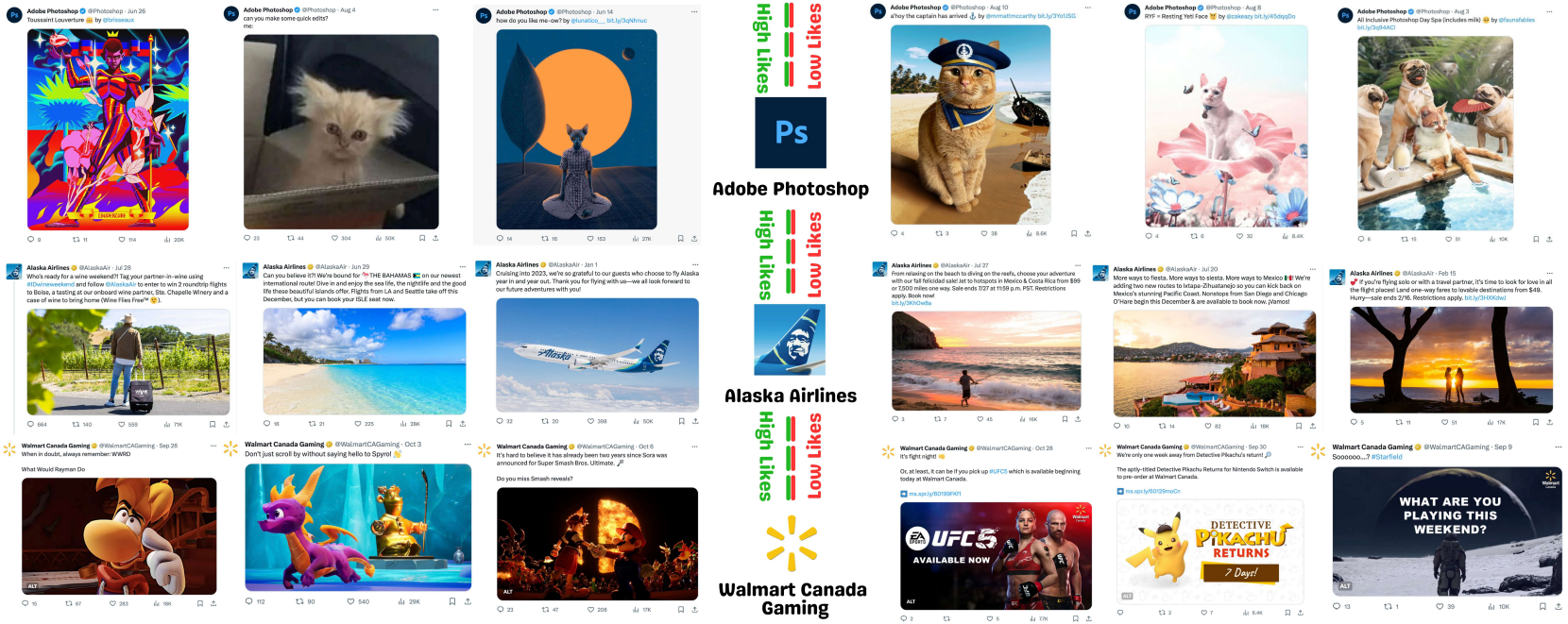


Figure 5.22: Sample media and tweets from enterprise accounts in the EngagingImageNet dataset. It can be noted, for example, in the Adobe Photoshop tweets, that the media does not differ significantly in aesthetics or objects themselves (all of them are cats). Despite that, there is much difference in the image KPIs, indicating that viewer engagement is distinct from other optimization objectives such as aesthetics or prompt adherence.

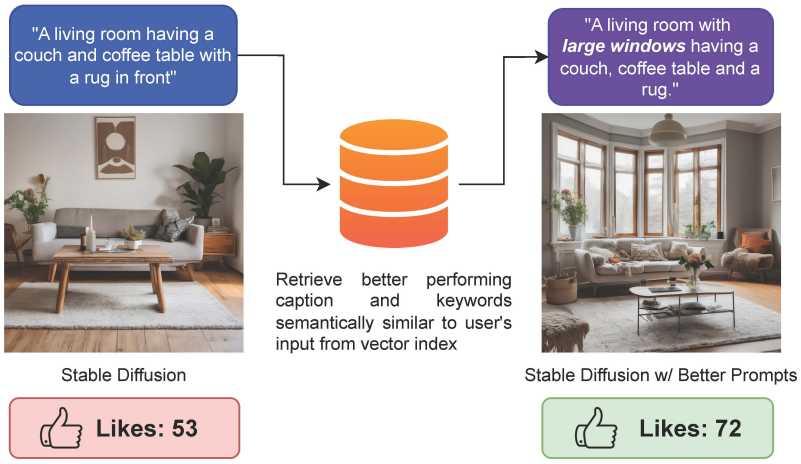


Figure 5.23: Retrieval framework for conditioning text-to-image models on higher engagement prompts as described in Section 5.2.4.1. The retrieved prompts may incorporate image characteristics that have been empirically shown to improve image engagement.

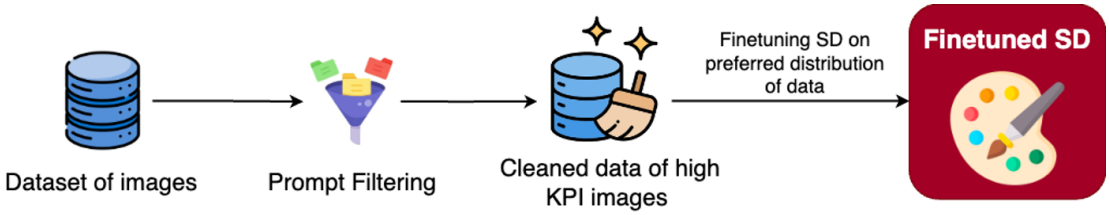


Figure 5.24: Illustration depicting supervised finetuning of stable diffusion model on high-liked images from EngagingImageNet dataset as described in Section 5.2.4.2. This method of finetuning U-Net module on preferred data distribution results in generating more engaging images.

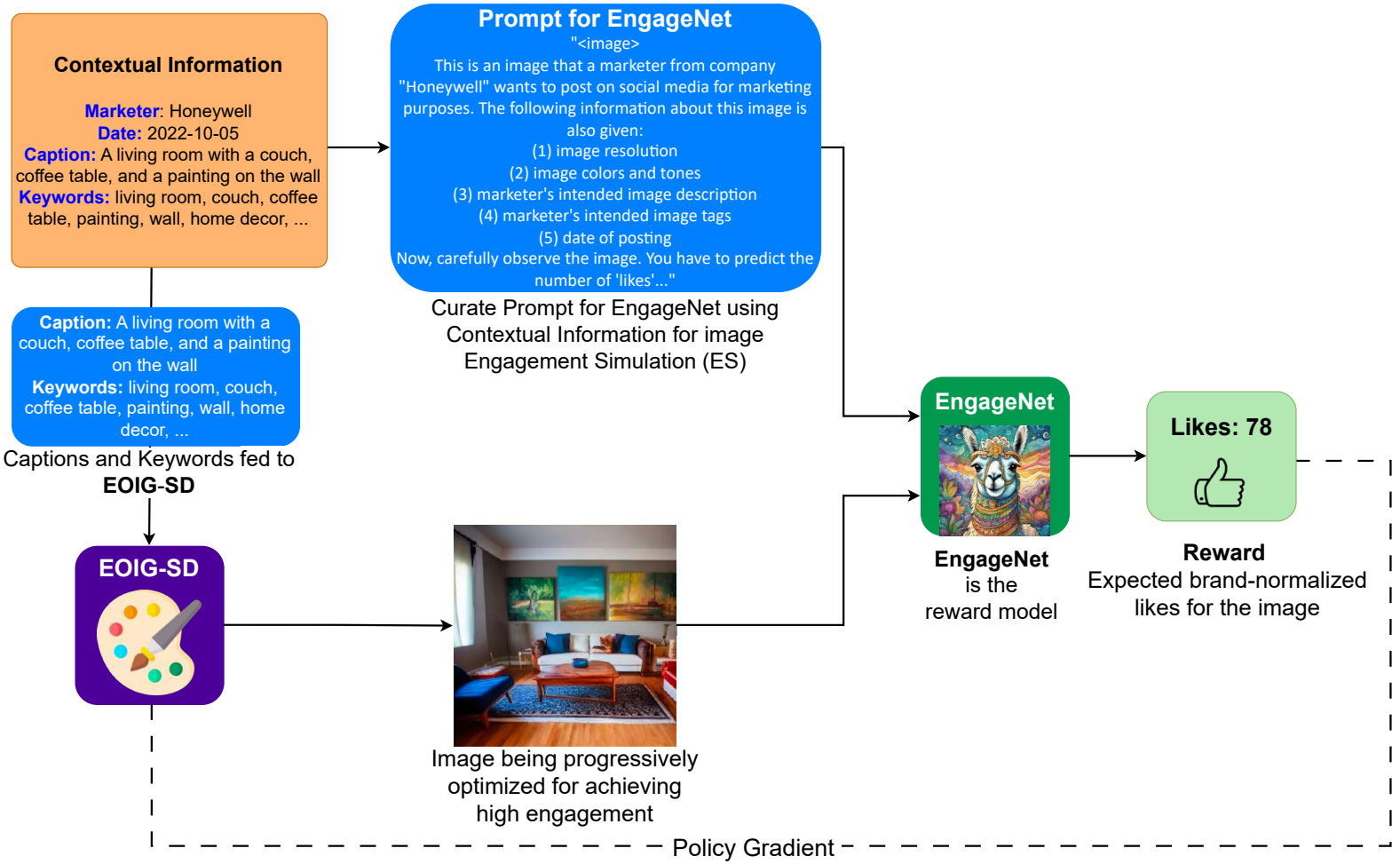


Figure 5.25: Aligning Stable Diffusion for higher engagement using DDPO algorithm (Black *et al.*, 2023) using Engagement Simulation (ES) as the reward function. Architecture of the proposed pipeline for training stable diffusion for the objective of engagement-optimised image generation (EOIG) using Engagement Simulation (ES) reward function as described in Section 5.2.4.3. EngageNet predicts the engagement level of images generated by stable diffusion. The scalar rewards are used to guide stable diffusion to produce progressively higher engagement images. The resulting diffusion model is called EOIG-SD (RLHF-ES).

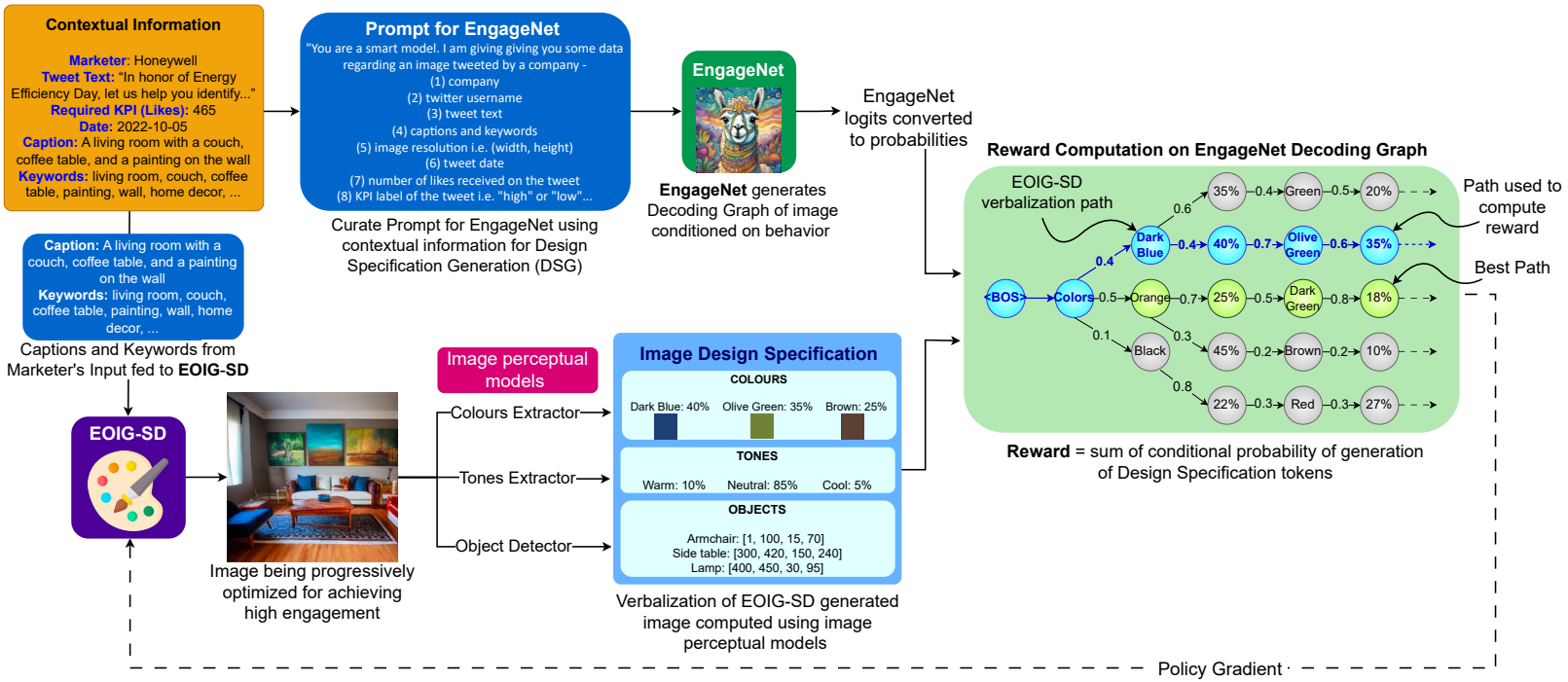


Figure 5.26: Aligning Stable Diffusion for higher engagement using DDPO algorithm (Black *et al.*, 2023) using Design Specification Generation (DSG) as the reward function. The architecture of the proposed pipeline for training stable diffusion for the objective of engagement-optimized image generation (EOIG) using Design Specification Generation (DSG) reward function as described in Section 5.2.4.3 and 5.2.7.7. EngageNet trained in this manner possesses the capability to generate verbal descriptions comprising colors, tones, objects, and their locations of an image based on conditioning factors such as the company, time, image caption and viewer likes. We leverage this EngageNet as a reward model to train stable diffusion such that the images generated by it have a design specification aligned with those of higher engagement image. EOIG-SD takes a prompt and generates an image, which then undergoes verbalization via image perception models. Its objective is to create images that, when verbalized, closely resemble the engagement-conditioned verbalization generated by EngageNet. The verbalized output of EOIG-SD is fed into the reward model. We ask EngageNet to predict the logits for this image verbalization, using which a reward is computed for EOIG-SD, indicating how closely this verbalized output aligns with EngageNet. This reward value serves as feedback for EOIG-SD in the form of policy gradient, aiding in its continual improvement and refinement within the image generation process. Thus, this pipeline trains EOIG-SD to generate engagement-optimized images by gradually aligning its output with EngageNet.

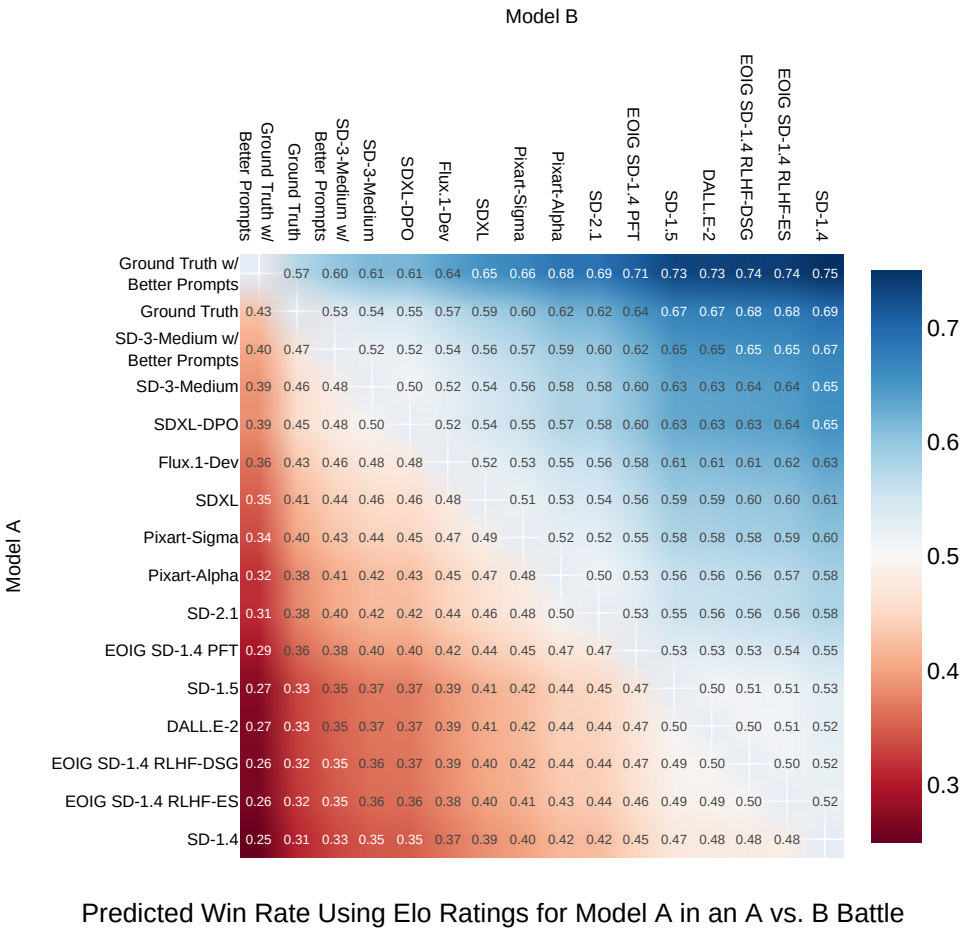


Figure 5.27: Win Rates of different models against each other in the Image Engagement Arena

Chapter 6

Conclusion and an Outlook for Future Work

This thesis has explored the intersection of communication theory, behavioral science, and artificial intelligence, with a particular focus on explaining, understanding, and optimizing human behavior through large-scale modeling approaches. Our work builds upon the fundamental seven-factor model of communication—communicator, message, channel, time of receipt, receiver, time of behavior, and receiver's behavior—while leveraging unprecedented access to digital behavioral data to advance both explanatory and predictive approaches to behavioral science. In the domain of persuasion strategy analysis, we have made significant contributions to understanding the mechanisms of influence in advertising. Through comprehensive research spanning marketing, social psychology, and machine learning literature, we developed the most extensive framework of generic persuasion strategies to date. This work was supported by the creation and release of the first datasets for studying persuasion strategies in both image and video advertisements.

We discover that existing Large Language Models (LLMs), despite their remarkable capabilities in various domains, are inherently limited in modeling behavior due to the systematic removal of behavioral data during training. To address this limitation, we developed the Large Content and Behavior Models (LCBM), which integrates all seven factors of communication to create more comprehensive models of human behavior. To support future research in this area, we released extensive behavior instruction fine-tuning data derived from over 40,000 YouTube videos and 168 million Twitter posts. Additionally, we established new benchmarks for evaluating joint content-behavior understanding, encompassing both predictive and descriptive tasks.

We also made significant strides in demonstrating how behavioral signals can enhance content understanding. Our research showed substantial improvements across 46 different tasks spanning 23 benchmark datasets across language, au-

dio, text, and video modalities. We proposed a scalable approach to enhance Vision Language Models (VLMs) without requiring significant architectural changes, making our improvements readily accessible to the broader research community. These results strongly validate our hypothesis that behavioral responses provide valuable signals for content understanding, opening new avenues for improving AI systems' comprehension capabilities.

In the realm of content generation, we made contributions towards generating performant content in both text and visual domains. Through our work on memorability optimization, we developed Henry, a model achieving a 44% improvement in memorability scores of the generated content from the starting point. This represents the first successful application of synthetic data in a domain previously lacking large-scale training resources. In the visual domain, we addressed the critical need for engagement-optimized image generation through the development of EngageNet and the creation of EngagingImageNet, a comprehensive dataset of 168 million tweets with associated media and engagement metrics. Our introduction of Engagement Arena, the first automated benchmark for assessing the engagement potential of text-to-image models, provides the research community with a valuable tool for evaluating and improving engagement-oriented image generation techniques.

6.1 Limitations and Future Work

Despite these significant contributions, our research has several important limitations that must be acknowledged. First, our models exhibit fundamental scale-dependent performance limitations, where behavioral specialization cannot overcome the advantages of larger parameter counts in pure content understanding tasks, as evidenced by LCBM's consistent underperformance compared to significantly larger models like GPT-4V. Second, we observe domain specificity constraints in behavioral transfer, with models trained on platform-specific data (e.g., YouTube) showing substantial performance degradation when applied to different domains (e.g., email marketing), suggesting that behavioral patterns are more context-dependent than initially hypothesized. Third, our approaches face scala-

bility challenges with longer and more complex content, particularly evident in the verbalization approach used in LCBM, where performance degrades significantly for content exceeding certain thresholds (videos over 300 seconds or emails with extensive visual content). Fourth, the quality and availability of behavioral signals create dependencies that limit model applicability—performance drops substantially for content with sparse engagement data or when behavioral metrics are influenced by external events and temporal variance. Fifth, our evaluation frameworks, particularly for memorability prediction, are constrained by relatively low human consistency scores (ranging from 0.55 to 0.78), indicating fundamental challenges in behavioral assessment that set theoretical performance ceilings. Finally, our datasets and models exhibit inherent biases stemming from platform demographics, self-selection effects, and algorithmic influences, which may limit generalizability across diverse populations and cultural contexts. These limitations underscore the need for continued research in developing more robust, scalable, and generalizable approaches to behavioral modeling.

Looking ahead, this research opens several promising directions for future work. The integration of behavioral data into AI systems could lead to more nuanced and context-aware models that better understand and predict human responses. Concretely, we visualize the following avenues for automated behavioral sciences in the near future:

1. **Infinite Personalization**: Before the invention of the printing press, each document had to be written with manual effort. Content production was the limiting factor in communication. The invention of the printing press made it possible to mass-produce content. However, delivery was still limited. While newspapers began to be printed, their area of influence was limited to a certain small geographical boundary. Delivery was the limiting factor then. Steam engines helped solve some of that problem. Still, the extent of delivery was limited, and the speed of delivery was slow. It was not until the invention of the internet and mobile devices that the delivery problem was completely solved. Now, anyone can instantly deliver any piece of content to any other person. The next limiting factor in communication is the time and human labor cost of producing content. This limits a communicator to send out the same message to all the receivers. Further, as both ours and several other research studies have shown, humans are bad at predicting the behavior of others; we need techniques to produce performant content. This will enable infinite personalization, a personalized way of communicating between a communicator and a receiver, with the aim of fulfilling the shared goals.

2. **Simulating Digital Humans and Digital Societies**: At the heart of social simulation lie two perspectives (Gilbert and Troitzsch, 2005): 1) the dynamic feedback or interaction among individuals, and 2) the states of the population, either as a collective whole or as distinct groups. By simulating social activities, researchers and practitioners can predict the future evolution of individuals and groups. In addition, they facilitate experimental environments through interventions. Social simulation can be implemented in two forms: digital humans (Park *et al.*, 2023; Chopard and Droz, 1998; Argyle *et al.*, 2023*a*) and digital societies (Khandelwal *et al.*, 2024; Bhattacharyya *et al.*, 2024; SI *et al.*, 2025; Khurana *et al.*, 2024; Santurkar *et al.*, 2023). In digital human simulation, either human-crafted rules or parameterized models are used to depict the behavior of individuals (referred to as agents) who interact with others, in societal simulation, equations or models are used to model the society as a whole including the societal non-linear interactions.

   The emergence of Large Language Models as behavioral proxies represents a paradigm shift in social science methodology. Recent breakthroughs have demonstrated that LLMs can be conditioned to embody specific demographic and psychological profiles, creating "silicon samples" that mirror human populations (Argyle *et al.*, 2023*b*; Park *et al.*, 2024). This capability extends beyond simple text generation to encompass the replication of complex human attitudes, decision-making patterns, and social behaviors with remarkable fidelity.

   The implications of this technological advancement are profound. LLMs can now be adapted to reflect the worldviews and opinions of populations shaped by specific information environments (Chu *et al.*, 2023), enabling researchers to explore how media consumption patterns influence collective beliefs and behaviors. The integration of persona variables—demographic, social, and behavioral characteristics—into language models (Hu and Collier, 2024) has opened new avenues for understanding individual differences in perception and judgment, particularly in contexts where human opinions naturally diverge.

   These developments converge toward a future where digital twins of human populations could serve as testing grounds for social interventions, policy proposals, and communication strategies. The ability to simulate complex social phenomena such as information cascades, opinion polarization, and collective decision-making (Wang *et al.*, 2025) promises to revolutionize how we study and predict human behavior at scale. Yet this power comes with the responsibility to ensure these simulations authentically represent the diversity of human experience, as early investigations reveal significant gaps between model outputs and the true breadth of human opinion (Santurkar *et al.*, 2023; Rescala *et al.*, 2024).

   The key to building these simulation models lies in leveraging the vast digital footprint left by these observable factors. Both physical and digital interactions contain these signals. For instance, consider a physical political banner displayed by the political campaign of Kamala Harris saying "For The People" in a busy city such as San Francisco and viewed by office-goers, receiving various reactions such as hopeful comments, visible disdain, or cold indifference. Analogously in the digital domain, a tweet by a figure like

Donald Trump saying "Make America Great Again" receives likes, retweets, and comments, whether positive or negative. However, digital signals are far more accessible and recorded in structured datasets, making them ideal for training a Foundation Model. Digital Analytics have been recording such digital signals for decades. Digital analytics involves collecting, analyzing, and interpreting data from digital platforms to capture user behavior. This data typically includes messages sent by a marketer in the form of websites, apps, or digital products and records actions such as clicks, page views, session durations, and navigation patterns, which provide insights into user behavior over a period of time. We have made some initial strides towards achieving this in our recent work (Bhattacharyya *et al.*, 2024).

3. **Measuring persuasiveness and engagement potential of automated agents**: Large Language Models (LLMs) have demonstrated proficiency in content generation and, more recently, in human persuasion through the production of persuasive content (Durmus *et al.*, 2024; Singh *et al.*, 2024*b*). The development of such systems that are capable of generating verifiably persuasive messages presents both opportunities and challenges for society. On one hand, such systems could positively impact domains like advertising and social good, such as addressing vaccine hesitancy (Sekar, 2021; Moore, Thomas, 2021). Conversely, these systems could have detrimental effects if used to influence political inclinations (Tappin *et al.*, 2023), propagate misinformation (Lukito, 2020), or manipulate consumer choices (Boerman *et al.*, 2017).

   The sophistication of LLMs in recognizing and generating persuasive content has reached remarkable levels, with models now capable of distinguishing argument quality and predicting individual responses with human-level accuracy (Rescala *et al.*, 2024). This capability raises fundamental questions about the nature of persuasion itself and challenges our understanding of what makes arguments compelling across different populations. The emergence of ensemble approaches that can surpass individual human performance in persuasion detection suggests we are approaching a threshold where artificial systems may possess more nuanced understanding of persuasive mechanisms than any single human observer. This convergence of artificial and human judgment capabilities (Santurkar *et al.*, 2023) underscores the critical importance of developing evaluation frameworks that capture the full spectrum of human diversity in persuasive response.

   Given these potential societal impacts, it is crucial to develop rigorous methods for studying, measuring, benchmarking, and monitoring the persuasive capabilities of AI models. We have made some initial strides towards achieving this in our recent works (Singh *et al.*, 2024*b*; Khurana *et al.*, 2024).

4. **Automatically explaining human behavior**: While the behavioral science communities are divided into prediction and explanation, and the communities are growing farther apart, the fundamental curiosity of humans is to learn more about themselves and their environment and how it operates. While predictions may be increasingly more and more accurate, if the mechanism is not well understood, the fundamental human curiosity is not satisfied. As a community, our ongoing commitment is to uncover the mechanisms underlying human behavior. However, we have to discover

methods that carry both higher predictive power and are scalable. This may be solved in the future by using advanced tools such as simulations and data from natural experiments to bridge the gap between prediction and explanation.

5. **Rethinking free will in the age of behavioral modeling**: A central and unresolved question in behavioral science is how we should define and understand free will in the context of human decision-making. One practical definition is that free will is the lack of predictability in human actions—if a model is able to predict an individual's next action to a high degree of accuracy, then the existence of true free will comes into question. Traditional views often assume that individuals possess agency and autonomy in their choices. However, the development of models such as LCBM (Khandelwal *et al.*, 2024) and advances in persuasion modeling (Singh *et al.*, 2024*b*; Khurana *et al.*, 2024) challenge this notion by demonstrating that human behavior can be systematically predicted, influenced, and even optimized using large-scale data and machine learning. This perspective aligns with a growing body of evidence from neuroscience and psychology, which suggests that many decisions are shaped by unconscious processes, prior experiences, and environmental cues—sometimes before conscious awareness is even possible (see, e.g., Libet's experiments (Libet, 1985; Libet *et al.*, 1993), studies on priming (Bargh *et al.*, 1996), and behavioral economics research on automaticity (Tversky and Kahneman, 1985; Ariely *et al.*, 2003; Johnson and Goldstein, 2003)). Such findings lend support to anti-free will arguments, raising profound questions about the true nature of autonomy.

   Our work in this thesis explores these issues from a modeling and machine learning perspective, showing how behavioral data and computational models can capture and even influence human actions at scale. Ultimately, the next frontier for science will be to rigorously investigate the degree to which human actions are governed by free will versus being determined by past experiences, environmental context, and external influences. Addressing this will require interdisciplinary collaboration, combining empirical research, philosophical inquiry, and advances in AI to better understand the boundaries and mechanisms of human agency.

6.2 Societal and Ethical Implications of Behavior Optimized AI

The development of AI systems capable of understanding, predicting, and optimizing human behavior represents a significant technological advancement with far-reaching societal implications. Our work throughout this thesis—spanning persuasion strategy analysis in advertising, behavioral signal integration for content understanding, content generation optimized for memorability and engagement,

and comprehensive behavioral modeling—demonstrates both the tremendous potential and the inherent risks of behavior-optimized AI. This section synthesizes the ethical considerations that emerge from our research and similar work in the field, providing a comprehensive framework for understanding the broader implications of this technology.

6.2.1 Dual-Use Nature and Manipulation Risks

Behavior-optimized AI systems exhibit a fundamental dual-use characteristic: the same technologies that can benefit society can also be misused for harmful purposes (Singh *et al.*, 2024*b*; Khurana *et al.*, 2024). Our research demonstrates that AI models can be trained to generate content optimized for specific behavioral outcomes—whether increasing memorability, engagement, or persuasive impact. While such capabilities offer tremendous potential for positive applications in education, public health campaigns, and user-centered design, they simultaneously create unprecedented opportunities for manipulation and deception.

The manipulation potential of these systems manifests in several critical domains. **Political manipulation** represents perhaps the most concerning risk, where behavior-optimized AI could be deployed to shape political inclinations through carefully crafted, engagement-maximized content that exploits psychological vulnerabilities rather than presenting balanced information (Tappin *et al.*, 2023). **Misinformation amplification** poses another significant threat, as these systems could be used to create and disseminate false information in formats optimized for viral spread and cognitive acceptance (Lukito, 2020). Additionally, **consumer manipulation** risks emerge when such technologies are used to encourage ill-informed purchasing decisions or exploit behavioral biases for commercial gain (Boerman *et al.*, 2017).

Our work has consistently acknowledged these risks and implemented safeguards including Acceptable Use Policies that explicitly prohibit deployment in high-risk domains such as political campaigning, spam generation, and deceptive advertising. However, the fundamental challenge remains: how can we quantitatively measure the ”manipulation potential” of a behavioral model and develop

robust safeguards that prevent misuse while preserving beneficial capabilities?

6.2.2 Privacy and Data Protection Challenges

The development of behavior-optimized AI systems necessitates the collection and analysis of vast amounts of behavioral data, raising significant privacy concerns. Throughout our research—from analyzing persuasion strategies in advertising datasets to integrating behavioral signals like eye movements, comments, and engagement metrics—we have implemented comprehensive privacy-preserving strategies, including systematic removal of personally identifiable information (PII), restriction of data collection to enterprise accounts identified through structured knowledge bases, and aggregation of behavioral signals to prevent individual re-identification.

However, the scale and granularity of behavioral data required for effective modeling present ongoing challenges. Even aggregated behavioral data can potentially reveal sensitive information about individuals or groups, particularly when combined with other data sources. The development of formal differential privacy mechanisms (Dwork *et al.*, 2014) and federated learning approaches (McMahan *et al.*, 2017) represents promising directions for addressing these concerns, but significant technical challenges remain in balancing privacy protection with model utility in high-dimensional, multimodal behavioral modeling scenarios.

6.2.3 Bias, Fairness, and Representation

Behavioral data inherently reflects the biases present in the platforms and populations from which it is collected. Our research has consistently acknowledged that data sourced from diverse platforms—including advertising datasets for persuasion analysis, social media platforms for engagement modeling, eye-tracking studies for attention modeling, and video platforms for behavioral understanding—may exhibit demographic skews, self-selection bias, and algorithmic influences that could lead to uneven model performance across different user groups or reinforce existing societal biases.

These biases manifest in multiple dimensions: **demographic biases** related to age, gender, and socioeconomic background may influence how persuasion strategies are interpreted and modeled; **cultural biases** may affect the recognition and effectiveness of behavioral interventions across different cultural contexts; and **temporal biases** may result from the dynamic nature of online platforms and changing user behaviors over time.

Addressing these challenges requires comprehensive bias mitigation strategies including adversarial de-biasing techniques (Zhang *et al.*, 2018*a*), fairness-aware learning algorithms (Hardt *et al.*, 2016), and systematic bias auditing using subgroup performance analysis and counterfactual fairness testing. However, the fundamental challenge of ensuring fair and representative behavioral modeling across diverse global populations remains an active area of research.

6.2.4 Autonomy, Consent, and Human Agency

The development of increasingly sophisticated behavior prediction and optimization systems raises profound questions about human autonomy and agency. If AI systems can accurately predict and influence human behavior, what does this mean for individual freedom of choice and decision-making? Our work touches on these philosophical questions, particularly in the context of free will and determinism, suggesting that human behavior may be more predictable and influenceable than commonly assumed.

The concept of informed consent becomes particularly complex in the context of behavior-optimized AI. While users may consent to data collection for service improvement, they may not fully understand how their behavioral data will be used to create models capable of predicting and influencing their future actions. This asymmetry of understanding and power raises questions about the validity of consent in these contexts and the need for more transparent and comprehensible disclosure practices.

6.2.5 Accountability and Governance Frameworks

The deployment of behavior-optimized AI systems requires robust governance frameworks that can address the complex ethical challenges while enabling beneficial applications. Our research has contributed to this discussion through various approaches: implementing acceptable use policies for persuasion analysis datasets, developing controlled evaluation frameworks for engagement optimization, establishing privacy-preserving protocols for behavioral data collection, and proposing "Constitutional AI" approaches for persuasion, where models are trained to adhere to auditable ethical principles developed in collaboration with ethicists, psychologists, and policymakers.

However, significant challenges remain in developing effective governance mechanisms. **Technical auditing** requires new methods for assessing model behavior, detecting potential misuse, and ensuring compliance with ethical guidelines. **Regulatory frameworks** must balance innovation with protection, requiring close collaboration between technologists, policymakers, and civil society organizations. **Industry standards** need to be developed to ensure responsible development and deployment practices across the technology sector.

6.2.6 Long-term Societal Implications

The widespread deployment of behavior-optimized AI systems could fundamentally alter the nature of human communication and social interaction. The possibility of "infinite personalization"—where every message is optimized for maximum effectiveness with specific individuals—could create unprecedented levels of persuasive power while potentially undermining the shared information environment necessary for democratic discourse.

The development of digital societies and human simulation capabilities raises additional concerns about the potential for creating echo chambers, amplifying existing divisions, or manipulating public opinion at scale. While these technologies offer tremendous potential for understanding and addressing social challenges, they also create new risks that must be carefully managed.

6.2.7 Toward Responsible Development

Addressing these challenges requires a multi-faceted approach combining technical innovation with ethical reflection and regulatory oversight. Key principles for responsible development include:

Transparency and Explainability: Developing models whose decision-making processes can be understood and audited, enabling meaningful oversight and accountability.

Participatory Design: Involving diverse stakeholders, including affected communities, in the design and deployment of behavior-optimized AI systems.

Continuous Monitoring: Implementing systems for ongoing assessment of model behavior, societal impact, and potential misuse.

Adaptive Governance: Creating flexible regulatory frameworks that can evolve with technological developments while maintaining core ethical principles.

The societal implications of behavior-optimized AI are profound and multi-faceted. While our research demonstrates the tremendous potential of these technologies for beneficial applications, it also highlights the critical importance of addressing ethical challenges proactively. The future development of this field must prioritize not only technical advancement but also the careful consideration of societal impact, ensuring that the power to understand and influence human behavior is wielded responsibly and in service of human flourishing.

6.3 Open Research Questions

To spur progress, we outline key open research questions, categorized into four thematic areas.

Core Modeling and Architectural Challenges

- **Architectural Innovation for Multimodal Behavior:** What architectural innovations are needed to effectively fuse textual, visual, and behavioral modalities in a single, coherent model? A central challenge is balancing the computational overhead of incorporating diverse behavioral signals

against the performance gains in downstream tasks.

- **Modeling Long-term and Sequential Behavioral Dynamics:** Current models often predict immediate, atomic actions (e.g., clicks, likes). A more profound understanding requires modeling long-term, sequential dynamics. How does a series of information exposures shape a user's beliefs or habits over months or years? This necessitates a shift from stateless to stateful user models that capture evolving internal states (e.g., knowledge, preferences, attitudes).
- **Data-Efficient Behavioral Modeling:** How can we develop accurate behavioral models in domains with scarce data? This involves exploring few-shot, zero-shot, and transfer learning to leverage knowledge from data-rich environments (e.g., social media) for data-poor contexts (e.g., public health campaigns, niche product markets, or B2B interactions).

Causality, Interpretability, and Intervention

- **Integrating Causality into Predictive Models:** A key frontier is the integration of causal inference with predictive modeling. Future work should focus on hybrid models that combine the predictive power of large-scale machine learning with the explanatory power of causal reasoning. This could involve techniques for automatically identifying and leveraging natural experiments within massive datasets or developing architectures that learn causal graphs from observational and interventional data, moving from 'what' users will do to 'why' they do it.
- **From Prediction to Automated Intervention:** How can we build systems that move beyond passive prediction to actively recommend or perform interventions that achieve specific behavioral outcomes? This requires developing frameworks that can optimize for long-term goals (e.g., user well-being, sustained engagement, public health adherence) in complex, dynamic environments, while navigating the ethical considerations of such automated influence.
- **Generating Persuasive and Trustworthy Explanations:** Beyond predicting behavior, can we generate faithful explanations for *why* people act as they do? Furthermore, can we design models whose own recommendations are accompanied by explanations that are not only interpretable but also persuasive and trust-inspiring? How can we measure the behavioral impact of such machine-generated explanations?

Ethical and Societal Considerations

- **Ethical Frameworks for Persuasive Technologies:** As AI systems become more persuasive, it is imperative to develop robust ethical frameworks to guide their deployment. An actionable research direction is the creation of 'Constitutional AI' for persuasion, where models are trained to adhere to auditable ethical principles (e.g., transparency, fairness, respect for autonomy), co-developed with ethicists, psychologists, and policymakers.
- **Quantifying and Mitigating Manipulation Risk:** How can we quanti-

tatively measure the 'manipulation potential' of a behavioral model? Can we develop adversarial training techniques, formal verification methods, or auditing protocols to build models that are robust against misuse for malicious influence while preserving their beneficial capabilities?
- **Building and Validating Digital Societies:** How can we construct and validate large-scale, multi-agent simulations of digital societies to study emergent social phenomena? Research is needed to model phenomena like opinion polarization and information cascades, and to ethically test the impact of potential policy or platform design interventions in these simulated worlds.

Bridging Theory, Practice, and People

- **Cross-Cultural and Demographic Generalization:** How do we ensure that behavioral models are fair and generalize across diverse cultures, languages, and demographic groups? This requires creating new benchmark datasets that capture global diversity and developing methods to detect and mitigate biases present in training data.
- **Theory-Driven and Interdisciplinary Modeling:** How can we more deeply integrate established theories from psychology, sociology, and economics into data-driven models? This involves creating systems that are not only empirically powerful but also scientifically grounded and interpretable, fostering collaborative frameworks that bridge the gap between social science theory and AI practice.
- **Real-World Validation and Deployment:** How can we establish effective partnerships to validate and deploy behavioral models in real-world settings while maintaining academic rigor and ethical standards? This includes developing best practices for collaboration between academia, industry, and government to ensure that research translates into responsible and beneficial applications.

Finally, as we stand at the cusp of what we identified as the fourth major phase in the study of communication, driven by unprecedented access to digital content and behavioral data, we should remember these sayings:

We're actually much better at planning the flight path of an interplanetary rocket (rocket science) than we are at managing the economy, merging two corporations, or even predicting how many copies of a book will sell (behavior prediction). So why is it that rocket science ***seems*** *hard, whereas problems having to do with people - which arguably are much harder - seem like they ought to be* ***just*** *a matter of common sense (easily predictable)?* - Duncan J. Watts

And,

Nothing in Nature is random (unpredictable). A thing appears random only through the incompleteness of our knowledge (ignorance). - Baruch Spinoza

Further, as we reflect on the scientific and empirical challenges to the notion of free will, it is instructive to recognize that these questions have been deeply explored in philosophical and religious traditions for millennia. The Bhagavad Gita, a foundational text of Indian philosophy, offers a profound perspective on agency and action. It suggests that what we perceive as free will may, in fact, be shaped by forces beyond our conscious control—nature, past experiences, and even the divine. This resonates with the findings we have made in this thesis, as well as the findings of modern behavioral science and neuroscience, which increasingly point to the limits of conscious agency.

For example, in Chapter 3, Verse 27, the Gita states:

> "Prakṛteḥ kriyamāṇāni guṇaiḥ karmāṇi sarvaśaḥ; Ahaṅkāra-vimūḍhātmā kartāham iti manyate."

"All actions are performed by the modes of material nature, but a person deluded by false identification with the ego thinks, 'I am the doer.' "

And in Chapter 11, Verses 33 and 34, Krishna tells Arjuna:

> "droṇaṁ cha bhīṣhmaṁ cha jayadrathaṁ cha karṇaṁ tathānyān api yodha-vīrān mayā hatāṁs tvaṁ jahi mā vyathiṣhṭhā yudhyasva jetāsi raṇe sapatnān"

"Drona, Bhishma, Jayadratha, Karna, and other great warriors have already been killed by Me. You only be an instrument (nimitta) in the fight."

These verses articulate a philosophy of action that acknowledges the limits of personal agency and encourages detachment from the fruits of action. The Gita's perspective is not one of fatalism, but rather an invitation to act with awareness of the larger forces at play—recognizing that while we must act, we are not the sole authors of our actions. This ancient wisdom aligns with contemporary scientific insights, suggesting the nature, causes, and limits of human agency.

Publications covered as part of thesis

1. Singla, Y. K., Parekh, S., Singh, S., Chen, C., Krishnamurthy, B., & Shah, R. R. (2022). MINIMAL: Mining models for data-free universal adversarial triggers. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).

2. Khurana, V., Kumar, Y., Hollenstein, N., Kumar, R., & Krishnamurthy, B. (2023). Synthesizing Human Gaze Feedback for Improved NLP Performance. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (pp. 1895-1908).

3. Kumar, Y., Jha, R., Gupta, A., Aggarwal, M., Garg, A., Malyan, T., Bhardwaj, A., Ratn Shah, R., Krishnamurthy, B., & Chen, C. (2023). Persuasion Strategies in Advertisements. Proceedings of the AAAI Conference on Artificial Intelligence, 37(1), 57-66. https://doi.org/10.1609/aaai.v37i1.25076

4. Bhattacharya, A., Singla, Y. K., Krishnamurthy, B., Shah, R. R., & Chen, C. (2023). A Video Is Worth 4096 Tokens: Verbalize Videos To Understand Them In Zero Shot. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9822–9839, Singapore. Association for Computational Linguistics (EMNLP). (**Nominated for the best paper award!**)

5. Khandelwal, A., Agrawal, A., Bhattacharyya, A., Singla, Y.K., Singh, S., Bhattacharya, U., Dasgupta, I., Petrangeli, S., Shah, R.R., Chen, C. and Krishnamurthy, B., 2024. Large Content And Behavior Models To Understand, Simulate, And Optimize Content And Behavior. International Conference on Learning Representations (ICLR). (**Spotlight and nominated for award!**)

6. S I, H., Singh, S., K Singla, Y., Krishnamurthy, B., Chen, C., Baths V., and Ratn Shah, R. (2024). Long-Term Ad Memorability: Understanding and Generating Memorable Ads. Proceedings of the Winter Conference on Applications of Computer Vision (WACV).

7. Khurana, V., Singla, Y.K., Subramanian, J., Shah, R.R., Chen, C., Xu, Z. and Krishnamurthy, B., 2025. Measuring and Improving Engagement of Text-to-Image Generation Models. International Conference on Learning Representations (ICLR).

8. Singh, S., SI, H., Singla, Y. K., Baths, V., Shah, R. R., Chen, C., and Krishnamurthy, B. (2025). Teaching Human Behavior Improves Content Understanding Abilities Of VLMs. International Conference on Learning Representations (ICLR).

Dataset contributions covered as part of this thesis

1. Persuasion Strategies for Images: https://midas-research.github.io/persuasion-advertisements/
2. Persuasion Strategies for Videos: https://behavior-in-the-wild.github.io/video-4096.html
3. Content Behavior Corpus (CBC): https://behavior-in-the-wild.github.io/LCBM.html
4. Long-term memorability of Advertisements:
   (a) LAMBDA: https://behavior-in-the-wild.github.io/memorability.html
   (b) UltraLAMBDA: https://behavior-in-the-wild.github.io/memorability.html
5. Behavior-LLaVA Instruction Fine-Tuning Dataset (BLIFT): https://behavior-in-the-wild.github.io/teaching-behavior-improves-content-understanding
6. EngagingImageNet: https://behavior-in-the-wild.github.io/measure-engagement

Other Publications During PhD

1. Singh, S., Singla, Y. K., SI, H., & Krishnamurthy, B. (2025). Measuring and Improving Persuasiveness of Large Language Models. The Thirteenth International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=NfCEVihkdC

2. Bhattacharyya, A., Singla, Y. K., Aggarwal, S., Menta, T., SR, N., & Krishnamurthy, B. (2025). Align via actions: Learning behavior aligns LLMs with human opinions in zero-shot. Association for Computational Linguistics Rolling Review. **Nominated for Best Paper Award**.

3. Choudhary, N., Goyal, P., Siwatch, D., Chandak, A., Mahajan, H., Khurana, V., & Singla, Y. K. (2025). AdQuestA: Knowledge-guided visual question answer framework for advertisements. Proceedings of the Winter Conference on Applications of Computer Vision (WACV).

4. Patnaik, S., Changwal, H., Aggarwal, M., Bhatia, S., Singla, Y. K., & Krishnamurthy, B. (2024). CABINET: Content relevance-based noise reduction for table question answering. Proceedings of the International Conference on Learning Representations (ICLR). **Spotlight**.

5. Anand, A., Nair, A., Prasad, K., Narayan, V., Lal, N., Mahata, D., Singla, Y. K., & Shah, R. (2024). Advances in citation text generation: Leveraging multi-source Seq2Seq models and large language models. Proceedings of the Conference on Information and Knowledge Management (CIKM).

6. Singla, Y. K. (2023). Can we use some advances in AI to teach AI? How could we make AI education more interdisciplinary? AI Matters as part of EAAI-23 Blue Sky Ideas in Artificial Intelligence Education from the AAAI/ACM SIGAI New and Future AI Educator Program.

7. Singla, Y. K., Singh, S., Parekh, S., Li, J. J., Shah, R. R., & Chen, C. (2023). Automatic essay scoring systems are both overstable and oversensitive: Explaining why and proposing defenses. Dialogue and Discourse Journal.

8. S., S., Pupneja, A., Mital, S., Shah, C., Bawkar, M., Gupta, L. P., Kumar, A., Singla, Y. K., Gupta, R., & Shah, R. R. (2023). H-AES: Towards automated essay scoring for Hindi. Proceedings of the Educational Advances in Artificial Intelligence (EAAI) at AAAI.

9. Ghosh, S., Kumar, S., Singla, Y. K., Shah, R. R., & Umesh, S. (2022). Span classification with structured information for disfluency detection in spoken utterances. Proceedings of Interspeech.

10. Singla, Y. K., Krishna, S., Shah, R. R., & Chen, C. (2022). Using sampling to estimate and improve performance of automated scoring systems with

guarantees. Proceedings of the AAAI Conference on Artificial Intelligence - Educational Advances in Artificial Intelligence (AAAI-EAAI).

11. Grover, M., Bamdev, P., Singla, Y. K., Vafaee, P., Hama, M., & Shah, R. R. (2022). Automated speech scoring system under the lens: Evaluating and interpreting the linguistic cues for language proficiency. International Journal of Artificial Intelligence in Education.

12. Mathur, A. N., Kumar, Y., Batra, D., Shah, R. R., Zimmermann, R., & Chen, C. (2021). Lifi: Towards linguistically informed frame interpolation. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

13. Singla, Y. K., Gupta, A., Bagga, S., Chen, C., Krishnamurthy, B., & Shah, R. R. (2021). Speaker-conditioned hierarchical modeling for automated speech scoring. Proceedings of the Conference on Information and Knowledge Management (CIKM).

14. Sikka, J., Satya, K., Kumar, Y., Uppal, S., Shah, R. R., & Zimmermann, R. (2020). Learning-based methods for code runtime complexity prediction. Proceedings of the European Conference on Information Retrieval (ECIR).

15. Sahrawat, D., Mahata, D., Zhang, R., Kulkarni, M., Sharma, A., Gosangi, R., Stent, A., Kumar, Y., Shah, R. R., & Zimmermann, R. (2020). Keyphrase extraction as a sequence labeling task using transformers. Proceedings of the European Conference on Information Retrieval (ECIR).

16. Kumar, Y., Sahrawat, D., Maheshwari, S., Mahata, D., Shah, R. R., Yin, Y., Zimmermann, R., & Stent, A. (2020). Harnessing GANs for zero-shot learning of new classes in visual speech recognition. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).

17. Srivastava, N., Saxena, A., Kumar, Y., Mahata, D., Shah, R. R., Stent, A., & Zimmermann, R. (2019). MobiVSR - Mobile application for visual speech recognition. Proceedings of Interspeech.

18. Uttam, S., Kumar, Y., Sahrawat, D., Aggarwal, M., Mahata, D., Shah, R. R., & Stent, A. (2019). Hush-Hush speak: Speech reconstruction using silent videos. Proceedings of Interspeech.

19. Kumar, Y., Jain, R., Mohd. Salik, K., Shah, R. R., Yin, Y., & Zimmermann, R. (2019). Lipper: Synthesizing thy speech using multi-view lipreading. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).

20. Kumar, Y., Aggarwal, S., Mahata, D., Shah, R. R., Kumaraguru, P., & Zimmermann, R. (2019). Get IT scored using AutoSAS - An automated system for scoring short answers. Proceedings of the Educational Advances in Artificial Intelligence (EAAI) at AAAI.

21. Kumar, Y., Jain, R., Mohd. Salik, K., Shah, R. R., Yin, Y., & Zimmermann, R. (2018). MyLipper: A personalized system for speech reconstruction using multi-view visual feeds. Proceedings of the International Symposium on Multimedia (ISM). **Best Poster Runner-up**.

22. Kumar, Y., Aggarwal, M., Nawal, P., Satoh, S., Shah, R. R., & Zimmermann, R. (2018). Harnessing AI for speech reconstruction using multi-view silent video feed. Proceedings of the ACM Multimedia Conference (ACMMM).

Patents During PhD

1. Kumar, Y., Singh, S., Indumathi, HSK., Krishnamurthy, B., & Bhattacharyya, A. (2025), Systems and methods for predicting content memorability (US Patent App. 18/538,988). U.S. Patent and Trademark Office.

2. Gupta, M., Kumar, Y., Gupta, R., Bothra, P., Hemani, M., Gupta, M., & Makkar, G. (2024), Generating digital content (U.S. Patent App. No. 20240362427A1). U.S. Patent and Trademark Office.

3. Kumar, Y., Singh, S., Park, S., Prasoon, P., Sainath, N., Joshi, N. S., Srikanth, N., Puri, N., Aggarwal, M., Subramanian, J., Palwe, G., Krishnamurthy, B., Rozen, M. W., Naware, M., & Chung, H. (2024). Digital content analysis (U.S. Patent App. No. 20240355020). U.S. Patent and Trademark Office.

4. Kumar, Y., & Khurana, V. (2024). Systems and methods for generating scanpaths (U.S. Patent App. No. 18/109,990). U.S. Patent and Trademark Office.

5. Kumar, Y., Khuc, V. N., Srivastava, V., Moorarka, U., Verma, S., Shahid, S., Bansal, S., Venkitachalam, S., Steimer, S., Karmakar, S., Srivastav, N., Puri, N., Naware, M., Jain, K. K., Singh, K. M., Chung, H., Bacila, H., Lordache, F. S., Pai, D., & Krishnamurthy, B. (2024). Determining user affinities for content generation applications (U.S. Patent No. 12,008,033). U.S. Patent and Trademark Office.

6. Kumar, Y., Ahlawat, V., Zhang, R., Aggarwal, M., Palwe, G. K., Krishnamurthy, B., & Khurana, V. (2024). Attention aware multi-modal model for content understanding (U.S. Patent App. No. 17/944,502). U.S. Patent and Trademark Office.

7. Bhattacharyya, A., & Kumar, Y. (2024). Generating video insights based on machine-generated text representations of videos (U.S. Patent App. No. 18/481,882). U.S. Patent and Trademark Office.

8. Kumar, Y., Singh, S., George, W. B., Liu, T. C., Basetty, S., Prasoon, P., Puri, N., Naware, M., Corlan, M., Butikofer, J. M., Chauhan, A., Singh, K. M., O'Reilly, J. P., Chung, H., Dest, L., Goudie-Nice, C. H., Pack, B. J., Krishnamurthy, B., Jain, K. K., Klimetschek, A., & Rozen, M. W. (2024). Content analytics as part of content creation (U.S. Patent No. 11,907,508 B1). U.S. Patent and Trademark Office.

9. Kumar, Y., & Krishnamurthy, B. (2023). Visual speech recognition for digital videos utilizing generative adversarial learning (U.S. Patent App. No. 17/650,020; CN Patent App. No. 202211407981.2A; DE Patent App. No. 102022131824.9A). U.S. Patent and Trademark Office, China National

Intellectual Property Administration, & German Patent and Trade Mark Office.

10. Kumar, Y. (2021). Pose-invariant visual speech recognition using a single view input (U.S. Patent No. 10,937,428). U.S. Patent and Trademark Office.

Awards

- Google PhD Fellow (1 among 40 candidates all over the world)
- The Indian Prime Minister PhD Fellowship (1 among 38 candidates all over India across all STEAM branches)
- University at Buffalo - PhD Best First-Year Achiever Award
- IIIT-Delhi Dean IRD Research Excellence Award for my PhD research for three consecutive years (2021-23)
- Adobe Outstanding Young Engineer Award
- Two spotlight paper awards at ICLR-24
- Nominated for Best Paper Award at EMNLP-23
- Best paper award nomination ACL Rolling Review-24
- Best poster paper runner up (*i.e.* 2nd position) award at ISM-2018
- Best Student Abstract Research Paper at AAAI-2019
- Best Adobe Sneaks Award - 2021 Project #CatchyContent. News Coverage: Forrester, Fast Company, and other media and research houses
- Awarded AI Future Educator Travel Award by EAAI/AAAI-23
- Heidelberg Laureate Young Researcher - 2021
- SIGIR Student Grant-2021
- Winner (*i.e.* 1st position) of ACM All-India Student Chapters Challenge 2019

List of Tables

2.1 The questions we asked to the non-expert annotators to help them identify persuasion strategy contained in the video advertisement. 38

2.2 Effect of different Modalities and Tasks on the accuracy and performance of the strategy prediction task. 45

2.3 Comparison of caption and object detection models. We noticed that BLIP while being more recent and trained on a larger dataset, generates more informatory captions for background objects which DenseCap successfully ignores. 46

2.4 Distribution of test, train, validation, and the total dataset 46

2.5 Comparison of Pyscenedetect (Breakthrough, 2023) with uniform sampling of choosing video frames. Based on downstream performance, we can see that uniform sampling works better than Pyscenedetect 51

2.6 Comparison of all the models across topic, emotion, and persuasion strategy detection tasks. We see that our framework, despite being zero-shot, outperforms finetuned video-based models on the topic classification, persuasion strategy detection and action and reason classification tasks and comes close on the emotion classification task. Further, the Roberta classifier trained on generated stories outperforms both finetuned and zero-shot models on most tasks. Best models are denoted in green and runner-ups in blue . . . 56

2.7 Comparison on story generation task on the video-story dataset. We see that our framework despite being zero-shot outperforms all the fine-tuned generative prior art on all metrics. Further, it also outperforms fine-tuned retrieval models, which choose from a fixed set of frame descriptions on most metrics. Best models are denoted in green and runner-ups in blue. 56

2.8 Comparison of the different zero-shot models on the action and reason generation tasks. Note that there are no fine-tuned generative models in the literature for this task and the number of annotated videos is too small to train a generative model. Best models are denoted in green. 57

2.9 Comparison of various models on the LVU benchmark. We see that our framework, despite being zero-shot, outperforms fine-tuned video-based models on 8/9 tasks. Best models are denoted in green and runner-ups in blue. 57

2.10 Comparison of various models on the HVU benchmark (Diba *et al.*, 2020). The models scores are as reported in (Diba *et al.*, 2020). We see that our framework, despite being zero-shot, outperforms fine-tuned video-based models on all the tasks. Best models are denoted in green and runner-ups in blue. 57

2.11 Ablation study of using only visual (caption) or audio (transcripts) and LLMs for downstream tasks. It can be noted that the overall model does not perform as well (compared to Table 2.6) when using only audio or scene description without generating story. . . . 59

2.12 Top-5 accuracy, and mAP for persuasion strategy detection task 59

2.13 Number of Samples required to generate Universal Adversarial Triggers for each Dataset. In a data-based approach like (Wallace *et al.*, 2019), validation set (column 2) is used to generate the UATs. The third column lists the number of queries we make to generate artificial samples. These artificial samples are then used to craft UATs. Note that no real samples are required for our method. 71

2.14 Class Impressions for BiLSTM-Word2Vec Sentiment Analysis Model. Note that the words in the class impression examples highly correspond to the respective sentiment classes. 72

2.15 The table reports the accuracy drop for the BiLSTM-Word2Vec sentiment analysis model after prepending 3-word adversarial triggers generated using MINIMAL and data-based methods. 73

2.16 Accuracy drop for transfer attack with data-free UAT generated by our method. We prepend 3-word adversarial triggers to the SST BiLSTM-ELMo model. 73

2.17 Class Impressions for ESIM model trained for the Natural Language Inference Task 74

2.18 We prepend a single word (Trigger) to SNLI hypotheses. We display the top 3 triggers created using both Validation set and Class Impressions for ESIM and show their performance on the DA. The original accuracies are mentioned in brackets. 75

2.19 Class Impressions for ALBERT model trained for the Microsoft Research Paraphrase Corpus 76

2.20 Accuracy drop for the ALBERT paraphrase identification model after prepending 3-word adversarial triggers generated using MINIMAL. 80

2.21 PMI percentiles for sample class impression words and their average 80

2.22 PMI percentiles for sample class impression words and their average 81

2.23 PMI percentiles for sample class impression words and their average 81

2.24 We prepend a single word trigger to SNLI hypotheses. We take the first word from all ground truth class impressions and evaluate them on class impressions of the target class. 81

3.1 **Behavior Simulation.** Mean RMSE and accuracy scores for scene-by-scene predictions of video replay values. Replay values are the normalized replay scores of each scene as provided by YouTube. The normalized scores are considered to 2 decimal places and multiplied by hundred to convert the score to an integer score in the range 0-100. RMSE is calculated for each video in the test set and the mean is calculated for this score and reported. The model is said to classify correctly if the absolute error between the predicted and ground truth value is less than or equal to 5. The scores are calculated in four regimes: past, future, random, and all-masked. In the past (future) regimes, first (last) 5-20% scenes are masked; in the random setting, 5-20% scenes are masked randomly, and in all masked setting, everything is masked. LCBM was behavior-fine-tuned (BFT) with 3,5,7,11 context window masking strategy, while GPT was compared with an in-context learning (ICL) setting. We note that behavior fine-tuned LCBM, while being at least 10x smaller than other models, performs the best. Best models are denoted in green and runner-ups in blue. 106

3.2 Details of the in-house Marketing Email dataset used to evaluate behavior generalization capabilities of the LCBM 109

3.3 **Behavior Simulation and Behavior Domain Adaptation**[‡]. Two-way classification accuracies for like prediction on Twitter. Given content, channel, and time, predict behavior (High, Low). We note that LCBM trained on Twitter and YouTube performs better than the one trained only on Twitter, showing signs of performance improvement by domain adaptation. 110

3.4 **Content Simulation and Behavior Domain Adaptation**[‡]. Given behavior, channel, time, tweet media caption as prompt, predict content (tweet text). We note that LCBM trained on Twitter and YouTube performs better than the one trained only on Twitter, showing signs of performance improvement by domain adaptation. 110

3.5 **Content Simulation.** In this task, the models have to choose the speech segment from a list of 25 options given the video description, non-masked scenes. and replay behavior. We see that despite being similar to masked language modeling (which is a content-only task), LCBM performs better than both Vicuna and GPT-3.5. Best models are denoted in green and runner-ups in blue 112

3.6 **Behavior Understanding.** In this task, the models have to simulate the sentiment of comments that a video would get by looking at only the video. Further, they also have to explain the reason for such sentiment. The responses were annotated by humans on a scale of 0-5 for the reason, with 0 being no response provided and 5 being the response matches exactly with the (ground truth) comments received on the video. Best models are denoted in green and runner-ups in blue 112

3.7 **Behavior Simulation.** RMSE, R^2, and accuracy scores for like/view ratio prediction task. To calculate accuracy, the model is said to classify correctly if the absolute error between the predicted and ground truth likes/views is less than or equal to 10%. BFT denotes behavior fine-tuning, and ICL stands for in-context learning. Replay values k-masked means a model which is trained by masking k consecutive values of the replay graph while doing BFT. We note that LCBM while being at least 10x smaller than the other models, performs the best. The best results over four runs are reported for all models. Best models are denoted in green and runner-ups in blue 114

3.8 **Content Understanding.** Comparison of several models, including behavior instruction tuned models before and after BFT. We compare the models across topic, emotion, and persuasion strategy detection tasks as per the framework given by Bhattacharyya *et al.* (2023). We see that our model outperforms similarly sized models (Vicuna, VideoChat) in most tasks. Best models are denoted in green and runner-ups in blue 114

3.9 **Behavior Domain Adaptation.** We test the generalization capability of LCBM on two tasks: (1) Behavior simulation on in-house Email Marketing Data, (2) Behavior simulation on the LVU benchmark. For (1), we train two versions of LCBM with the in-house Email Marketing data: one was trained on YouTube videos and further BFT on a few email samples (*domain-adapted*), and the other was BFT on a larger set of emails, but not including YouTube data (*in-domain*)§. We report the RMSE and R^2 scores for this task. For (2), we compare LCBM with other state-of-the-art results and GPT-3. In (1), we note that the domain-adapted LCBM performs better than the in-domain LCBM in both settings. We posit that YouTube data helps LCBM understand how a company's viewers like to hear from it, giving LCBM an edge over a model trained on a small amount of the same data (600 unique emails). In (2), LCBM performs better than the existing state-of-the-art. Surprisingly, GPT-3.5 does better than LCBM on this task. From both (1) and (2), we gather that a model trained on certain YouTube behaviors performs better on other behaviors, thus showing promise of domain-adaptation in the behavior modality. Best models are denoted in green and runner-ups in blue. . . 115

4.1 In-domain Evaluation of Scanpath Generation on the CELER dataset (Berzak *et al.*, 2022). 140

4.2 Cross-domain Evaluation of Scanpath Generation on the Dataset by (Mishra *et al.*, 2016*a*). 143

4.3 Cross-domain Evaluation of Scanpath Generation on the Dataset by (Mishra *et al.*, 2017). 143

4.4 Sentiment analysis and sarcasm detection results on the dataset by (Mishra *et al.*, 2016*a*). Model configuration refers to the type of scanpath included in train and test data. 145

4.5 Results of training NLP models with and without scanpaths on the GLUE benchmark tasks. Including scanpaths leads to consistent improvements across all the NLP tasks. 146

4.6 Ablation on using comments and/or perception signals from Salicon 161

4.7 Ablation on different sampling ratios and epochs of training. Sampling ratio is the ratio of behaviour data to multimodal instruct data. Performance is the average increase in 0-shot accuracy on 6 tasks with 250 samples each from the eval set. These tasks include image emotion recognition, video emotion recognition, persuasion strategy classification, MSRVTT, HVU and MSVD-QA 162

4.8 Comparison of various models on the Long Video Understanding benchmark (Wu and Krahenbuhl, 2021) consisting of 9 VQA tasks. We see that Behavior-LLaVA improves on LLaMA-Vid on 9/9 tasks with an average improvement of 21.49%. Further, it outperforms the state-of-the-art in 5/9 tasks. 164

4.9 Comparison of various models on three video emotion understanding benchmarks (Video Emotion8 (Jiang *et al.*, 2014), CAER (Lee *et al.*, 2019), Ekman-6 (Xu *et al.*, 2016*a*)). The main goal of comparing on these benchmarks is to demonstrate Behavior-LLaVA's understanding of complex tasks like video emotions of long-form videos. We see that Behavior-LLaVA improves on LLaMA-Vid on 3/3 benchmarks with an average improvement score of 51.85% in zero-shot and 3.39% in fine-tuned settings. Further, it outperforms the current state-of-the-art on 3/3 benchmarks in zero-shot and 1/3 in fine-tuned settings. 167

4.10 Evaluation on audio and text modalities. We evaluate on the audio summarization benchmark (Han *et al.*, 2023) for audio and IMDB sentiment benchmark for text (Maas *et al.*, 2011). 167

4.11 Comparison of various models on two video understanding benchmarks (Hussain *et al.*, 2017; Kumar *et al.*, 2023*b*) consisting of 5 tasks related to video advertisements understanding. The main goal of comparing on these benchmarks is to demonstrate Behavior-LLaVA's understanding of complex videos. We see that Behavior-LLaVA improves on LLaMA-Vid on 5/5 tasks with an average improvement score of 43.18% in zero-shot and 27.64% in fine-tuned settings. Further, it outperforms the current state-of-the-art on 3/5 tasks in zero-shot and 5/5 in fine-tuned settings. Full results are presented in the Table 4.12. 168

4.12 Comparison of various models on two video understanding benchmarks (Hussain *et al.*, 2017; Kumar *et al.*, 2023*b*) consisting of 5 tasks related to video advertisements understanding. The main goal of comparing on these benchmarks is to demonstrate Behavior-LLaVA's understanding of complex videos. We see that Behavior-LLaVA improves on LLaMA-Vid on 5/5 tasks with an average improvement score of 43.18% in zero-shot and 27.64% in fine-tuned settings. Further, it outperforms the current state-of-the-art on 3/5 tasks in zero-shot and 5/5 in fine-tuned settings. 169

4.13 Comparison of various models on seven video and image memorability benchmarks (Memento10k (Newman *et al.*, 2020), VideoMem (Cohendet *et al.*, 2019), LaMem (Khosla *et al.*, 2015), SUN (Isola *et al.*, 2011), MemCat (Goetschalckx and Wagemans, 2019), MediaEval (Kiziltepe *et al.*, 2021*a*), LAMBDA (SI *et al.*, 2025)). The main goal of comparing on these benchmarks is to demonstrate Behavior-LLaVA's understanding of complex and high-level tasks like memorability simulation. We see that Behavior-LLaVA improves on LLaMA-Vid on 7/7 benchmarks with an average improvement score of 186.4% in zero-shot and 39% in fine-tuned settings after seeing 25% train data. Further, it performs similarly to the current state-of-the-art on 7/7 benchmarks in the fine-tuned settings while still seeing only 25% data. 170

4.14 Comparison of various models on the image dense captioning task. The main goal of this task is to demonstrate Behavior-LLaVA's image captioning ability. Despite not being explicitly trained on this task, Behavior-LLaVA performs better than both Ad-LLaVA and LLaMA-Vid on Detail and Quality aspects while losing marginally on correctness. On the aspects of detail and quality, it even outperforms the much larger model of LLaVA-1.6 (34B). 171

4.15 Comparison of various models on four image emotion understanding benchmarks (IAPSa-8 (Mikels *et al.*, 2005) Abstract (Machajdik and Hanbury, 2010), Emotion6 (Peng *et al.*, 2015), Emoset (Yang *et al.*, 2023)). The main goal of comparing on these benchmarks is to demonstrate Behavior-LLaVA's understanding of complex tasks like image emotions. We see that Behavior-LLaVA improves on LLaMA-Vid on 4/4 benchmarks with an average improvement score of 29.14% in zero-shot and 8.95% in fine-tuned settings. Further, it outperforms the current state-of-the-art on 4/4 benchmarks in the fine-tuned settings. 172

4.16 Comparison of various models on three conventional video question answering benchmarks consisting of question answers related to action understanding. The main goal of comparing on this benchmark is to show that Behavior-LLaVA does not perform worse on low-level understanding tasks like action recognition. We see that Behavior-LLaVA marginally improves on LLaMA-Vid on 2/3 benchmarks. Further, it performs equivalent to the state-of-the-art in 3/3 benchmarks. 172

4.17 Comparison of various models on the Holistic Video Understanding benchmark (Diba *et al.*, 2020) consisting of 7 VQA tasks. We see that Behavior-LLaVA improves on LLaMA-Vid on 6/7 tasks with an average improvement of 5.88%. Further, it performs equivalent to the state-of-the-art in 6/7 tasks. State of the art is achieved by generating a story and asking Behavior-LLaVA to answer questions based on the generated story. 173

4.18 Improvement on downstream content understanding tasks by introducing more behaviour signals. Brackets [] denote the new behaviour that we include. Replay graphs (Khandelwal *et al.*, 2024). Mem-Recalls (SI *et al.*, 2025) Evaluation done on Multi-shot video summarization (Han *et al.*, 2023) and MomentDETR (Lei *et al.*, 2021) 178

5.1 Comparison of all the major (image and video) memorability datasets available in the literature along with LAMBDA (ours). The datasets are compared on the following axes: number of samples, type of memorability (short-term (ST) and long-term (LT)), memory retrieval process (recall or recognition), type of content (images/videos and their type), duration with which the sample was shown on the participants' screen, whether audio was present or not, human consistency achieved in the study, and the protocol followed in the study to collect the data. **Memento10k** - (Newman *et al.*, 2020), **VideoMem** - (Cohendet *et al.*, 2019), **LaMem** - (Khosla *et al.*, 2015), **SUN** - (Isola *et al.*, 2011), **MemCat** - (Goetschalckx and Wagemans, 2019), **MediaEval** - (Kiziltepe *et al.*, 2021*b*) 192

5.2 Results of Henry (our model) on eight datasets compared with the current best models reported in the literature and GPT-3.5. Human consistency values are also listed in the top row for reference. It can be observed that our model achieves state-of-the-art performance across all datasets. Best models are denoted in green and runner-ups in blue. References for the seven literature SOTA models in the format {`dataset: SOTA model citation`} are: LaMem: (Hagen and Espeseth, 2023), MemCat: (Hagen and Espeseth, 2023), SUN: (Fajtl *et al.*, 2018), Merged Image datasets: (Hagen and Espeseth, 2023), Memento10k: (Dumont *et al.*, 2023), VideoMem: (Dumont *et al.*, 2023), MediaEval: (Lu and Wu, 2021) 199

5.3 **Ad Generation**: Results of Henry-SEED compared with in-context-learning (ICL) GPT-3.5, 4 on Ad-Memorability and Ad generation quality. See §5.1.5 for details of the metrics computed. We see that Henry-SEED generated ads are more memorable than ads generated using 15x larger GPT-3.5 and GPT-4. We test ad quality using GPT-4 as judge and then test the top-two models using human annotators. GPT-4 as a judge rates GPT-4 and Henry-SEED as the top two models. Subsequently, we ask humans to select between the original and generated ad stories. We observed that human annotators preferred Henry-SEED ads more than the original ads 3/5 times, while GPT-4 generated ads are preferred 2/5 times over the original ads. Further, we note that an increase in the amount of training data for Henry-SEED increases its performance across all metrics. Figs. 5.4-5.6 and Listings 5.1-5.10 contain some qualitative samples generated using Henry-SEED. 208

5.4 **Ad Generation**: Perplexity comparison (refer §5.1.11) of LLaVA and Henry-SEED on low/medium/high memorable ads from LAMBDA test set. We see that untrained LLaVA does not favor memorable ads. Further, we note that when synthetic data is included during training, the ratio of perplexity on low and high ads grows from 2.79 to 3.01. 211

5.5 To augment the scene understanding of LLM, we verbalize video scenes and images using a diverse set of cognitive and perception tools and pass it to the LLM in the format shown in the table. For image memorability datasets, we use the following semantic categories: caption, color, photo style, emotion, clutter, human presence, object tags, OCR, and aesthetics. For video scene memorability datasets, we use the following semantic categories: caption, color, emotion, human presence, object tags, ASR, OCR, Audio-type, Logo-presence. We use the following models to extract the features: OCR (Du *et al.*, 2020), clutter (Khurana *et al.*, 2023), ASR (Radford *et al.*, 2022), Photo style (Li *et al.*, 2023*b*), human presence (Liu *et al.*, 2023*c*), emotion (Singh *et al.*, 2024*a*), caption (Li *et al.*, 2023*b*), aesthetics (Ke *et al.*, 2023), colors (Qin *et al.*, 2020*a*), object tags (Zhang *et al.*, 2023*c*), audio-type (Giannakopoulos, 2015), and logo presence (Zhang *et al.*, 2023*c*). Black colored text is the verbalization template, and red text indicates the model outputs. 223

5.6 Ablation across data to understand how memorability prediction generalizes across the type of memory, datasets, modality (image/video), and brands. The reported values are correlations between model and human memorability scores. A few trends can be observed from the table: (i) STM generalizes better on LTM in zero-shot than vice versa (rows 1 and 2), (ii) Henry trained on either videos or images generalizes to both (rows 3 and 4), (iii) There is a significant performance loss in modeling memorability for brands not seen during training (row 5), (iv) Zero-shot generalization to Memento (video) and Memcat (image) is near to the current trained state of the art literature models on Memento (Dumont *et al.*, 2023) and Memcat (Hagen and Espeseth, 2023) (rows 6 and 7). 225

5.7 Ablation across architectural choices. "-" denotes non-speech dataset. A few trends are visible from the table: (i) Despite having a vision branch, object tags and colors have a net positive impact on the overall performance (rows 2,3,4), (ii) For LTM (LAMBDA, VideoMem (LT)), dropping cognitive features such as emotion, aesthetics, and clutter cause a larger performance drop than dropping visual features such as objects and colors. The trend is the opposite for STM (Lamem, Memcat, VideoMem (ST), Memento10k). . . 225

5.8 Ablation on modeling behavior simulation (BS) or memorability prediction and Content Simulation (CS) on memorable ad generation together. For memorability prediction, we again show the Spearman rank correlation on the test set similar to Table 5.2; for generation, we measure the change in memorability according to Henry Oracle similar to Table 5.3. We observe that mixing the two tasks together increases the performance across both tasks. . . . 226

5.9 A comparison of datasets containing image preferences 247

5.10 Pearson correlation between model predicted scores and image engagement measured by account normalized likes. 249

5.11 Results reveal the significant gains achieved in improving the engagement of low-liked subset of the EngagingImageNet dataset by enhancing the image descriptions fed to a text-to-image model, as described in Section 5.2.4.1. 253

5.12 Comparing the performance gains on the EngagingImageNet test dataset, resulting from train-time engagement-optimization methods applied on stable diffusion, as described in Sections 5.2.4.2 and 5.2.4.3. 256

5.13 Pearson correlation coefficients (r) and associated p-values for the relationship between marketer-allocated advertisement budget and five key performance indicators (KPIs): Impressions, Clicks, Cost Per Click (CPC), Cost Per Thousand Impressions (CPM), and Cost Per Purchase (CPP). Budget allocation serves as a proxy for marketer confidence in advertisement efficacy. Data were collected from a Fortune 500 company's marketing campaigns (n > 1,000 advertisements) over a 12-month period. Results suggest no statistically significant correlation between marketing spend and advertisement performance across all measured KPIs, indicating potential limitations in expert marketers' ability to predict advertisement success. 262

5.14 Distribution of ground truth EngagingImageNet images 264

5.15 Performance of all models on the engagement-optimized design specification generation (DSG) task across different engagement-level images in EngagingImageNet data. It is noteworthy that (i) EngageNet outperforms larger sized GPT-3.5, 4 and also the sized Llama model fine-tuned on the same data (without including engagement tokens). (ii) In-context learning does not work well in the engagement-conditioned design specification generation domain. 271

5.16 Results for the performance of various models on the EngagingImageNet for the engagement-optimized image generation task. Results are computed on the EngageNet Design Specification Generation (DSG) reward (§5.2.7.8) as well as other metrics reported in the literature. 277

List of Figures

1.1 Communication process can be defined by seven factors: Communicator, Message, Time of message, Channel, Receiver, Time of effect, and Effect. Any message is created to serve an end goal. For marketers, the end goal is to bring in the desired receiver effect (behavior) (like clicks, purchases, likes, and customer retention). The figure presents the key elements in the communication pipeline - the marketer, message, channel, receivers, and finally, the receiver effect. 1

1.2 Interaction of the communication modalities in a mass communication scenario like a speech. 2

1.3 Interaction of the communication modalities in a mass communication scenario on a social media platform like Twitter. 3

1.4 Interaction of the communication modalities in a person-to-person communication scenario on a channel like Whatsapp. 4

1.5 Interaction of the communication modalities in person-to-person communication with bidirectional communication. 5

1.6 Levels of content analysis. The figure lists tasks and their sample outputs arranged in a hierarchy (Shannon and Weaver, 1949). This is roughly based on levels of language. Notably, humans are good at predicting the first three levels but not the last level (Tetlock, 2017; Collaborative, 2023; Tan *et al.*, 2014; Isola *et al.*, 2013). . 11

1.7 Communication process can be defined by seven factors: Communicator, Message, Time of message, Channel, Receiver, Time of effect, and Effect. Any message is created to serve an end goal. In this thesis, we explore the two main concerns of behavioral sciences: understanding (or explanation) and prediction. The figure shows the links between the different chapters and how they link together to form the two core pillars of understanding and explanation. 13

2.1 Different persuasion strategies are used for marketing the same product (footwear in this example). The strategies are in red words and to be defined by us in the paper. 21

2.2 Examples of videos with their annotated persuasion strategies. Relevant keyframes and ASR captions are shown in the figure, along with the annotated strategies. These two videos can be watched at `https://bit.ly/3Ie3JG0`, `https://bit.ly/3OgtLwj` 21

2.3 Example of various rhetoric strategies used in advertisements . . 23

2.4 Persuasion strategies in advertisements. Marketers use both text and vision modalities to create ads containing different messaging strategies. Different persuasion strategies are constituted by using various rhetorical devices such as slogans, symbolism, colors, emotions, allusion. . . 27

2.5 Distribution of Persuasion Strategies in the image persuasion strategy dataset. The top-3 strategies are Concreteness, Eager, and Fashionable. 31

2.6 Image with a segmentation mask depicting the strategies *Emotion:Cheerful*, *Emotion:Eager* and *Trustworthiness*. 34

2.7 Advertisements containing humans and concreteness 35

2.8 Dice correlation between topics and strategies. Topics are taken from the Pitts Ad dataset and further similar topics are combined to get these values. 36

2.9 Distribution of persuasion strategies in our video persuasion strategy dataset 37

2.10 Some samples from the Pitts Ads dataset along with the ground truth and predicted action-reason statement, topic and sentiments. 39

2.11 Architecture of the Persuasion Strategy Prediction model. To capture the different rhetoric devices, we extract features for the image, text, and symbolism modalities and then apply cross-modal attention fusion to leverage the interdependence of the different devices. Further, the model trains over two tasks: persuasion strategies and the reasoning task of action-reason prediction. 43

2.12 Incremental effect of introducing new data through active learning; Results for prediction of persuasion strategies on the test set 46

2.13 An example of a story generated by the proposed pipeline along with the predicted outputs of the video-understanding tasks on the generated story. The generated story captures information across scenes, characters, event sequences, dialogues, emotions, and the environment. This helps the downstream models to get adequate information about the video to reason about it correctly. The original video can be watched at `https://youtu.be/_amwPjAcoC8`. 52

2.14 The overview of our framework to generate a story from a video and perform downstream video-understanding tasks. First, we sample keyframes from the video which are verbalized using BLIP-2. We also extract OCR from all the frames. Next, using the channel name and ID, we query Wikidata to get company and product information. Next, we obtain automatically generated captions from Youtube videos using the Youtube API. All of these are concatenated as a single prompt and given as input to an LLM and ask it to generate the story of the advertisement. Using the generated story, we then perform the downstream tasks of emotion and topic classification and persuasion strategy identification. This video can be watched at `https://youtu.be/ZBLkTALi1CI`. 54

2.15 Two step process to generate universal adversarial triggers. First, we generate multiple class impressions for each class c. For this, we take multiple initialization sequences differing in starting word and length. After generating class impressions, we use them as our dataset for generating universal adversarial triggers. 67

2.16 Class Impression Generation (CIG) Algorithm. We start with an initial sequence of *"the the ... the"* and continuously update it based on its gradient with respect to output probabilities (Eq. 2.4). The final sequence we get represents the class impression CI^c for the class c. 69

2.17 Iterative Universal Trigger Generation (UTG) Algorithm. . . . 79

2.18 Mean Entropy of class impression words and 350 words randomly selected from the SST, SNLI, and MRPC dataset vocabularies. 80

2.19 t-SNE plots for SST, SNLI, and MRPC class impression words. The words from the different class impressions form distinct clusters based on their classes for all three datasets. The clusters are shown in different colors based on their classes. 82

2.20 Attack success rate as a function of trigger length 83

3.1 Communication process can be defined by seven factors: Communicator, Message, Time of message, Channel, Receiver, Time of effect, and Effect. Any message is created to serve an end goal. For marketers, the end goal is to bring in the desired receiver effect (behavior) (like clicks, purchases, likes, and customer retention). The figure presents the key elements in the communication pipeline - the marketer, message, channel, receivers, and finally, the receiver effect. 89

3.2 Encoding and predicting content (images, videos, and text) and behavior in the language space. Large Content Behavior Models (LCBMs), once trained, can enable a host of different applications, including behavior simulation, content understanding, content-behavior optimization, and content-behavior understanding. 91

3.3 Comparison of GPT-3.5, GPT-4, Vicuna-13B, and LCBM-13B on: (a) Behavior Simulation accuracy on two types of behaviors: replay value prediction and likes/views prediction. The task is, given the video content and channel information, to predict replay values corresponding to each scene and the ratio of likes to views. (b) Content simulation and behavior understanding tasks. The task for content simulation is, given the channel information and scene-level behavior, to predict the scene content. Given information on the video platform and the video content, the task of behavior understanding is to predict and explain the sentiments of the viewers and the commenters. Six evaluators scored the models' explanations between 0-5 to get the predicted sentiment and explanation scores by comparing the ratings and reasons with the user comments. The annotators did not know which model gave the reasoning. (c) Content understanding tasks. We evaluate four tasks: emotion, topic, and persuasion strategy prediction, and action-and-reason understanding. (d) Behavior Simulation on in-house Email Marketing Data (R^2 score) and Twitter likes (accuracy), and Content Simulation on Twitter tweet prediction (BLEU/ROUGE scores). It can be noted that on the behavior simulation, content simulation, and behavior understanding tasks, LCBM performs better than 3-shot GPT-3.5 and 10-shot GPT-4 (covering a larger area. On the content understanding tasks, while LCBM outperforms similar-sized Vicuna models, GPT-3.5 performs better. However, we also note that GPT-3.5 and GPT-4 are at least 12 times larger than LCBM-13B. Further, we show the behavior domain adaptation results in Table 3.9, 3.3, 3.4. 91

3.4 Five factors of communication: Communicator, Message, Channel, Receiver, and Effect. 93

3.5 A few examples showing LCBM's ability to understand and explain human behavior of scene replayability. We compare it against human-provided explanations of the same. 95

3.6 A few examples showing LCBM's ability to understand and explain human behavior of audience sentiment. We also compare it against other models like Vicuna and GPT-3.5. 96

3.7 Encoding and predicting content (images, videos, and text) and behavior in the language space. Strategy to behavior instruction finetune (BFT) LLMs to create LCBMs. We capture visual concepts through the visual encoder (EVA-CLIP), and world knowledge is through an LLM (Llama). To leverage the rich knowledge of LLMs, we use GMHRA and QFormer to convert visual tokens of ViT to language tokens that Llama can understand. Further, we find that verbalizing the visual stimulus helps Llama to gather information more explicitly than what is provided by ViT+QFormer. We finetune the combined model end-to-end to predict 1) behavior given content and 2) content given behavior. Snowflake and fire symbols denote the frozen and unfrozen parts of the architecture. 101

3.8 The in-house marketing emails used in the Email dataset look similar to the ones shown here. 118

4.1 Generated scanpaths over text samples taken from various natural language processing (NLP) tasks. The green circles denote the important words characteristic of that task. The circles' size denotes the fixation duration, and the arrows depict the saccadic movements. As can be seen, linguistically important words often have a higher fixation duration and revisit. Regressions (word revisits) also appear in the examples. 128

4.2 (Intent-aware) Scanpath samples generated by conditioning scanpath generation on different downstream natural language tasks. Note that the conditioned scanpaths are heavily biased to words important for that downstream task. 130

4.3 The architecture of the proposed **ScanTextGAN** model. The model consists of a conditional generator and a discriminator playing a zero-sum game. The generator is trained by two cognitively inspired losses: text content reconstruction and scanpath content reconstruction. 133

4.4 Comparison of *real* and *synthesized* scanpaths corresponding to a few text samples. The proposed ScanTextGAN model generates the latter. 141

4.5 The architecture of the proposed Intent-Aware **ScanTextGAN** model. The model consists of a conditional generator and a discriminator playing a zero-sum game. Two cognitively inspired losses train the generator: scanpath (Task-1) and text (Task-2) reconstruction, a loss from the downstream intent of the natural language task (Task-3), and finally, the loss from the adversarial zero-sum game (Task-4). Variations of scanpaths are generated based on the downstream natural language task. 148

4.6 Saliency samples generated by conditioning scanpath generation on different downstream natural language tasks. It can be observed that the conditioned saliency pays much more attention to words important for that downstream task. 149

4.7 The diagram depicts the five factors of communication in the context of an example YouTube video https://www.youtube.com/watch?v=eT8hO4e2iTM and where lies the free lunch. The receiver effect is not used while training Large Vision and Language Models. However, it contains many important signals that can help in understanding the content. The figure shows several comments containing temporal, cognitive, character, context, and user opinion information useful for understanding the video. 151

4.8 Behavior-LLaVA is trained to answer behavioral questions like simulating user comments and likes on the video. The model, once trained, shows superior performance than LLaMA-Vid and other VLMs on content-related tasks like emotion recognition, action recognition, question answering, persuasion strategy classification, *etc.* The original video was showcased in SuperBowl-2024 and is posted on YouTube on the URL https://www.youtube.com/watch?v=OU7BJc96lI4. The video is titled "Perfect 10: The Kia big game commercial featuring the 2024 Kia EV9" by Kia America. 152

4.9 Behaviour-LLava achieves much higher zero-shot performance compared to Ad-LLaVA and the base model LLaMA-VID across a diverse suite of image, video, and audio benchmarks. 154

4.10 Behavior Instruction fine-tuning template for the video: https://www.youtube.com/watch?v=BKPQkjRF4yY 158

4.11 Percentage performance improvement over an untrained LLaMA-Vid model, compared across various sampling ratios at different checkpoints. The 1:1 sampling ratio shows the best empirical performance. Performance is averaged over 0-shot accuracy improvements on six tasks with 250 samples each from the evaluation set. These tasks include image emotion recognition, video emotion recognition, and persuasion strategy classification, MSRVTT, HVU and MSVD-QA. The figure also shows the benefit of our data filtering process. While training on unfiltered BLIFT also improves the result over baseline performance of Llama-Vid but data filtering adds more improvement on top of it. 159

4.12 Dense caption generated by Behavior-LLaVA for an artistic drawing of a tree. 173

4.13 Dense caption generated by Behavior-LLaVA for a Nike ad. The red-colored text highlights the most important aspects of the video captured by Behavior-LLaVA, demonstrating an understanding of aesthetics, characters, world knowledge, emotion, and spatial relationships. 174

4.14 Dense caption generated by Behavior-LLaVA for a painting of a soldier. The model captures many qualitative aspects that are usually missed in common captioning tasks. 174

4.15 Dense caption generated by Behavior-LLaVA for the video of a Volkswagen ad. The original video is posted at URL: `https://www.youtube.com/watch?v=kyuGXPNr-T0`. The red-colored text highlights the most important aspects of the video captured by Behavior-LLaVA, demonstrating an understanding of aesthetics, characters, world knowledge, emotion, and spatial relationships. More such examples are given in Figs. 4.12, 4.13, 4.14, and Figs. 4.16, 4.17 for images and videos respectively. 175

4.16 Dense caption generated by Behavior-LLaVA for the video of the official trailer of the game Red Dead Redemption 2. The original video is posted at URL: https://www.youtube.com/watch?v=eaW0tYpxyp0. 176

4.17 Dense caption generated by Behavior-LLaVA for the video of Argentina vs France FIFA World Cup Qatar 2022 Highlights. The original video is posted at URL: https://www.youtube.com/watch?v=zhEWqfP6V_w. 177

5.1 Correlations between *content factors* (a-d), *interaction factors* (e-g), and *customer behavior factors* (h-j) with memorability on LAMBDA samples. While emotion has a high correlation with memory, other content factors do not have much correlation. Further, while there is little correlation between the order of videos seen and memorability; with time, participants' memory of the videos shows a forgetting trend. Video popularity, as measured by YouTube likes/views, shows a slight positive correlation with memory. Average brand relevance has a strong positive correlation with memory, with top sectors being remembered as food, entertainment, and tech. Speech, silence and music have little effect with silence having the highest positive correlation with recall. Silence ratio is measured as the percentage of silence in a video, similarly for music and speech. 193

5.2 Predicting memorability by encoding visual information (via visual encoder EVA-CLIP), cognitive concepts (via verbalization module), and world knowledge (through fine-tuned Llama). We instruction fine-tune the combined model end to end to predict user memorability. Snowflake and fire symbols denote the frozen and unfrozen parts of the architecture. 195

5.3 Overview of our SEED method for memorable ad generation. Our self-alignment consists of three steps: (i) **Self-instruction creation**: We first collect 5 million high-quality ads from YouTube, Facebook, and other mediums. Henry (trained on the complete train+test sets of LAMBDA) is then used to rate this curated set in an LLM-as-a-Judge fashion. (ii) **Self-curation**: We select marketing-like and high-memorability samples from the UltraLAMBDA and LAMBDA datasets. (iii) Instruction fine-tuning: Henry-SEED is trained on the self-curated set using two tasks: Behavior Simulation and Content Simulation. 203

5.4 Henry-SEED Prompt: *Generate the detailed description of a 30-second memorable advertisement titled "Brainly Keep Learning 30sec Final 16x9" for the brand Brainly.* Link to the original ad: `https://www.youtube.com/watch?v=kytRXyWXivU` Original Memorability score: 85. Memorability score of Generated Ad: 99. 210

5.5 Henry-SEED Prompt: *Generate the detailed description of a 50 second memorable advertisement titled "Shining a Light on Women's Rights | The Truth Has a Voice | The New York Times" for the brand The New York Times* Link to the original ad: `https://www.youtube.com/watch?v=bPblzhUzTeg` Original memorability score: 65. Memorability score of Generated Ad: 91. 213

5.6 Henry-SEED Prompt: *Generate the detailed description of a 18 second memorable advertisement titled "Maytag Overnight Wash and Dry" for the brand Costco.* Link to the original ad: `https://www.youtube.com/watch?v=uT72lJhUUS0` Original memorability score: 76. Memorability score of the generated Ad: 83 214

5.7 Word Cloud (resembling Henry) for the GPT-4 reasoning on the 75/88 generations where it rates Henry-SEED Generated Ads to be better than the Original. 222

5.8 Airbnb advertisement showing the visual concepts of two adults, and the text “Our guest room is paying for our wedding”. “World knowledge” captured by LLMs helps identify the two adults as partners, and helps relate the text with the two adults and the Airbnb logo to infer what the ad is talking about. 224

5.9 The top three rows show the keyframes from videos in our dataset, LAMBDA, arranged from most to least memorable. The bottom two rows show brands arranged from the most memorable brands to the least. 224

5.10 The graph shows the relationship of the year the ad is uploaded on youtube vs the recall. 226

5.11 Graphs showing the importance of the amount of synthetic data on (a) Ad memorability score and (b) Ad quality for the generated ads. As we can see from the graphs, both the ad memorability and quality increase with the increase in the amount of synthetic data. 227

5.12 The study protocol we followed for our long term memorability human study. All the previous works follow a game-like annotation protocol, where the study participants compete with each other to get best memorability scores and a participant is excluded from the study if their annotations fall below a certain threshold. We follow a more natural way in which participants fill an initial questionnaire, then watch 10 ads with attention checks on day 1 and in subsequent days, receive a form asking them to fill in what do they remember seeing. Further, using Stable Diffusion, we also ask them to recreate the scenes they remember. 233

5.13 Some images from the EngagingImageNet dataset. We constructed pairs of similar images posted within a 45 days interval by the same account. In each pair shown in the figure, the left image corresponds to lower likes and the right one received higher likes. However, existing image generation metrics like Aesthetics, PickScore, Human Preference Score, ImageReward, *etc.*, exhibit image preference in the opposite direction as actual user engagement. . . . 242

5.14 Figure illustrating the steps involved in the creation of the EngagingImageNet dataset. 245

5.15 Visual Instruction Finetuning of EngageNet on EngagingImageNet dataset. The EngageNet model is trained to predict the KPI of an image on a 0-100 scale, conditioned on marketer provided metadata comprising the company, image resolution, image colours and tones with their spatial coverage, marketer's intended image description and tags, and the date of releasing the image on social media. . 251

5.16 Comparison of generated images - EOIG-SD *vs* Base stable diffusion. Engagement optimization helps the model to learn to generate persuasion skills. EOIG-SD generates better product photography (a,c,d), model photography (b), generates images with social appeal and social identity (a,c), and learns temporal patterns (e) (prominently Christmas-themed image of dog) 256

5.17 Rankings and Elo ratings of various text-to-image models in the proposed Image Engagement Arena. 260

5.18 Aesthetic Score distribution across High and Low KPI images in EngagingImageNet dataset 265

5.19 Plots showing variation of number of tweets and likes with time for a few companies in the EngagingImageNet dataset 270

5.20 Reward curves for the performance alignment of stable diffusion on EngagingImageNet (train (a) and validation (b) sets) 276

5.21 Any message is created to serve an end goal. For marketers, the end goal is to bring in the desired receiver effect (behavior) (like clicks, purchases, likes, and customer retention). The figure presents the key elements in the communication pipeline - the marketer, message, channel, receivers, and finally, the receiver effect. Traditionally, image generation is optimized on metrics such as aesthetics and FID. For effective communication, the image generation process needs to be optimized on the receiver effect (other than the traditional metrics). 279

5.22 Sample media and tweets from enterprise accounts in the EngagingImageNet dataset. It can be noted, for example, in the Adobe Photoshop tweets, that the media does not differ significantly in aesthetics or objects themselves (all of them are cats). Despite that, there is much difference in the image KPIs, indicating that viewer engagement is distinct from other optimization objectives such as aesthetics or prompt adherence. 280

5.23 Retrieval framework for conditioning text-to-image models on higher engagement prompts as described in Section 5.2.4.1. The retrieved prompts may incorporate image characteristics that have been empirically shown to improve image engagement. 281

5.24 Illustration depicting supervised finetuning of stable diffusion model on high-liked images from EngagingImageNet dataset as described in Section 5.2.4.2. This method of finetuning U-Net module on preferred data distribution results in generating more engaging images. 281

5.25 Aligning Stable Diffusion for higher engagement using DDPO algorithm (Black *et al.*, 2023) using Engagement Simulation (ES) as the reward function. Architecture of the proposed pipeline for training stable diffusion for the objective of engagement-optimised image generation (EOIG) using Engagement Simulation (ES) reward function as described in Section 5.2.4.3. EngageNet predicts the engagement level of images generated by stable diffusion. The scalar rewards are used to guide stable diffusion to produce progressively higher engagement images. The resulting diffusion model is called EOIG-SD (RLHF-ES). 282

5.26 Aligning Stable Diffusion for higher engagement using DDPO algorithm (Black *et al.*, 2023) using Design Specification Generation (DSG) as the reward function. The architecture of the proposed pipeline for training stable diffusion for the objective of engagement-optimized image generation (EOIG) using Design Specification Generation (DSG) reward function as described in Section 5.2.4.3 and 5.2.7.7. EngageNet trained in this manner possesses the capability to generate verbal descriptions comprising colors, tones, objects, and their locations of an image based on conditioning factors such as the company, time, image caption and viewer likes. We leverage this EngageNet as a reward model to train stable diffusion such that the images generated by it have a design specification aligned with those of higher engagement image. EOIG-SD takes a prompt and generates an image, which then undergoes verbalization via image perception models. Its objective is to create images that, when verbalized, closely resemble the engagement-conditioned verbalization generated by EngageNet. The verbalized output of EOIG-SD is fed into the reward model. We ask EngageNet to predict the logits for this image verbalization, using which a reward is computed for EOIG-SD, indicating how closely this verbalized output aligns with EngageNet. This reward value serves as feedback for EOIG-SD in the form of policy gradient, aiding in its continual improvement and refinement within the image generation process. Thus, this pipeline trains EOIG-SD to generate engagement-optimized images by gradually aligning its output with EngageNet. 283

5.27 Win Rates of different models against each other in the Image Engagement Arena 284

Bibliography

1. (). List of dirty, naughty, obscene, and otherwise bad words. https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words. 13th April, 2024.

2. (2024*a*). Reddit pics: Rules. https://www.reddit.com/r/pics/wiki/rule8/. Accessed: April 11, 2024.

3. (2024*b*). Reddit pics: Wiki index. https://new.reddit.com/r/pics/wiki/index/. Accessed: April 11, 2024.

4. (2024*c*). Reddit videos: Rules. https://www.reddit.com/r/videos/wiki/rules. Accessed: April 11, 2024.

5. **Adobe** (2024). Adobe Firefly. https://www.adobe.com/products/firefly.html. Accessed: February 9, 2024.

6. **Akagunduz, E.**, **A. G. Bors**, and **K. K. Evans** (2019). Defining image memorability using the visual memory schema. *IEEE transactions on pattern analysis and machine intelligence*, **42**(9), 2165–2178.

7. **Alalwan, A. A.**, **N. P. Rana**, **Y. K. Dwivedi**, and **R. Algharabat** (2017). Social media in marketing: A review and analysis of the existing literature. *Telematics and informatics*, **34**(7), 1177–1190.

8. **Althoff, T.**, **C. Danescu-Niculescu-Mizil**, and **D. Jurafsky**, How to ask for a favor: A case study on the success of altruistic requests. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 8. 2014.

9. **Anand, P.**, **J. King**, **J. Boyd-Graber**, **E. Wagner**, **C. Martell**, **D. Oard**, and **P. Resnik**, Believe me—we can do this! annotating persuasive acts in blog text. *In Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. 2011.

10. **Arev, I.**, **H. S. Park**, **Y. Sheikh**, **J. Hodgins**, and **A. Shamir** (2014). Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics (TOG)*, **33**(4), 1–11.

11. **Arevalo, J.**, **T. Solorio**, **M. M. y Gómez**, and **F. A. González** (2017). Gated multimodal units for information fusion.

12. **Argyle, L. P.**, **E. C. Busby**, **N. Fulda**, **J. R. Gubler**, **C. Rytting**, and **D. Wingate** (2023*a*). Out of one, many: Using language models to simulate human samples. *Political Analysis*, **31**(3), 337–351. ISSN 1476-4989. URL http://dx.doi.org/10.1017/pan.2023.2.

13. **Argyle, L. P.**, **E. C. Busby**, **N. Fulda**, **J. R. Gubler**, **C. Rytting**, and **D. Wingate** (2023*b*). Out of one, many: Using language models to simulate human samples. *Political Analysis*, **31**(3), 337–351.

14. **Ariely, D.**, **G. Loewenstein**, and **D. Prelec** (2003). "coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly journal of economics*, **118**(1), 73–106.

15. **Aronson, E.**, **J. A. Turner**, and **J. M. Carlsmith** (1963). Communicator credibility and communication discrepancy as determinants of opinion change. *The Journal of Abnormal and Social Psychology*, **67**(1), 31.

16. **Asch, S. E.** (1948). The doctrine of suggestion, prestige and imitation in social psychology. *Psychological review*, **55**(5), 250.

17. **Assens, M.**, **X. Giro-i Nieto**, **K. McGuinness**, and **N. E. O'Connor**, Pathgan: Visual scanpath prediction with generative adversarial networks. *In* **L. Leal-Taixé** and **S. Roth** (eds.), *Computer Vision – ECCV 2018 Workshops*. Springer International Publishing, 2019.

18. **Asur, S.** and **B. A. Huberman**, Predicting the future with social media. *In 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, volume 1. IEEE, 2010.

19. **Atalay, A. S.**, **S. E. Kihal**, and **F. Ellsaesser** (2023). Creating effective marketing messages through moderately surprising syntax. *Journal of Marketing*, 00222429231153582.

20. **Atkinson, R. C.** and **R. M. Shiffrin**, Human memory: A proposed system and its control processes. *In Psychology of learning and motivation*, volume 2. Elsevier, 1968, 89–195.

21. **Bai, C.**, **H. Chen**, **S. Kumar**, **J. Leskovec**, and **V. Subrahmanian** (2021). M2p2: Multimodal persuasion prediction using adaptive fusion. *IEEE Transactions on Multimedia*.

22. **Bai, Y.**, **A. Jones**, **K. Ndousse**, **A. Askell**, **A. Chen**, **N. DasSarma**, **D. Drain**, **S. Fort**, **D. Ganguli**, **T. Henighan**, *et al.* (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

23. **Bain, M.**, **A. Nagrani**, **G. Varol**, and **A. Zisserman**, Frozen in time: A joint video and image encoder for end-to-end retrieval. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

24. **Bankes, S. C.** (2002). Agent-based modeling: A revolution? *Proceedings of the National Academy of Sciences*, **99**(suppl_3), 7199–7200.

25. **Bargh, J. A.**, **M. Chen**, and **L. Burrows** (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of personality and social psychology*, **71**(2), 230.

26. **Barrett, M.**, **J. Bingel**, **N. Hollenstein**, **M. Rei**, and **A. Søgaard**, Sequence classification with human attention. *In CNLL*. Association for Computational Linguistics, 2018*a*. URL https://aclanthology.org/K18-1030.

27. **Barrett, M.**, **J. Bingel**, **F. Keller**, and **A. Søgaard**, Weakly supervised part-of-speech tagging using eye-tracking data. *In ACL (Volume 2: Short Papers)*. ACL, 2016*a*. URL https://aclanthology.org/P16-2094.

28. **Barrett, M.**, **A. V. González-Garduño**, **L. Frermann**, and **A. Søgaard**, Unsupervised induction of linguistic categories with records of reading, speaking, and writing. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2018*b*. URL https://aclanthology.org/N18-1184.

29. **Barrett, M.**, **F. Keller**, and **A. Søgaard**, Cross-lingual transfer of correlations between parts of speech and gaze features. *In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 2016*b*. URL https://aclanthology.org/C16-1126.

30. **Behjati, M.**, **S. Moosavi-Dezfooli**, **M. S. Baghshah**, and **P. Frossard**, Universal adversarial attacks on text classifiers. *In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2019. URL https://doi.org/10.1109/ICASSP.2019.8682430.

31. **Bellman, S.**, **S. Arismendez**, and **D. Varan** (2021). Can muted video advertising be as effective as video advertising with sound? *SN Business & Economics*, **1**(1), 27.

32. **Bertenthal, B. I.** (1996). Origins and early development of perception, action, and representation. *Annual review of psychology*, **47**(1), 431–459.

33. **Berzak, Y.**, **C. Nakamura**, **A. Smith**, **E. Weng**, **B. Katz**, **S. Flynn**, and **R. Levy** (2022). Celer: A 365-participant corpus of eye movements in l1 and l2 english reading. *Open Mind*.

34. **Bhattacharya, C. B.** and **S. Sen** (2003). Consumer–company identification: A framework for understanding consumers' relationships with companies. *Journal of marketing*, **67**(2), 76–88.

35. **Bhattacharyya, A.**, **S. Agrawal**, **Y. K. Singla**, **S. Nikitha**, **T. R. Menta**, and **B. Krishnamurthy** (2024). Socia: Training large language models to simulate social constructs and interventions. *arXiv*.

36. **Bhattacharyya, A.**, **Y. K. Singla**, **B. Krishnamurthy**, **R. R. Shah**, and **C. Chen**, A video is worth 4096 tokens: Verbalize videos to understand them in zero shot. *In* **H. Bouamor**, **J. Pino**, and **K. Bali** (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 2023. URL https://aclanthology.org/2023.emnlp-main.608.

37. **Bicknell, K.** and **R. Levy**, Why readers regress to previous words: A statistical analysis. *In Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33. 2011.

38. **Biderman, S.**, **K. Bicheno**, and **L. Gao** (2022). Datasheet for the pile. *arXiv preprint arXiv:2201.07311*.

39. **Biderman, S.**, **H. Schoelkopf**, **Q. G. Anthony**, **H. Bradley**, **K. O'Brien**, **E. Hallahan**, **M. A. Khan**, **S. Purohit**, **U. S. Prashanth**, **E. Raff**, *et al.*, Pythia: A suite for analyzing large language models across training and scaling. *In International Conference on Machine Learning*. PMLR, 2023.

40. **Biderman, S.**, **H. Schoelkopf**, **L. Sutawika**, **L. Gao**, **J. Tow**, **B. Abbasi**, **A. F. Aji**, **P. S. Ammanamanchi**, **S. Black**, **J. Clive**, *et al.* (2024). Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

41. **Black, K.**, **M. Janner**, **Y. Du**, **I. Kostrikov**, and **S. Levine** (2023). Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

42. **Blohm, M.**, **G. Jagfeld**, **E. Sood**, **X. Yu**, and **N. T. Vu**, Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension. *In CNLL*. ACL, 2018.

43. **Bobadilla, J.**, **F. Ortega**, **A. Hernando**, and **A. Gutiérrez** (2013). Recommender systems survey. *Knowledge-based systems*, **46**, 109–132.

44. **Boerman, S. C.**, **S. Kruikemeier**, and **F. J. Zuiderveen Borgesius** (2017). Online behavioral advertising: A literature review and research agenda. *Journal of advertising*, **46**(3), 363–376.

45. **Bowman, S. R.**, **G. Angeli**, **C. Potts**, and **C. D. Manning**, A large annotated corpus for learning natural language inference. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2015. URL `https://www.aclweb.org/anthology/D15-1075`.

46. **Boyd, A.**, **K. W. Bowyer**, and **A. Czajka**, Human-aided saliency maps improve generalization of deep learning. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022.

47. **Breakthrough** (2023). Pyscenedetect: Video scene cut detection and analysis tool. URL `https://github.com/Breakthrough/PySceneDetect`.

48. **Brehm, J. W.** (1966). A theory of psychological reactance.

49. **Breiman, L.** (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, **16**(3), 199–231.

50. **Briñol, P.** and **R. E. Petty**, A history of attitudes and persuasion research. *In Handbook of the history of social psychology*. Psychology Press, 2012, 283–320.

51. **Brockmann, D.**, **L. Hufnagel**, and **T. Geisel** (2006). The scaling laws of human travel. *Nature*, **439**(7075), 462–465.

52. **Brown, P. F.**, **S. A. Della Pietra**, **V. J. Della Pietra**, **J. C. Lai**, and **R. L. Mercer** (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics*, **18**(1), 31–40.

53. **Brown, T. B.**, **B. Mann**, **N. Ryder**, **M. Subbiah**, **J. Kaplan**, **P. Dhariwal**, **A. Neelakantan**, **P. Shyam**, **G. Sastry**, **A. Askell**, **S. Agarwal**, **A. Herbert-Voss**, **G. Krueger**, **T. Henighan**, **R. Child**, **A. Ramesh**, **D. M. Ziegler**, **J. Wu**, **C. Winter**, **C. Hesse**, **M. Chen**, **E. Sigler**, **M. Litwin**, **S. Gray**, **B. Chess**, **J. Clark**, **C. Berner**, **S. McCandlish**, **A. Radford**, **I. Sutskever**, and **D. Amodei** (2020). Language models are few-shot learners.

54. **Burger, J. M.** (1999). The foot-in-the-door compliance procedure: A multiple-process analysis and review. *Personality and social psychology review*, **3**(4), 303–325.

55. **Bylinskii, Z.**, **P. Isola**, **C. Bainbridge**, **A. Torralba**, and **A. Oliva** (2015). Intrinsic and extrinsic effects on image memorability. *Vision research*, **116**, 165–178.

56. **Caba Heilbron, F.**, **V. Escorcia**, **B. Ghanem**, and **J. Carlos Niebles**, Activitynet: A large-scale video benchmark for human activity understanding. *In Proceedings of the ieee conference on computer vision and pattern recognition*. 2015.

57. **Cambria, E.**, **B. Schuller**, **Y. Xia**, and **C. Havasi** (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, **28**(2), 15–21.

58. **Carrière-Swallow, Y.** and **F. Labbé** (2013). Nowcasting with google trends in an emerging market. *Journal of Forecasting*, **32**(4), 289–298.

59. **Chaiken, S.** (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, **39**(5).

60. **Chandra, S.**, **S. Verma**, **W. M. Lim**, **S. Kumar**, and **N. Donthu** (2022). Personalization in personalized marketing: Trends and ways forward. *Psychology & Marketing*, **39**(8), 1529–1562.

61. **Changpinyo, S.**, **P. Sharma**, **N. Ding**, and **R. Soricut**, Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

62. **Chen, D.** and **W. B. Dolan**, Collecting highly parallel data for paraphrase evaluation. *In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 2011.

63. **Chen, J.**, **C. Ge**, **E. Xie**, **Y. Wu**, **L. Yao**, **X. Ren**, **Z. Wang**, **P. Luo**, **H. Lu**, and **Z. Li** (2024*a*). Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*.

64. **Chen, J.** and **D. Yang** (2021*a*). Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(14), 12648–12656. URL https://ojs.aaai.org/index.php/AAAI/article/view/17498.

65. **Chen, J.** and **D. Yang**, Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35. 2021*b*.

66. **Chen, J.**, **J. YU**, **C. GE**, **L. Yao**, **E. Xie**, **Z. Wang**, **J. Kwok**, **P. Luo**, **H. Lu**, and **Z. Li**, Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *In The Twelfth International Conference on Learning Representations*. 2024*b*. URL https://openreview.net/forum?id=eAKmQPe3m1.

67. **Chen, L.**, **K. Lu**, **A. Rajeswaran**, **K. Lee**, **A. Grover**, **M. Laskin**, **P. Abbeel**, **A. Srinivas**, and **I. Mordatch** (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, **34**, 15084–15097.

68. **Chen, Q.**, **X. Zhu**, **Z.-H. Ling**, **S. Wei**, **H. Jiang**, and **D. Inkpen**, Enhanced LSTM for natural language inference. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 2017. URL https://www.aclweb.org/anthology/P17-1152.

69. **Chen, S.**, **X. He**, **L. Guo**, **X. Zhu**, **W. Wang**, **J. Tang**, and **J. Liu** (2023). Valor: Vision-audio-language omni-perception pretraining model and dataset. URL https://arxiv.org/abs/2304.08345.

70. **Chen, T.**, **B. Xu**, **C. Zhang**, and **C. Guestrin** (2016). Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

71. **Chen, X.**, **H. Fang**, **T.-Y. Lin**, **R. Vedantam**, **S. Gupta**, **P. Dollár**, and **C. L. Zitnick** (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

72. **Chen, Z.** and **W. Sun**, Scanpath prediction for visual attention using ior-roi lstm. *In IJCAI*. 2018.

73. **Cheng, J.**, **L. Adamic**, **P. A. Dow**, **J. M. Kleinberg**, and **J. Leskovec**, Can cascades be predicted? *In Proceedings of the 23rd international conference on World wide web*. 2014*a*.

74. **Cheng, M.-M.**, **N. J. Mitra**, **X. Huang**, **P. H. Torr**, and **S.-M. Hu** (2014*b*). Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, **37**(3), 569–582.

75. **Chiang, W.-L.**, **Z. Li**, **Z. Lin**, **Y. Sheng**, **Z. Wu**, **H. Zhang**, **L. Zheng**, **S. Zhuang**, **Y. Zhuang**, **J. E. Gonzalez**, **I. Stoica**, and **E. P. Xing** (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. URL https://lmsys.org/blog/2023-03-30-vicuna/.

76. **Chiang, W.-L.**, **L. Zheng**, **Y. Sheng**, **A. N. Angelopoulos**, **T. Li**, **D. Li**, **H. Zhang**, **B. Zhu**, **M. Jordan**, **J. E. Gonzalez**, and **I. Stoica** (2024). Chatbot arena: An open platform for evaluating llms by human preference. URL https://arxiv.org/abs/2403.04132.

77. **Choi, H.** and **H. Varian** (2012). Predicting the present with google trends. *Economic record*, **88**, 2–9.

78. **Chopard, B.** and **M. Droz** (1998). Cellular automata. *Modelling of Physical*, **1**.

79. **Christiano, P. F.**, **J. Leike**, **T. Brown**, **M. Martic**, **S. Legg**, and **D. Amodei** (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, **30**.

80. **Chu, E.**, **J. Andreas**, **S. Ansolabehere**, and **D. Roy** (2023). Language models trained on media diets can predict public opinion. *arXiv preprint arXiv:2303.16779*.

81. **Chung, H. W.**, **L. Hou**, **S. Longpre**, **B. Zoph**, **Y. Tay**, **W. Fedus**, **Y. Li**, **X. Wang**, **M. Dehghani**, **S. Brahma**, **A. Webson**, **S. S. Gu**, **Z. Dai**, **M. Suzgun**, **X. Chen**, **A. Chowdhery**, **A. Castro-Ros**, **M. Pellat**, **K. Robinson**, **D. Valter**, **S. Narang**, **G. Mishra**, **A. Yu**, **V. Zhao**, **Y. Huang**, **A. Dai**, **H. Yu**, **S. Petrov**, **E. H. Chi**, **J. Dean**, **J. Devlin**, **A. Roberts**, **D. Zhou**, **Q. V. Le**, and **J. Wei** (2022). Scaling instruction-finetuned language models.

82. **Cialdini, R. B.** and **R. B. Cialdini**, *Influence: The psychology of persuasion*, volume 55. Collins New York, 2007.

83. **Cialdini, R. B.** and **N. J. Goldstein** (2004). Social influence: Compliance and conformity. *Annual review of psychology*, **55**(1), 591–621.

84. **Cjwbw** (2022). Waifudiffusion. `https://replicate.com/cjwbw/waifu-diffusion`.

85. **Clark, M. S.** (1984). Record keeping in two types of relationships. *Journal of personality and social psychology*, **47**(3).

86. **Clark, M. S.** and **J. Mills** (1979). Interpersonal attraction in exchange and communal relationships. *Journal of personality and social psychology*, **37**(1).

87. **Clark, M. S.**, **J. Mills**, and **M. C. Powell** (1986). Keeping track of needs in communal and exchange relationships. *Journal of personality and social psychology*, **51**(2), 333.

88. **Clifton Jr, C.**, **A. Staub**, and **K. Rayner** (2007). Eye movements in reading words and sentences. *Eye movements*.

89. **Cohen, J.** (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

90. **Cohendet, R.**, **C.-H. Demarty**, **N. Q. Duong**, and **M. Engilberge**, Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

91. **Collaboration, O. S.** (2015). Estimating the reproducibility of psychological science. *Science*, **349**(6251), aac4716.

92. **Collaborative, T. F.** (2023). Insights into the accuracy of social scientists' forecasts of societal change. *Nature human behaviour*, **7**(4), 484–501.

93. **Cornia, M.**, **L. Baraldi**, **G. Serra**, and **R. Cucchiara** (2018). Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, **14**(2), 1–21.

94. **Cristino, F.**, **S. Mathôt**, **J. Theeuwes**, and **I. D. Gilchrist** (2010). Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*.

95. **Danescu-Niculescu-Mizil, C.**, **J. Cheng**, **J. Kleinberg**, and **L. Lee**, You had me at hello: How phrasing affects memorability. *In* **H. Li**, **C.-Y. Lin**, **M. Osborne**, **G. G. Lee**, and **J. C. Park** (eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jeju Island, Korea, 2012. URL https://aclanthology.org/P12-1094.

96. **Dao, T.**, **D. Y. Fu**, **S. Ermon**, **A. Rudra**, and **C. Ré**, FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *In Advances in Neural Information Processing Systems*. 2022.

97. **de Belen, R. A. J.**, **T. Bednarz**, and **A. Sowmya**, Scanpathnet: A recurrent mixture density network for scanpath prediction. *In IEEE CVPR*. 2022.

98. **De Montjoye, Y.-A.**, **L. Radaelli**, **V. K. Singh**, and **A. S. Pentland** (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, **347**(6221), 536–539.

99. **De Toni, G.**, **C. Consonni**, and **A. Montresor** (2021). A general method for estimating the prevalence of influenza-like-symptoms with wikipedia data. *Plos one*, **16**(8), e0256858.

100. **Demberg, V.** and **F. Keller** (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*.

101. **Deutsch, M.** and **H. B. Gerard** (1955). A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, **51**(3), 629.

102. **Devlin, J.**, **M.-W. Chang**, **K. Lee**, and **K. Toutanova**, BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2019. URL https://www.aclweb.org/anthology/N19-1423.

103. **Dhakal, V.**, **A. M. Feit**, **P. O. Kristensson**, and **A. Oulasvirta**, Observations on typing from 136 million keystrokes. *In Proceedings of the 2018 CHI conference on human factors in computing systems*. 2018.

104. **Dhariwal, P.** and **A. Nichol** (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, **34**, 8780–8794.

105. **Diba, A.**, **M. Fayyaz**, **V. Sharma**, **M. Paluri**, **J. Gall**, **R. Stiefelhagen**, and **L. Van Gool**, Large scale holistic video understanding. *In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020.

106. **Ding, K.**, **R. Wang**, and **S. Wang**, Social media popularity prediction: A multiple feature fusion approach with deep neural networks. *In Proceedings of the 27th ACM International Conference on Multimedia*. 2019.

107. **Dolan, W. B.** and **C. Brockett**, Automatically constructing a corpus of sentential paraphrases. *In Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. 2005. URL https://www.aclweb.org/anthology/I05-5002.

108. **Donadello, I.**, **M. Dragoni**, and **C. Eccher**, Explaining reasoning algorithms with persuasiveness: a case study for a behavioural change system. *In Proceedings of the 35th Annual ACM Symposium on Applied Computing*. 2020.

109. **Donadello, I.**, **A. Hunter**, **S. Teso**, and **M. Dragoni**, Machine learning for utility prediction in argument-based computational persuasion. *In proceedings of the AAAI conference on artificial intelligence*, volume 36. 2022.

110. **Dosovitskiy, A.**, **L. Beyer**, **A. Kolesnikov**, **D. Weissenborn**, **X. Zhai**, **T. Unterthiner**, **M. Dehghani**, **M. Minderer**, **G. Heigold**, **S. Gelly**, *et al.* (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

111. **Douze, M.**, **A. Guzhva**, **C. Deng**, **J. Johnson**, **G. Szilvasy**, **P.-E. Mazaré**, **M. Lomeli**, **L. Hosseini**, and **H. Jégou** (2024). The faiss library.

112. **Du, Y.**, **C. Li**, **R. Guo**, **X. Yin**, **W. Liu**, **J. Zhou**, **Y. Bai**, **Z. Yu**, **Y. Yang**, **Q. Dang**, *et al.* (2020). Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*.

113. **Dumont, T.**, **J. S. Hevia**, and **C. L. Fosco**, Modular memorability: Tiered representations for video memorability prediction. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.

114. **Durmus, E.** and **C. Cardie**, Exploring the role of prior beliefs for argument persuasion. *In NAACL: Human Language Technologies, Volume 1 (Long Papers)*. ACL, New Orleans, Louisiana, 2018. URL https://aclanthology.org/N18-1094.

115. **Durmus, E.**, **L. Lovitt**, **A. Tamkin**, **S. Ritchie**, **J. Clark**, and **D. Ganguli** (2024). Measuring the persuasiveness of language models. URL https://www.anthropic.com/news/measuring-model-persuasiveness.

116. **Dwork, C.**, **A. Roth**, *et al.* (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, **9**(3–4), 211–407.

117. **Dzedzickis, A.**, **A. Kaklauskas**, and **V. Bucinskas** (2020). Human emotion recognition: Review of sensors and methods. *Sensors*, **20**(3), 592.

118. **Ebbinghaus, H.**, *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot, 1885.

119. **Ebrahimi, J.**, **A. Rao**, **D. Lowd**, and **D. Dou**, HotFlip: White-box adversarial examples for text classification. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2018. URL https://www.aclweb.org/anthology/P18-2006.

120. **Economist, T.** (2023). Why reddit users are protesting against the site's leadership. *The Economist Explains*. URL https://www.economist.com/the-economist-explains/2023/06/26/why-reddit-users-are-protesting-against-the-sites-leadership. Accessed: April 11, 2024.

121. **El Baff, R.**, **H. Wachsmuth**, **K. Al Khatib**, and **B. Stein**, Analyzing the Persuasive Effect of Style in News Editorial Argumentation. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2020. URL https://aclanthology.org/2020.acl-main.287.

122. **Engbert, R.**, **A. Nuthmann**, **E. M. Richter**, and **R. Kliegl** (2005). Swift: a dynamical model of saccade generation during reading. *Psychological review*.

123. **Epley, N.**, **A. Waytz**, and **J. T. Cacioppo** (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, **114**(4), 864.

124. **Esser, P.**, **S. Kulal**, **A. Blattmann**, **R. Entezari**, **J. Müller**, **H. Saini**, **Y. Levi**, **D. Lorenz**, **A. Sauer**, **F. Boesel**, *et al.*, Scaling rectified flow transformers for high-resolution image synthesis. *In Forty-first International Conference on Machine Learning*. 2024.

125. **Everaert, M. N.**, **M. Bocchio**, **S. Arpa**, **S. Süsstrunk**, and **R. Achanta**, Diffusion in style. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

126. **Fajtl, J.**, **V. Argyriou**, **D. Monekosso**, and **P. Remagnino**, Amnet: Memorability estimation with attention. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

127. **Fan, S.**, **Z. Shen**, **M. Jiang**, **B. L. Koenig**, **J. Xu**, **M. S. Kankanhalli**, and **Q. Zhao**, Emotional attention: A study of image sentiment and visual attention. *In Proceedings of the IEEE Conference on computer vision and pattern recognition*. 2018.

128. **Forbes** (2022). Agencies agree 2021 was a record year for ad spending, with more growth expected in 2022. https://www.forbes.com/sites/bradadgate/2021/12/08/agencies-agree-2021-was-a-record-year-for-ad-spending-with-more-growth-expected-in-2022/. Accessed on December 8, 2023.

129. **Fournier, S.** (1998). Consumers and their brands: Developing relationship theory in consumer research. *Journal of consumer research*, **24**(4), 343–373.

130. **Fouts, R. S.**, **M. L. A. Jensvold**, and **D. H. Fouts** (2002). 35 chimpanzee signing: Darwinian realities and cartesian delusions. *The Cognitive Animal*, 285.

131. **Frank, S.**, **E. Bugliarello**, and **D. Elliott** (2021). Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*.

132. **Fredrikson, M.**, **S. Jha**, and **T. Ristenpart**, Model inversion attacks that exploit confidence information and basic countermeasures. *In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 2015.

133. **Freedman, J. L.** and **S. C. Fraser** (1966). Compliance without pressure: the foot-in-the-door technique. *Journal of personality and social psychology*, **4**(2), 195.

134. **Friedman, D.** and **Y. A. Feldman**, Knowledge-based cinematography and its applications. *In ECAI*, volume 16. Citeseer, 2004.

135. **Furnham, A.** and **H. C. Boo** (2011). A literature review of the anchoring effect. *The journal of socio-economics*, **40**(1), 35–42.

136. **Ge, Y.**, **Y. Ge**, **Z. Zeng**, **X. Wang**, and **Y. Shan** (2023). Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*.

137. **Generous, N.**, **G. Fairchild**, **A. Deshpande**, **S. Y. Del Valle**, and **R. Priedhorsky** (2014). Global disease monitoring and forecasting with wikipedia. *PLoS computational biology*, **10**(11), e1003892.

138. **Giannakopoulos, T.** (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, **10**(12).

139. **Giffin, K.** (1967). The contribution of studies of source credibility to a theory of interpersonal trust in the communication process. *Psychological bulletin*, **68**(2), 104.

140. **Gilad-Bachrach, R.**, **A. Navot**, and **N. Tishby** (2005). Query by committee made real. *Advances in neural information processing systems*, **18**.

141. **Gilbert, N.** and **K. Troitzsch**, *Simulation for the social scientist*. McGraw-Hill Education (UK), 2005.

142. **Goetschalckx, L.**, **A. Andonian**, **A. Oliva**, and **P. Isola**, Ganalyze: Toward visual definitions of cognitive image properties. *In Proceedings of the ieee/cvf international conference on computer vision*. 2019.

143. **Goetschalckx, L.** and **J. Wagemans** (2019). Memcat: a new category-based image set quantified on memorability. *PeerJ*, **7**, e8169.

144. **Goodfellow, I.**, **J. Pouget-Abadie**, **M. Mirza**, **B. Xu**, **D. Warde-Farley**, **S. Ozair**, **A. Courville**, and **Y. Bengio** (2020). Generative adversarial networks. *Communications of the ACM*, **63**(11), 139–144.

145. **Google** (2021). The google natural language api. `https://cloud.google.com/natural-language#natural-language-api-demo`. Accessed: 28 March 2022.

146. **Graves, A.** and **J. Schmidhuber** (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, **18**(5-6), 602–610.

147. **Gray, G. W.** (1946). The "precepts of kagemni and ptah-hotep". *Quarterly Journal of Speech*, **32**(4), 446–454.

148. **Griffin, D. R.**, *Animal minds: Beyond cognition to consciousness*. University of Chicago Press, 2001.

149. **Grifoni, P.**, *Multimodal human computer interaction and pervasive services*. IGI Global, 2009.

150. **Gururangan, S.**, **S. Swayamdipta**, **O. Levy**, **R. Schwartz**, **S. Bowman**, and **N. A. Smith**, Annotation artifacts in natural language inference data. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2018.

151. **Habernal, I.** and **I. Gurevych**, What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. *In EMNLP*. 2016.

152. **Hackforth, R.**, *Plato: Phaedrus*. 119. Cambridge University Press, 1972.

153. **Hadoux, E.**, **A. Hunter**, and **S. Polberg** (2021). Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. *arXiv preprint arXiv:2101.11870*.

154. **Hagen, T.** and **T. Espeseth** (2023). Image memorability prediction with vision transformers. *arXiv preprint arXiv:2301.08647*.

155. **Hahn, M.** and **F. Keller**, Modeling human reading with neural attention. *In EMNLP*. Association for Computational Linguistics, Austin, Texas, 2016. URL https://aclanthology.org/D16-1009.

156. **Han, M.**, **X. Chang**, **H. Wang**, and **L. Yang** (2023). Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*.

157. **Hardt, M.**, **E. Price**, and **N. Srebro** (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, **29**.

158. **Hare, B.**, **J. Call**, **B. Agnetta**, and **M. Tomasello** (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, **59**(4), 771–785.

159. **He, H.**, **D. Chen**, **A. Balakrishnan**, and **P. Liang** (2018). Decoupling strategy and generation in negotiation dialogues. *arXiv preprint arXiv:1808.09637*.

160. **He, S.**, **H. R. Tavakoli**, **A. Borji**, and **N. Pugeault**, Human attention in image captioning: Dataset and analysis. *In ICCV*. 2019.

161. **Herlocker, J. L.**, **J. A. Konstan**, **L. G. Terveen**, and **J. T. Riedl** (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, **22**(1), 5–53.

162. **Hern, A.** (2023). Reddit moderator protest spreads across thousands of communities. *The Guardian*. URL https://www.theguardian.com/technology/2023/dec/30/reddit-moderator-protest-communities-social-media. Accessed: April 11, 2024.

163. **Heusel, M.**, **H. Ramsauer**, **T. Unterthiner**, **B. Nessler**, and **S. Hochreiter** (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, **30**.

164. **Hibbert, S.**, **A. Smith**, **A. Davies**, and **F. Ireland** (2007). Guilt appeals: Persuasion knowledge and charitable giving. *Psychology & Marketing*, **24**(8), 723–742.

165. **Hidey, C.**, **E. Musi**, **A. Hwang**, **S. Muresan**, and **K. McKeown**, Analyzing the semantic types of claims and premises in an online persuasive forum. *In Proceedings of the 4th Workshop on Argument Mining*. 2017.

166. **Ho, J.**, **A. Jain**, and **P. Abbeel** (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, **33**, 6840–6851.

167. **Ho, J.** and **T. Salimans** (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

168. **Hofman, J. M.**, **A. Sharma**, and **D. J. Watts** (2017). Prediction and explanation in social systems. *Science*, **355**(6324), 486–488.

169. **Hollenstein, N.**, **M. Barrett**, **M. Troendle**, **F. Bigiolli**, **N. Langer**, and **C. Zhang** (2019). Advancing nlp with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

170. **Hollenstein, N.**, **E. Chersoni**, **C. L. Jacobs**, **Y. Oseki**, **L. Prévot**, and **E. Santus**, CMCL 2021 shared task on eye-tracking prediction. *In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, Online, 2021*a*. URL https://aclanthology.org/2021.cmcl-1.7.

171. **Hollenstein, N.**, **F. Pirovano**, **C. Zhang**, **L. Jäger**, and **L. Beinborn**, Multilingual language models predict human reading behavior. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2021*b*. URL https://aclanthology.org/2021.naacl-main.10.

172. **Hollenstein, N.** and **C. Zhang**, Entity recognition at first sight: Improving NER with eye movement information. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2019*a*. URL https://aclanthology.org/N19-1001.

173. **Hollenstein, N.** and **C. Zhang**, Entity recognition at first sight: Improving NER with eye movement information. *In NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2019*b*.

174. **Hovland, C. I.**, **I. L. Janis**, and **H. H. Kelley** (1953). Communication and persuasion.

175. **Hu, T.** and **N. Collier** (2024). Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

176. **Huan, Z.**, **Y. Wang**, **X. Zhang**, **L. Shang**, **C. Fu**, and **J. Zhou**, Data-free adversarial perturbations for practical black-box attack. *In Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2020.

177. **Hussain, Z.**, **M. Zhang**, **X. Zhang**, **K. Ye**, **C. Thomas**, **Z. Agha**, **N. Ong**, and **A. Kovashka**, Automatic understanding of image and video advertisements. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

178. **Isola, P.**, **J. Xiao**, **D. Parikh**, **A. Torralba**, and **A. Oliva** (2013). What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, **36**(7), 1469–1482.

179. **Isola, P.**, **J. Xiao**, **A. Torralba**, and **A. Oliva**, What makes an image memorable? *In CVPR 2011*. IEEE, 2011.

180. **Iyer, R. R.** and **K. Sycara** (2019). An unsupervised domain-independent framework for automated detection of persuasion tactics in text. *arXiv preprint arXiv:1912.06745*.

181. **Jarodzka, H.**, **K. Holmqvist**, and **M. Nyström**, A vector-based, multidimensional scanpath similarity measure. *In Proceedings of the 2010 symposium on eye-tracking research & applications*. 2010.

182. **Jelinek, F.**, **R. L. Mercer**, **L. R. Bahl**, and **J. K. Baker** (2005). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, **62**(S1), S63–S63. ISSN 0001-4966. URL `https://doi.org/10.1121/1.2016299`.

183. **Jiang, A. Q.**, **A. Sablayrolles**, **A. Mensch**, **C. Bamford**, **D. S. Chaplot**, **D. de las Casas**, **F. Bressand**, **G. Lengyel**, **G. Lample**, **L. Saulnier**, **L. R. Lavaud**, **M.-A. Lachaux**, **P. Stock**, **T. L. Scao**, **T. Lavril**, **T. Wang**, **T. Lacroix**, and **W. E. Sayed** (2023). Mistral 7b.

184. **Jiang, M.**, **X. Boix**, **G. Roig**, **J. Xu**, **L. Van Gool**, and **Q. Zhao** (2016). Learning to predict sequences of human visual fixations. *IEEE transactions on neural networks and learning systems*.

185. **Jiang, M.**, **S. Huang**, **J. Duan**, and **Q. Zhao**, Salicon: Saliency in context. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

186. **Jiang, Y.-G.**, **B. Xu**, and **X. Xue**, Predicting emotions in user-generated videos. *In Proceedings of the AAAI conference on artificial intelligence*, volume 28. 2014.

187. **Johnson, E. J.** and **D. Goldstein** (2003). Do defaults save lives?

188. **Johnson, J.**, **M. Douze**, and **H. Jégou** (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, **7**(3), 535–547.

189. **Joo, J.**, **W. Li**, **F. F. Steen**, and **S.-C. Zhu**, Visual persuasion: Inferring communicative intents of images. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

190. **Joo, J.**, **F. F. Steen**, and **S.-C. Zhu**, Automated facial trait judgment and election outcome prediction: Social dimensions of face. *In Proceedings of the IEEE international conference on computer vision*. 2015.

191. **Joshi, A.**, **V. Sharma**, and **P. Bhattacharyya**, Harnessing context incongruity for sarcasm detection. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, 2015. URL https://aclanthology.org/P15-2124.

192. **Just, M.** and **P. A. Carpenter** (1980). A theory of reading: From eye fixations to comprehension. URL https://kilthub.cmu.edu/articles/journal_contribution/A_theory_of_reading_From_eye_fixations_to_comprehension/6613262.

193. **Kang, W.-C.**, **J. Ni**, **N. Mehta**, **M. Sathiamoorthy**, **L. Hong**, **E. Chi**, and **D. Z. Cheng** (2023). Do llms understand user preferences? evaluating llms on user rating prediction. *arXiv preprint arXiv:2305.06474*.

194. **Karessli, N.**, **Z. Akata**, **B. Schiele**, and **A. Bulling**, Gaze embeddings for zero-shot image classification. *In IEEE CVPR*. 2017.

195. **Ke, J.**, **K. Ye**, **J. Yu**, **Y. Wu**, **P. Milanfar**, and **F. Yang**, Vila: Learning image aesthetics from user comments with vision-language pretraining. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

196. **Keller, P. A.**, **I. M. Lipkus**, and **B. K. Rimer** (2003). Affect, framing, and persuasion. *Journal of Marketing Research*, **40**(1), 54–64.

197. **Kennedy, A.**, **J. Pynte**, **W. S. Murray**, and **S.-A. Paul** (2013). Frequency and predictability effects in the dundee corpus: An eye movement analysis. *Quarterly Journal of Experimental Psychology*, **66**(3), 601–618.

198. **Khandelwal, A.**, **A. Agrawal**, **A. Bhattacharyya**, **Y. Kumar**, **S. Singh**, **U. Bhattacharya**, **I. Dasgupta**, **S. Petrangeli**, **R. R. Shah**, **C. Chen**, and **B. Krishnamurthy**, Large content and behavior models to understand, simulate, and optimize content and behavior. *In The Twelfth International Conference on Learning Representations*. 2024. URL https://openreview.net/forum?id=TrKq4Wlwcz.

199. **Khosla, A.**, **W. A. Bainbridge**, **A. Torralba**, and **A. Oliva**, Modifying the memorability of face photographs. *In Proceedings of the IEEE international conference on computer vision*. 2013.

200. **Khosla, A.**, **A. Das Sarma**, and **R. Hamid**, What makes an image popular? *In Proceedings of the 23rd international conference on World wide web*. 2014.

201. **Khosla, A.**, **A. S. Raju**, **A. Torralba**, and **A. Oliva**, Understanding and predicting image memorability at a large scale. *In Proceedings of the IEEE international conference on computer vision*. 2015.

202. **Khrulkov, V.** and **I. V. Oseledets**, Art of singular vectors and universal adversarial perturbations. *In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018.

203. **Khurana, V.**, **Y. Kumar**, **N. Hollenstein**, **R. Kumar**, and **B. Krishnamurthy**, Synthesizing human gaze feedback for improved NLP performance. *In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 2023. URL https://aclanthology.org/2023.eacl-main.139.

204. **Khurana, V.**, **Y. K. Singla**, **J. Subramanian**, **R. R. Shah**, **C. Chen**, **Z. Xu**, and **B. Krishnamurthy** (2024). Measuring and improving engagement of text-to-image generation models. *arXiv preprint arXiv:2311.10995*.

205. **Kim, S.**, **J.-Y. Jiang**, **M. Nakada**, **J. Han**, and **W. Wang**, Multimodal post attentive profiling for influencer marketing. *In Proceedings of The Web Conference 2020*. 2020.

206. **Kirby, S.**, **H. Cornish**, and **K. Smith** (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, **105**(31), 10681–10686.

207. **Kirstain, Y.**, **A. Polyak**, **U. Singer**, **S. Matiana**, **J. Penna**, and **O. Levy** (2023). Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, **36**, 36652–36663.

208. **Kiziltepe, R. S.**, **L. Sweeney**, **M. G. Constantin**, **F. Doctor**, **A. G. S. de Herrera**, **C.-H. Demarty**, **G. Healy**, **B. Ionescu**, and **A. F. Smeaton** (2021*a*). An annotated video dataset for computing video memorability. *Data in Brief*, **39**, 107671.

209. **Kiziltepe, R. S.**, **L. Sweeney**, **M. G. Constantin**, **F. Doctor**, **A. G. S. de Herrera**, **C.-H. Demarty**, **G. Healy**, **B. Ionescu**, and **A. F. Smeaton** (2021*b*). An annotated video dataset for computing video memorability. *Data in Brief*, **39**, 107671. URL https://doi.org/10.1016%2Fj.dib.2021.107671.

210. **Klerke, S.**, **Y. Goldberg**, and **A. Søgaard**, Improving sentence compression by learning to predict gaze. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 2016*a*. URL https://aclanthology.org/N16-1179.

211. **Klerke, S.**, **Y. Goldberg**, and **A. Søgaard**, Improving sentence compression by learning to predict gaze. *In NAACL: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 2016*b*. URL https://aclanthology.org/N16-1179.

212. **Klimt, B.** and **Y. Yang**, The enron corpus: A new dataset for email classification research. *In European conference on machine learning*. Springer, 2004.

213. **Knowles, E. S.** and **J. A. Linn**, *Resistance and persuasion*. Psychology Press, 2004.

214. **Kotseruba, I.** and **J. K. Tsotsos** (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*.

215. **Krebs, J. R.** and **R. Dawkins** (1984). Animal signals: mind-reading and manipulation.

216. **Kreutzer, J.**, **S. Khadivi**, **E. Matusov**, and **S. Riezler** (2018). Can neural machine translation be improved with user feedback? *arXiv preprint arXiv:1804.05958*.

217. **Krishna, R.**, **K. Hata**, **F. Ren**, **L. Fei-Fei**, and **J. Carlos Niebles**, Dense-captioning events in videos. *In Proceedings of the IEEE international conference on computer vision*. 2017*a*.

218. **Krishna, R.**, **Y. Zhu**, **O. Groth**, **J. Johnson**, **K. Hata**, **J. Kravitz**, **S. Chen**, **Y. Kalantidis**, **L.-J. Li**, **D. A. Shamma**, *et al.* (2017*b*). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, **123**, 32–73.

219. **Krumme, C.**, **A. Llorente**, **M. Cebrian**, **A. Pentland**, and **E. Moro** (2013). The predictability of consumer visitation patterns. *Scientific reports*, **3**(1), 1645.

220. **Kruthiventi, S. S.**, **V. Gudisa**, **J. H. Dholakiya**, and **R. V. Babu**, Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

221. **Kumar, Y.**, **S. Aggarwal**, **D. Mahata**, **R. R. Shah**, **P. Kumaraguru**, and **R. Zimmermann**, Get IT scored using autosas - an automated system for scoring short answers. *In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019. URL https://doi.org/10.1609/aaai.v33i01.33019662.

222. **Kumar, Y.**, **R. Jha**, **A. Gupta**, **M. Aggarwal**, **A. Garg**, **T. Malyan**, **A. Bhardwaj**, **R. R. Shah**, **B. Krishnamurthy**, and **C. Chen**, Persuasion strategies in advertisements. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37. 2023*a*.

223. **Kumar, Y.**, **R. Jha**, **A. Gupta**, **M. Aggarwal**, **A. Garg**, **T. Malyan**, **A. Bhardwaj**, **R. R. Shah**, **B. Krishnamurthy**, and **C. Chen** (2023*b*). Persuasion strategies in advertisements. *Proceedings of the AAAI Conference on Artificial Intelligence*.

224. **Kümmerer, M.** and **M. Bethge** (2021). State-of-the-art in human scanpath prediction. *arXiv preprint arXiv:2102.12239*.

225. **Kümmerer, M.**, **M. Bethge**, and **T. S. Wallis** (2022). Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*.

226. **Kyle-Davidson, C.**, **A. G. Bors**, and **K. K. Evans** (2022). Modulating human memory for complex scenes with artificially generated images. *Scientific Reports*, **12**(1), 1583.

227. **Labs, B. F.** (2024). Announcing black forest labs. URL https://blackforestlabs.ai/announcing-black-forest-labs/.

228. **Lan, Z.**, **M. Chen**, **S. Goodman**, **K. Gimpel**, **P. Sharma**, and **R. Soricut**, ALBERT: A lite BERT for self-supervised learning of language representations. *In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

229. **Lasswell, H. D.** (1948). The structure and function of communication in society. *The communication of ideas*, **37**(1), 136–139.

230. **Lasswell, H. D.**, *Propaganda technique in world war I*. MIT press, 1971.

231. **Lavidge, R. J.** and **G. A. Steiner** (1961). A model for predictive measurements of advertising effectiveness. *Journal of marketing*, **25**(6), 59–62.

232. **Lee, A. Y.**, **P. A. Keller**, and **B. Sternthal** (2010). Value from regulatory construal fit: The persuasive impact of fit between consumer goals and message concreteness. *Journal of Consumer Research*, **36**(5), 735–747.

233. **Lee, J.**, **S. Kim**, **S. Kim**, **J. Park**, and **K. Sohn**, Context-aware emotion recognition networks. *In Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

234. **Lee, K.**, **H. Liu**, **M. Ryu**, **O. Watkins**, **Y. Du**, **C. Boutilier**, **P. Abbeel**, **M. Ghavamzadeh**, and **S. S. Gu** (2023). Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*.

235. **Lei, J.**, **T. L. Berg**, and **M. Bansal** (2021). Qvhighlights: Detecting moments and highlights in videos via natural language queries. URL https://arxiv.org/abs/2107.09609.

236. **Levenshtein, V.** (1965). Leveinshtein distance.

237. **Levesque, N.** and **F. Pons** (2020). The human brand: A systematic literature review and research agenda. *Journal of Customer Behaviour*, **19**(2), 143–174.

238. **Lewis, D. D.** and **J. Catlett**, Heterogeneous uncertainty sampling for supervised learning. *In Machine learning proceedings 1994*. Elsevier, 1994, 148–156.

239. **Li, C.** (2010). Primacy effect or recency effect? a long-term memory test of super bowl commercials. *Journal of Consumer Behaviour: An International Research Review*, **9**(1), 32–44.

240. **Li, J.**, **E. Durmus**, and **C. Cardie**, Exploring the role of argument structure in online debate persuasion. *In EMNLP*. Association for Computational Linguistics, Online, 2020*a*. URL https://aclanthology.org/2020.emnlp-main.716.

241. **Li, J.**, **R. Ji**, **H. Liu**, **X. Hong**, **Y. Gao**, and **Q. Tian**, Universal perturbation attack against image retrieval. *In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019. URL https://doi.org/10.1109/ICCV.2019.00500.

242. **Li, J.**, **D. Li**, **S. Savarese**, and **S. Hoi** (2023*a*). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

243. **Li, J.**, **D. Li**, **S. Savarese**, and **S. Hoi** (2023*b*). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

244. **Li, J.**, **D. Li**, **C. Xiong**, and **S. Hoi**, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *In ICML*. 2022.

245. **Li, J.**, **Y. Wong**, **Q. Zhao**, and **M. S. Kankanhalli** (2020*b*). Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, **22**(2), 554–565.

246. **Li, K.**, **Y. He**, **Y. Wang**, **Y. Li**, **W. Wang**, **P. Luo**, **Y. Wang**, **L. Wang**, and **Y. Qiao** (2023*c*). Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

247. **Li, K.**, **Y. Wang**, **G. Peng**, **G. Song**, **Y. Liu**, **H. Li**, and **Y. Qiao**, Uniformer: Unified transformer for efficient spatial-temporal representation learning. *In International Conference on Learning Representations*. 2021.

248. **Li, Y.**, **C. Wang**, and **J. Jia** (2023*d*). Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

249. **Liang, P. P.**, **A. Zadeh**, and **L.-P. Morency** (2022). Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*.

250. **Libet, B.** (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and brain sciences*, **8**(4), 529–539.

251. **Libet, B.**, **B. Libet**, **C. A. Gleason**, **E. W. Wright**, and **D. K. Pearl** (1993). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential) the unconscious initiation of a freely voluntary act. *Neurophysiology of consciousness*, 249–268.

252. **Lin, T.-Y.**, **P. Goyal**, **R. Girshick**, **K. He**, and **P. Dollár**, Focal loss for dense object detection. *In Proceedings of the IEEE international conference on computer vision*. 2017.

253. **Lin, T.-Y.**, **M. Maire**, **S. Belongie**, **J. Hays**, **P. Perona**, **D. Ramanan**, **P. Dollár**, and **C. L. Zitnick**, Microsoft coco: Common objects in context. *In ECCV*. Springer, 2014.

254. **Liu, H.**, **C. Li**, **Q. Wu**, and **Y. J. Lee** (2023*a*). Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

255. **Liu, R.**, **C. Li**, **Y. Ge**, **Y. Shan**, **T. H. Li**, and **G. Li** (2023*b*). One for all: Video conversation is feasible without video instruction tuning. *arXiv preprint arXiv:2309.15785*.

256. **Liu, S.**, **Z. Zeng**, **T. Ren**, **F. Li**, **H. Zhang**, **J. Yang**, **C. Li**, **J. Yang**, **H. Su**, **J. Zhu**, *et al.* (2023*c*). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

257. **Liu, W.**, **D. Anguelov**, **D. Erhan**, **C. Szegedy**, **S. Reed**, **C.-Y. Fu**, and **A. C. Berg**, Ssd: Single shot multibox detector. *In ECCV*. Springer, 2016.

258. **Long, Y.**, **Q. Lu**, **R. Xiang**, **M. Li**, and **C.-R. Huang**, A cognition based attention model for sentiment analysis. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2017. URL https://aclanthology.org/D17-1048.

259. **Longpre, L.**, **E. Durmus**, and **C. Cardie**, Persuasion of the undecided: Language vs. the listener. *In Proceedings of the 6th Workshop on Argument Mining*. 2019.

260. **Lu, Y.** and **X. Wu**, Cross-modal interaction for video memorability prediction. *In* **S. Hicks**, **K. Pogorelov**, **A. Lommatzsch**, **A. G. S. de Herrera**, **P. Martin**, **S. Z. Hassan**, **A. Porter**, **A. Kasem**, **S. Andreadis**, **M. Lux**, **M. G. Ocaña**, **A. Liu**, and **M. A. Larson** (eds.), *Working Notes Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021*, volume 3181 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL https://ceur-ws.org/Vol-3181/paper18.pdf.

261. **Lu, Y.** and **X. Wu** (2022). Video storytelling based on gated video memorability filtering. *Electronics Letters*, **58**(15), 576–578. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ell2.12525.

262. **Lukin, S.**, **P. Anand**, **M. Walker**, and **S. Whittaker**, Argument strength is in the eye of the beholder: Audience effects in persuasion. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, 2017. URL https://aclanthology.org/E17-1070.

263. **Lukito, J.** (2020). Coordinating a multi-platform disinformation campaign: Internet research agency activity on three us social media platforms, 2015 to 2017. *Political Communication*, **37**(2), 238–255.

264. **Luu, K.**, **C. Tan**, and **N. A. Smith** (2019). Measuring online debaters' persuasive skill from text over time. *Transactions of the Association for Computational Linguistics*, **7**, 537–550.

265. **Lynn, M.** (1991). Scarcity effects on value: A quantitative review of the commodity theory literature. *Psychology & Marketing*, **8**(1), 43–57.

266. **Maas, A. L.**, **R. E. Daly**, **P. T. Pham**, **D. Huang**, **A. Y. Ng**, and **C. Potts**, Learning word vectors for sentiment analysis. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 2011. URL http://www.aclweb.org/anthology/P11-1015.

267. **Maaz, M.**, **H. Rasheed**, **S. Khan**, and **F. S. Khan** (2023). Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

268. **Machajdik, J.** and **A. Hanbury**, Affective image classification using features inspired by psychology and art theory. *In Proceedings of the 18th ACM international conference on Multimedia*. 2010.

269. **Mack, P.**, *A History of Renaissance rhetoric 1380-1620*. OUP Oxford, 2011.

270. **Mai, L.-W.** and **G. Schoeller** (2009). Emotions, attitudes and memorability associated with tv commercials. *Journal of Targeting, Measurement and Analysis for Marketing*, **17**, 55–63.

271. **Malavolta, M.**, **E. Trimarco**, **V. Groznik**, and **A. Sadikov**, Awareness of being tested and its effect on reading behaviour. *In International Conference on Artificial Intelligence in Medicine*. Springer, 2022.

272. **Marcinkiewicz, M. A.** (1994). Building a large annotated corpus of english: The penn treebank. *Using Large Corpora*.

273. **Martin, D.**, **A. Serrano**, **A. W. Bergman**, **G. Wetzstein**, and **B. Masia** (2022). Scangan360: A generative model of realistic scanpaths for 360° images. *IEEE Transactions on Visualization and Computer Graphics*.

274. **Martin, J.-C.**, **S. Grimard**, and **K. Alexandri**, On the annotation of the multimodal behavior and computation of cooperation between modalities. *In Proceedings of the Workshop on Multimodal Communicatiom and Context in Embodied Agents, Fifth International Conference on Autonomous Agents*. 2001.

275. **Martin, T.**, **J. M. Hofman**, **A. Sharma**, **A. Anderson**, and **D. J. Watts**, Exploring limits to prediction in complex social systems. *In Proceedings of the 25th international conference on world wide web*. 2016.

276. **Mathias, S.**, **D. Kanojia**, **A. Mishra**, and **P. Bhattacharya**, A survey on using gaze behaviour for natural language processing. *In* **C. Bessiere** (ed.), *IJCAI*. 2020. Survey track.

277. **Matthies, F.** and **A. Søgaard**, With blinkers on: Robust prediction of eye movements across readers. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 2013. URL https://aclanthology.org/D13-1075.

278. **Mazloom, M.**, **R. Rietveld**, **S. Rudinac**, **M. Worring**, and **W. Van Dolen**, Multimodal popularity prediction of brand-related social media posts. *In Proceedings of the 24th ACM international conference on Multimedia*. 2016.

279. **McCarney, R.**, **J. Warner**, **S. Iliffe**, **R. Van Haselen**, **M. Griffin**, and **P. Fisher** (2007). The hawthorne effect: a randomised, controlled trial. *BMC medical research methodology*, **7**(1), 1–8.

280. **McCroskey, J. C.**, *An introduction to rhetorical communication*. Routledge, 2015.

281. **McGinnies, E.** and **C. D. Ward** (1980). Better liked than right: Trustworthiness and expertise as factors in credibility. *Personality and Social Psychology Bulletin*, **6**(3), 467–472.

282. **McGuire, W. J.** (1964). Inducing resistance to persuasion. some contemporary approaches. *CC Haaland and WO Kaelber (Eds.), Self and Society. An Anthology of Readings, Lexington, Mass.(Ginn Custom Publishing) 1981, pp. 192-230.*.

283. **McGuire, W. J.** and **D. Papageorgis** (1961). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *Journal of Abnormal Psychology*, **62**(2).

284. **McMahan, B.**, **E. Moore**, **D. Ramage**, **S. Hampson**, and **B. A. y Arcas**, Communication-efficient learning of deep networks from decentralized data. *In Artificial intelligence and statistics*. PMLR, 2017.

285. **McMahan, H. B.**, **G. Holt**, **D. Sculley**, **M. Young**, **D. Ebner**, **J. Grady**, **L. Nie**, **T. Phillips**, **E. Davydov**, **D. Golovin**, *et al.*, Ad click prediction: a view from the trenches. *In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013.

286. **McQuail, D.** and **S. Windahl**, *Communication models for the study of mass communications*. Routledge, 2015.

287. **Meier, R. L.**, The measurement of social change. *In Papers presented at the the March 3-5, 1959, western joint computer conference*. 1959.

288. **Meng, R.**, **S. Zhao**, **S. Han**, **D. He**, **P. Brusilovsky**, and **Y. Chi**, Deep keyphrase generation. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 2017. URL https://www.aclweb.org/anthology/P17-1054.

289. **Mestyán, M.**, **T. Yasseri**, and **J. Kertész** (2013). Early prediction of movie box office success based on wikipedia activity big data. *PloS one*, **8**(8), e71226.

290. **Meyers-Levy, J.** and **P. Malaviya** (1999). Consumers' processing of persuasive advertisements: An integrative framework of persuasion theories. *Journal of marketing*, **63**(4_suppl1), 45–60.

291. **Micaelli, P.** and **A. J. Storkey**, Zero-shot knowledge transfer via adversarial belief matching. *In* **H. M. Wallach**, **H. Larochelle**, **A. Beygelzimer**, **F. d'Alché-Buc**, **E. B. Fox**, and **R. Garnett** (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fe663a72b27bdc613873fbbb512f6f67-Abstract.html.

292. **Michel, P.**, **X. Li**, **G. Neubig**, and **J. Pino**, On evaluation of adversarial perturbations for sequence-to-sequence models. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2019. URL https://www.aclweb.org/anthology/N19-1314.

293. **Mikels, J. A.**, **B. L. Fredrickson**, **G. R. Larkin**, **C. M. Lindberg**, **S. J. Maglio**, and **P. A. Reuter-Lorenz** (2005). Emotional category data on images from the international affective picture system. *Behavior research methods*, **37**, 626–630.

294. **Mikolov, T.**, **K. Chen**, **G. Corrado**, and **J. Dean** (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

295. **Mikolov, T.**, **E. Grave**, **P. Bojanowski**, **C. Puhrsch**, and **A. Joulin**, Advances in pre-training distributed word representations. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL https://www.aclweb.org/anthology/L18-1008.

296. **Milgram, S.** (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology*, **67**(4), 371.

297. **Milgram, S.** and **C. Gudehus** (1978). Obedience to authority.

298. **Miller, G. R.** (1966). On defining communication: Another stab. *Journal of Communication*.

299. **Miritello, G.**, **R. Lara**, **M. Cebrian**, and **E. Moro** (2013). Limited communication capacity unveils strategies for human interaction. *Scientific reports*, **3**(1), 1950.

300. **Mishra, A.**, **D. Kanojia**, and **P. Bhattacharyya**, Predicting readers' sarcasm understandability by modeling gaze behavior. *In AAAI*. 2016*a*.

301. **Mishra, A.**, **D. Kanojia**, **S. Nagar**, **K. Dey**, and **P. Bhattacharyya**, Harnessing cognitive features for sarcasm detection. *In ACL (Volume 1: Long Papers)*. ACL, 2016*b*. URL https://aclanthology.org/P16-1104.

302. **Mishra, A.**, **D. Kanojia**, **S. Nagar**, **K. Dey**, and **P. Bhattacharyya**, Leveraging cognitive features for sentiment analysis. *In SIGNLL*. ACL, 2016*c*. URL https://aclanthology.org/K16-1016.

303. **Mishra, A.**, **D. Kanojia**, **S. Nagar**, **K. Dey**, and **P. Bhattacharyya** (2017). Scanpath complexity: Modeling reading effort using gaze information. *AAAI*. URL https://ojs.aaai.org/index.php/AAAI/article/view/11159.

304. **Misyak, J.**, **T. Noguchi**, and **N. Chater** (2016). Instantaneous conventions: The emergence of flexible communicative signals. *Psychological science*, **27**(12), 1550–1561.

305. **Mitchell, M.**, **S. Wu**, **A. Zaldivar**, **P. Barnes**, **L. Vasserman**, **B. Hutchinson**, **E. Spitzer**, **I. D. Raji**, and **T. Gebru**, Model cards for model reporting. *In Proceedings of the conference on fairness, accountability, and transparency*. 2019.

306. **Moore, Thomas** (2021). Hhs plans mega $250m 'defeat despair' COVID-19 campaign. URL https://www.prweek.com/article/1693203/hhs-plans-mega-250m-defeat-despair-covid-19-campaign.

307. **Moosavi-Dezfooli, S.**, **A. Fawzi**, **O. Fawzi**, and **P. Frossard**, Universal adversarial perturbations. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017. URL https://doi.org/10.1109/CVPR.2017.17.

308. **Mopuri, K. R.**, **A. Ganeshan**, and **R. V. Babu** (2018*a*). Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, **41**(10), 2452–2465.

309. **Mopuri, K. R.**, **U. Garg**, and **V. B. Radhakrishnan**, Fast feature fool: A data independent approach to universal adversarial perturbations. *In British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017.

310. **Mopuri, K. R.**, **P. K. Uppala**, and **R. V. Babu**, Ask, acquire, and attack: Data-free uap generation using class impressions. *In Proceedings of the European Conference on Computer Vision (ECCV)*. 2018*b*.

311. **Nakano, R.**, **J. Hilton**, **S. Balaji**, **J. Wu**, **L. Ouyang**, **C. Kim**, **C. Hesse**, **S. Jain**, **V. Kosaraju**, **W. Saunders**, *et al.* (2021). Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

312. **Nayak, G. K.**, **K. R. Mopuri**, **V. Shaj**, **V. B. Radhakrishnan**, and **A. Chakraborty**, Zero-shot knowledge distillation in deep networks. *In* **K. Chaudhuri** and **R. Salakhutdinov** (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*. PMLR, 2019. URL http://proceedings.mlr.press/v97/nayak19a.html.

313. **Newman, A.**, **C. Fosco**, **V. Casser**, **A. Lee**, **B. McNamara**, and **A. Oliva**, Multimodal memorability: Modeling effects of semantics and decay on video memorability. *In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020.

314. **Newstead, K.** and **J. Romaniuk** (2010). Cost per second: The relative effectiveness of 15-and 30-second television advertisements. *Journal of Advertising Research*, **50**(1), 68–76.

315. **Nichol, A. Q.**, **P. Dhariwal**, **A. Ramesh**, **P. Shyam**, **P. Mishkin**, **B. Mcgrew**, **I. Sutskever**, and **M. Chen**, Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *In International Conference on Machine Learning*. PMLR, 2022.

316. **Norlin, G.**, **L. Van Hook**, *et al.*, *Isocrates*, volume 1. Harvard University Press, 1928.

317. **Norris, D.** (2017). Short-term memory and long-term memory are still different. *Psychological bulletin*, **143**(9), 992.

318. **OpenAI** (2023). Gpt-4 technical report.

319. **Osgood, C. E.**, **G. J. Suci**, and **P. H. Tannenbaum**, *The measurement of meaning*. 47. University of Illinois press, 1957.

320. **Ouyang, L.**, **J. Wu**, **X. Jiang**, **D. Almeida**, **C. L. Wainwright**, **P. Mishkin**, **C. Zhang**, **S. Agarwal**, **K. Slama**, **A. Ray**, *et al.* (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

321. **Parikh, A.**, **O. Täckström**, **D. Das**, and **J. Uszkoreit**, A decomposable attention model for natural language inference. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2016. URL https://www.aclweb.org/anthology/D16-1244.

322. **Park, J. S.**, **J. C. O'Brien**, **C. J. Cai**, **M. R. Morris**, **P. Liang**, and **M. S. Bernstein** (2023). Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

323. **Park, J. S.**, **C. Q. Zou**, **A. Shaw**, **B. M. Hill**, **C. Cai**, **M. R. Morris**, **R. Willer**, **P. Liang**, and **M. S. Bernstein** (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

324. **Patro, B.** and **V. P. Namboodiri**, Differential attention for visual question answering. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

325. **Penedo, G.**, **Q. Malartic**, **D. Hesslow**, **R. Cojocaru**, **A. Cappelli**, **H. Alobeidli**, **B. Pannier**, **E. Almazrouei**, and **J. Launay** (2023). The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

326. **Peng, K.-C.**, **T. Chen**, **A. Sadovnik**, and **A. C. Gallagher**, A mixed bag of emotions: Model, predict, and transfer emotion distributions. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

327. **Pennington, J.**, **R. Socher**, and **C. Manning**, GloVe: Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 2014. URL https://www.aclweb.org/anthology/D14-1162.

328. **Perner, J.** and **H. Wimmer** (1985). "john thinks that mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, **39**(3), 437–471.

329. **Peters, M.**, **M. Neumann**, **M. Iyyer**, **M. Gardner**, **C. Clark**, **K. Lee**, and **L. Zettlemoyer**, Deep contextualized word representations. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2018. URL https://www.aclweb.org/anthology/N18-1202.

330. **Petty, R. E.** and **J. T. Cacioppo**, The elaboration likelihood model of persuasion. *In Communication and persuasion*. Springer, 1986, 1–24.

331. **Petty, R. E.**, **J. T. Cacioppo**, and **D. Schumann** (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of consumer research*, **10**(2), 135–146.

332. **Petty, R. E.**, **D. T. Wegener**, and **L. R. Fabrigar** (1997). Attitudes and attitude change. *Annual review of psychology*, **48**(1), 609–647.

333. **Pinkney, J.** (2022). Text-to-pokemon. https://replicate.com/lambdal/text-to-pokemon.

334. **Plank, B.**, Keystroke dynamics as signal for shallow syntactic parsing. *In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016*a*.

335. **Plank, B.**, Keystroke dynamics as signal for shallow syntactic parsing. *In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 2016*b*. URL https://aclanthology.org/C16-1059.

336. **PLUTCHIK, R.**, Chapter 1 - a general psychoevolutionary theory of emotion. *In* **R. Plutchik** and **H. Kellerman** (eds.), *Theories of Emotion*. Academic Press, 1980. ISBN 978-0-12-558701-3, 3–33. URL https://www.sciencedirect.com/science/article/pii/B9780125587013500077.

337. **Podell, D.**, **Z. English**, **K. Lacey**, **A. Blattmann**, **T. Dockhorn**, **J. Müller**, **J. Penna**, and **R. Rombach** (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

338. **Poliak, A.**, **J. Naradowsky**, **A. Haldar**, **R. Rudinger**, and **B. Van Durme**, Hypothesis only baselines in natural language inference. *In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, New Orleans, Louisiana, 2018. URL https://www.aclweb.org/anthology/S18-2023.

339. **Prinz, W.** (1997). Perception and action planning. *European journal of cognitive psychology*, **9**(2), 129–154.

340. **PromptHero** (2023). Openjourneyv2. https://huggingface.co/prompthero/openjourney-v2.

341. **Pryzant, R.**, **Y. Chung**, and **D. Jurafsky** (2017). Predicting sales from the language of product descriptions. *eCOM@ SIGIR*, **2311**.

342. **Putrevu, S.**, **J. Tan**, and **K. R. Lord** (2004). Consumer responses to complex advertisements: The moderating role of need for cognition, knowledge, and gender. *Journal of Current Issues & Research in Advertising*, **26**(1), 9–24.

343. **Qian, R.**, **T. Meng**, **B. Gong**, **M.-H. Yang**, **H. Wang**, **S. Belongie**, and **Y. Cui**, Spatiotemporal contrastive video representation learning. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

344. **Qin, X.**, **Z. Zhang**, **C. Huang**, **M. Dehghan**, **O. Zaiane**, and **M. Jagersand**, U2-net: Going deeper with nested u-structure for salient object detection. volume 106. 2020*a*.

345. **Qin, X.**, **Z. Zhang**, **C. Huang**, **M. Dehghan**, **O. R. Zaiane**, and **M. Jagersand** (2020*b*). U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, **106**, 107404.

346. **Qiu, H.**, **L. He**, and **F. Wang**, Dual focus attention network for video emotion recognition. *In 2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020.

347. **Radford, A.**, **J. W. Kim**, **C. Hallacy**, **A. Ramesh**, **G. Goh**, **S. Agarwal**, **G. Sastry**, **A. Askell**, **P. Mishkin**, **J. Clark**, *et al.*, Learning transferable visual models from natural language supervision. *In International conference on machine learning*. PMLR, 2021.

348. **Radford, A.**, **J. W. Kim**, **T. Xu**, **G. Brockman**, **C. McLeavey**, and **I. Sutskever** (2022). Robust speech recognition via large-scale weak supervision.

349. **Radford, A.**, **J. W. Kim**, **T. Xu**, **G. Brockman**, **C. McLeavey**, and **I. Sutskever**, Robust speech recognition via large-scale weak supervision. *In International Conference on Machine Learning*. PMLR, 2023.

350. **Radford, A.**, **K. Narasimhan**, **T. Salimans**, **I. Sutskever**, *et al.* (2018). Improving language understanding by generative pre-training.

351. **Radford, A.**, **J. Wu**, **R. Child**, **D. Luan**, **D. Amodei**, **I. Sutskever**, *et al.* (2019). Language models are unsupervised multitask learners. *OpenAI blog*, **1**(8), 9.

352. **Rafailov, R.**, **A. Sharma**, **E. Mitchell**, **C. D. Manning**, **S. Ermon**, and **C. Finn** (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, **36**.

353. **Raffel, C.**, **N. Shazeer**, **A. Roberts**, **K. Lee**, **S. Narang**, **M. Matena**, **Y. Zhou**, **W. Li**, and **P. J. Liu** (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, **21**(1), 5485–5551.

354. **Rajbhandari, S.**, **J. Rasley**, **O. Ruwase**, and **Y. He**, Zero: Memory optimizations toward training trillion parameter models. *In SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2020.

355. **Ramesh, A.**, **P. Dhariwal**, **A. Nichol**, **C. Chu**, and **M. Chen** (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, **1**(2), 3.

356. **Ramesh, A.**, **M. Pavlov**, **G. Goh**, **S. Gray**, **C. Voss**, **A. Radford**, **M. Chen**, and **I. Sutskever**, Zero-shot text-to-image generation. *In International conference on machine learning*. Pmlr, 2021.

357. **Rapp, C.** (2002). Aristotle's rhetoric.

358. **Rasley, J.**, **S. Rajbhandari**, **O. Ruwase**, and **Y. He**, Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. *In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020.

359. **Rayner, K.**, **K. Chace**, **T. Slattery**, and **J. Ashby** (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading - SCI STUD READ*.

360. **Reddit, Inc.** (2024). Reddit content policy. https://www.redditinc.com/policies/content-policy. Accessed: April 11, 2024.

361. **Redmon, J.**, **S. Divvala**, **R. Girshick**, and **A. Farhadi**, You only look once: Unified, real-time object detection. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

362. **Regan, D. T.** (1971). Effects of a favor and liking on compliance. *Journal of experimental social psychology*, **7**(6), 627–639.

363. **Reich, D. R.**, **P. Prasse**, **C. Tschirner**, **P. Haller**, **F. Goldhammer**, and **L. A. Jäger**, Inferring native and non-native human reading comprehension and subjective text difficulty from scanpaths in reading. *In ETRA*. Association for Computing Machinery, 2022.

364. **Ren, J.**, **S. Rajbhandari**, **R. Y. Aminabadi**, **O. Ruwase**, **S. Yang**, **M. Zhang**, **D. Li**, and **Y. He**, {ZeRO-Offload}: Democratizing {Billion-Scale} model training. *In 2021 USENIX Annual Technical Conference (USENIX ATC 21)*. 2021.

365. **Ren, Y.** and **D. Xiong**, CogAlign: Learning to align textual neural representations to cognitive language processing signals. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2021. URL https://aclanthology.org/2021.acl-long.291.

366. **Rescala, P.**, **M. H. Ribeiro**, **T. Hu**, and **R. West** (2024). Can language models recognize convincing arguments? *arXiv preprint arXiv:2404.00750*.

367. **Roethlisberger, F. J.** and **W. J. Dickson**, *Management and the Worker*, volume 5. Psychology press, 2003.

368. **Rombach, R.**, **A. Blattmann**, **D. Lorenz**, **P. Esser**, and **B. Ommer**, High-resolution image synthesis with latent diffusion models. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

369. **Romero, E.**, **M. Chica**, **S. Damas**, and **W. Rand** (2023). Two decades of agent-based modeling in marketing: a bibliometric analysis. *Progress in Artificial Intelligence*, 1–17.

370. **Rosenthal, S.** and **K. Mckeown** (2017). Detecting influencers in multiple online genres. *ACM Transactions on Internet Technology (TOIT)*, **17**(2), 1–22.

371. **Rothman, A. J.**, **S. C. Martino**, **B. T. Bedell**, **J. B. Detweiler**, and **P. Salovey** (1999). The systematic influence of gain-and loss-framed messages on interest in and use of different types of health behavior. *Personality and Social Psychology Bulletin*, **25**(11), 1355–1369.

372. **Saharia, C.**, **W. Chan**, **S. Saxena**, **L. Li**, **J. Whang**, **E. L. Denton**, **K. Ghasemipour**, **R. Gontijo Lopes**, **B. Karagol Ayan**, **T. Salimans**, *et al.* (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, **35**, 36479–36494.

373. **Salganik, M. J.**, *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

374. **Salimans, T.**, **I. Goodfellow**, **W. Zaremba**, **V. Cheung**, **A. Radford**, and **X. Chen** (2016). Improved techniques for training gans. *Advances in neural information processing systems*, **29**.

375. **Santurkar, S.**, **E. Durmus**, **F. Ladhak**, **C. Lee**, **P. Liang**, and **T. Hashimoto**, Whose opinions do language models reflect? *In International Conference on Machine Learning*. PMLR, 2023.

376. **Schmidt, S.** and **M. Eisend** (2015). Advertising repetition: A meta-analysis on effective frequency in advertising. *Journal of Advertising*, **44**(4), 415–428.

377. **Schuhmann, C.**, **R. Beaumont**, **R. Vencu**, **C. Gordon**, **R. Wightman**, **M. Cherti**, **T. Coombes**, **A. Katta**, **C. Mullis**, **M. Wortsman**, *et al.* (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, **35**, 25278–25294.

378. **Schulman, J.**, **F. Wolski**, **P. Dhariwal**, **A. Radford**, and **O. Klimov** (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

379. **Sekar, K.**, *Domestic Funding for COVID-19 Vaccines:.*. Congressional Research Service, 2021.

380. **Shaikh, O.**, **J. Chen**, **J. Saad-Falcon**, **P. Chau**, and **D. Yang**, Examining the ordering of rhetorical strategies in persuasive requests. *In Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2020. URL https://aclanthology.org/2020.findings-emnlp.116.

381. **Shannon, C. E.** and **W. Weaver**, *The mathematical theory of communication.*. The mathematical theory of communication. University of Illinois Press, Champaign, IL, US, 1949.

382. **Sharma, P.**, **N. Ding**, **S. Goodman**, and **R. Soricut**, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018.

383. **Shmueli, G.** (2010). To explain or to predict? *Statistical Science*.

384. **SI, H.**, **S. Singh**, **Y. K. Singla**, **A. Bhattacharyya**, **V. Baths**, **C. Chen**, **R. R. Shah**, and **B. Krishnamurthy**, Long-term ad memorability: Understanding and generating memorable ads. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2025.

385. **Siarohin, A.**, **G. Zen**, **C. Majtanovic**, **X. Alameda-Pineda**, **E. Ricci**, and **N. Sebe**, How to make an image more memorable? a deep style transfer approach. *In Proceedings of the 2017 ACM on international conference on multimedia retrieval*. 2017.

386. **Siddiquie, B.**, **D. Chisholm**, and **A. Divakaran**, Exploiting multimodal affect and semantics to identify politically persuasive web videos. *In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 2015.

387. **Singh, S.**, **H. SI**, **Y. K. Singla**, **V. Baths**, **R. R. Shah**, **C. Chen**, and **B. Krishnamurthy** (2024*a*). Teaching human behavior improves content understanding abilities of vlms. *arXiv preprint arXiv:2405.00942*.

388. **Singh, S.**, **Y. K. Singla**, **H. SI**, and **B. Krishnamurthy** (2024*b*). Measuring and improving persuasiveness of generative models. *arXiv preprint arXiv:2410.02653*.

389. **Singla, Y. K.**, **S. Parekh**, **S. Singh**, **C. Chen**, **B. Krishnamurthy**, and **R. R. Shah**, Minimal: Mining models for universal adversarial triggers. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36. 2022.

390. **Smith, J. M.** and **D. Harper**, *Animal signals*. Oxford University Press, 2003.

391. **Smith, J. M.** and **E. Szathmary**, *The major transitions in evolution*. OUP Oxford, 1997.

392. **Socher, R.**, **A. Perelygin**, **J. Wu**, **J. Chuang**, **C. D. Manning**, **A. Ng**, and **C. Potts**, Recursive deep models for semantic compositionality over a sentiment treebank. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 2013. URL https://www.aclweb.org/anthology/D13-1170.

393. **Song, C.**, **Z. Qu**, **N. Blumm**, and **A.-L. Barabási** (2010). Limits of predictability in human mobility. *Science*, **327**(5968), 1018–1021.

394. **Song, L.**, **X. Yu**, **H.-T. Peng**, and **K. Narasimhan**, Universal adversarial attacks with natural triggers for text classification. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2021. URL https://www.aclweb.org/anthology/2021.naacl-main.291.

395. **Sood, E.**, **S. Tannert**, **P. Müller**, and **A. Bulling** (2020). Improving natural language processing tasks with human gaze-guided neural attention. *NeurIPS*.

396. **Squire, L. R.** (2009). The legacy of patient hm for neuroscience. *Neuron*, **61**(1), 6–9.

397. **Stab, C.** and **I. Gurevych** (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics*, **43**(3), 619–659.

398. **Stiennon, N.**, **L. Ouyang**, **J. Wu**, **D. Ziegler**, **R. Lowe**, **C. Voss**, **A. Radford**, **D. Amodei**, and **P. F. Christiano** (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, **33**, 3008–3021.

399. **Strack, F.** and **T. Mussweiler** (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of personality and social psychology*, **73**(3), 437.

400. **Strzyz, M.**, **D. Vilares**, and **C. Gómez-Rodríguez**, Towards making a dependency parser see. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2019. URL https://aclanthology.org/D19-1160.

401. **Sun, C.**, **A. Myers**, **C. Vondrick**, **K. Murphy**, and **C. Schmid**, Videobert: A joint model for video and language representation learning. *In Proceedings of the IEEE/CVF international conference on computer vision*. 2019*a*.

402. **Sun, Q.**, **Y. Fang**, **L. Wu**, **X. Wang**, and **Y. Cao** (2023). Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

403. **Sun, W.**, **Z. Chen**, and **F. Wu** (2019*b*). Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE transactions on pattern analysis and machine intelligence*, **43**(6), 2101–2118.

404. **Tan, C.**, **L. Lee**, and **B. Pang**, The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014.

405. **Tan, C.**, **V. Niculae**, **C. Danescu-Niculescu-Mizil**, and **L. Lee**, Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *In Proceedings of the 25th international conference on world wide web*. 2016.

406. **Tappin, B. M.**, **C. Wittenberg**, **L. B. Hewitt**, **A. J. Berinsky**, and **D. G. Rand** (2023). Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences*, **120**(25), e2216261120.

407. **Tetlock, P. E.**, Expert political judgment. *In Expert Political Judgment*. Princeton University Press, 2017.

408. **Tong, Z.**, **Y. Song**, **J. Wang**, and **L. Wang** (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training.

409. **Touvron, H.**, **T. Lavril**, **G. Izacard**, **X. Martinet**, **M.-A. Lachaux**, **T. Lacroix**, **B. Rozière**, **N. Goyal**, **E. Hambro**, **F. Azhar**, *et al.* (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

410. **Tramèr, F.**, **F. Zhang**, **A. Juels**, **M. K. Reiter**, and **T. Ristenpart**, Stealing machine learning models via prediction apis. *In 25th {USENIX} Security Symposium ({USENIX} Security 16)*. 2016.

411. **Tumasjan, A.**, **T. Sprenger**, **P. Sandner**, and **I. Welpe**, Predicting elections with twitter: What 140 characters reveal about political sentiment. *In Proceedings of the international AAAI conference on web and social media*, volume 4. 2010.

412. **Tversky, A.** and **D. Kahneman** (1974). Judgment under uncertainty: Heuristics and biases. *science*, **185**(4157), 1124–1131.

413. **Tversky, A.** and **D. Kahneman**, The framing of decisions and the psychology of choice. *In Behavioral decision making*. Springer, 1985, 25–41.

414. **Vakratsas, D.** and **T. Ambler** (1999). How advertising works: what do we really know? *Journal of marketing*, **63**(1), 26–43.

415. **Van Lange, P. A.**, **E. T. Higgins**, and **A. W. Kruglanski** (2011). Handbook of theories of social psychology. *Handbook of Theories of Social Psychology*, 1–568.

416. **Varan, D.**, **M. Nenycz-Thiel**, **R. Kennedy**, and **S. Bellman** (2020). The effects of commercial length on advertising impact: What short advertisements can and cannot deliver. *Journal of Advertising Research*, **60**(1), 54–70.

417. **Vargheese, J. P.**, **M. Collinson**, and **J. Masthoff**, Exploring susceptibility measures to persuasion. *In Persuasive Technology. Designing for Future Change: 15th International Conference on Persuasive Technology, PERSUASIVE 2020, Aalborg, Denmark, April 20–23, 2020, Proceedings 15*. Springer, 2020.

418. **Villarroel Ordenes, F.**, **D. Grewal**, **S. Ludwig**, **K. D. Ruyter**, **D. Mahr**, and **M. Wetzels** (2019). Cutting through content clutter: How speech and image acts drive consumer sharing of social media brand messages. *Journal of Consumer Research*, **45**(5), 988–1012.

419. **von Platen, P.**, **S. Patil**, **A. Lozhkov**, **P. Cuenca**, **N. Lambert**, **K. Rasul**, **M. Davaadorj**, and **T. Wolf** (2023). Stablediffusion text-to-image fine-tuning — huggingface diffusers documentation. `https://huggingface.co/docs/diffusers/v0.13.0/en/training/text2image`.

420. **Vrandečić, D.** and **M. Krötzsch** (2014). Wikidata: A free collaborative knowledgebase. *Commun. ACM*, **57**(10), 78–85. ISSN 0001-0782. URL `https://doi.org/10.1145/2629489`.

421. **Wachsmuth, H.**, **N. Naderi**, **Y. Hou**, **Y. Bilu**, **V. Prabhakaran**, **T. A. Thijm**, **G. Hirst**, and **B. Stein**, Computational argumentation quality assessment in natural language. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 2017.

422. **Wallace, B.**, **M. Dang**, **R. Rafailov**, **L. Zhou**, **A. Lou**, **S. Purushwalkam**, **S. Ermon**, **C. Xiong**, **S. Joty**, and **N. Naik**, Diffusion model alignment using direct preference optimization. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

423. **Wallace, E.**, **S. Feng**, **N. Kandpal**, **M. Gardner**, and **S. Singh**, Universal adversarial triggers for attacking and analyzing NLP. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2019. URL `https://www.aclweb.org/anthology/D19-1221`.

424. **Wang, A.**, **A. Singh**, **J. Michael**, **F. Hill**, **O. Levy**, and **S. Bowman**, GLUE: A multi-task benchmark and analysis platform for natural language understanding. *In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 2018*a*. URL https://aclanthology.org/W18-5446.

425. **Wang, J.**, **J. Tang**, **M. Yang**, **X. Bai**, and **J. Luo**, Improving ocr-based image captioning by incorporating geometrical relationship. *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021*a*.

426. **Wang, K.**, **M. Bansal**, and **J.-M. Frahm**, Retweet wars: Tweet popularity prediction via dynamic multimodal regression. *In 2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018*b*.

427. **Wang, L.**, **J. Zhang**, **H. Yang**, **Z.-Y. Chen**, **J. Tang**, **Z. Zhang**, **X. Chen**, **Y. Lin**, **H. Sun**, **R. Song**, *et al.* (2025). User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, **43**(2), 1–37.

428. **Wang, S.**, **J. Zhang**, and **C. Zong**, Learning sentence representation with guidance of human attention. *In IJCAI*. AAAI Press, 2017. ISBN 9780999241103.

429. **Wang, W.**, **Q. Lai**, **H. Fu**, **J. Shen**, **H. Ling**, and **R. Yang** (2021*b*). Salient object detection in the deep learning era: An in-depth survey. *IEEE T-PAMI*.

430. **Wang, W.**, **J. Shen**, **X. Dong**, and **A. Borji**, Salient object detection driven by fixation prediction. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018*c*.

431. **Wang, X.**, **W. Shi**, **R. Kim**, **Y. Oh**, **S. Yang**, **J. Zhang**, and **Z. Yu**, Persuasion for good: Towards a personalized persuasive dialogue system for social good. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2019.

432. **Wang, Y.**, **K. Li**, **Y. Li**, **Y. He**, **B. Huang**, **Z. Zhao**, **H. Zhang**, **J. Xu**, **Y. Liu**, **Z. Wang**, **S. Xing**, **G. Chen**, **J. Pan**, **J. Yu**, **Y. Wang**, **L. Wang**, and **Y. Qiao** (2022). Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

433. **Waugh, N. C.** and **D. A. Norman** (1965). Primary memory. *Psychological review*, **72**(2), 89.

434. **Wegener, D. T.**, **R. E. Petty**, **B. T. Detweiler-Bedell**, and **W. B. G. Jarvis** (2001). Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology*, **37**(1), 62–69.

435. **Wei, Z.**, **Y. Liu**, and **Y. Li**, Is this post persuasive? ranking argumentative comments in online forum. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2016.

436. **Wiedemann, N.**, **V. Wüest**, **A. Loquercio**, **M. Müller**, **D. Floreano**, and **D. Scaramuzza** (2023). Training efficient controllers via analytic policy gradient.

437. **Wikidata contributors** (ongoing). Wikidata. https://www.wikidata.org/.

438. **Wood, W.** (2000). Attitude change: Persuasion and social influence. *Annual review of psychology*, **51**(1), 539–570.

439. **Wu, C.-Y.** and **P. Krahenbuhl**, Towards long-form video understanding. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

440. **Wu, J.**, **L. Ouyang**, **D. M. Ziegler**, **N. Stiennon**, **R. Lowe**, **J. Leike**, and **P. Christiano** (2021). Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

441. **Wu, X.**, **H. Sun**, **J. Xue**, **R. Zhai**, **X. Kong**, **J. Nie**, and **L. He** (2023*a*). emotions: A large-scale dataset for emotion recognition in short videos.

442. **Wu, X.**, **K. Sun**, **F. Zhu**, **R. Zhao**, and **H. Li** (2023*b*). Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*.

443. **Wu, X.**, **K. Sun**, **F. Zhu**, **R. Zhao**, and **H. Li**, Human preference score: Better aligning text-to-image models with human preference. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023*c*.

444. **Xia, M.**, **E. Kochmar**, and **T. Briscoe**, Automatic learner summary assessment for reading comprehension. *In NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019.

445. **Xiao, F.**, **K. Kundu**, **J. Tighe**, and **D. Modolo**, Hierarchical self-supervised representation learning for movie understanding. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

446. **Xiong, C.**, **V. Zhong**, and **R. Socher**, Dynamic coattention networks for question answering. *In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rJeKjwvclx.

447. **Xu, B.**, **Y. Fu**, **Y.-G. Jiang**, **B. Li**, and **L. Sigal** (2016*a*). Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Transactions on Affective Computing*, **9**(2), 255–270.

448. **Xu, H.**, **J. Zhang**, **J. Cai**, **H. Rezatofighi**, and **D. Tao** (2022*a*). Gmflow: Learning optical flow via global matching.

449. **Xu, J.**, **X. Liu**, **Y. Wu**, **Y. Tong**, **Q. Li**, **M. Ding**, **J. Tang**, and **Y. Dong** (2024). Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, **36**.

450. **Xu, J.**, **T. Mei**, **T. Yao**, and **Y. Rui**, Msr-vtt: A large video description dataset for bridging video and language. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016*b*.

451. **Xu, L.**, **Z. Wang**, **B. Wu**, and **S. Lui**, Mdan: Multi-level dependent attention network for visual emotion analysis. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022*b*.

452. **Yang, A.**, **A. Miech**, **J. Sivic**, **I. Laptev**, and **C. Schmid** (2022). Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, **35**, 124–141.

453. **Yang, D.**, **J. Chen**, **Z. Yang**, **D. Jurafsky**, and **E. Hovy**, Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. *In NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019.

454. **Yang, J.**, **Q. Huang**, **T. Ding**, **D. Lischinski**, **D. Cohen-Or**, and **H. Huang**, Emoset: A large-scale visual emotion dataset with rich attributes. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

455. **Yang, J.**, **J. Li**, **X. Wang**, **Y. Ding**, and **X. Gao** (2021). Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing*, **30**, 7432–7445.

456. **Yang, L.**, **K. Tang**, **J. Yang**, and **L.-J. Li**, Dense captioning with joint inference and visual context. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

457. **Ye, Y.**, **B. Hui**, **M. Yang**, **B. Li**, **F. Huang**, and **Y. Li** (2023). Large language models are versatile decomposers: Decompose evidence and questions for table-based reasoning. *arXiv preprint arXiv:2301.13808*.

458. **Yu, Y.**, **J. Choi**, **Y. Kim**, **K. Yoo**, **S.-H. Lee**, and **G. Kim**, Supervising neural attention models for video captioning by human gaze data. *In IEEE CVPR*. 2017.

459. **Zhang, A.**, **B. Culbertson**, and **P. Paritosh**, Characterizing online discussion using coarse discourse sequences. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 11. 2017.

460. **Zhang, B. H.**, **B. Lemoine**, and **M. Mitchell**, Mitigating unwanted biases with adversarial learning. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018*a*.

461. **Zhang, C.**, **P. Benz**, **C. Lin**, **A. Karjauv**, **J. Wu**, and **I. S. Kweon** (2021). A survey on universal adversarial attack. *arXiv preprint arXiv:2103.01498*.

462. **Zhang, F.**, **D. Litman**, and **K. Forbes-Riley**, Inferring discourse relations from pdtb-style discourse labels for argumentative revision classification. *In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016.

463. **Zhang, H.**, **X. Li**, and **L. Bing** (2023*a*). Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

464. **Zhang, L.**, **S. Wang**, and **B. Liu** (2018*b*). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **8**(4), e1253.

465. **Zhang, R.**, **J. Han**, **C. Liu**, **A. Zhou**, **P. Lu**, **H. Li**, **P. Gao**, and **Y. Qiao**, Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. *In The Twelfth International Conference on Learning Representations*. 2023*b*.

466. **Zhang, Y.**, **X. Huang**, **J. Ma**, **Z. Li**, **Z. Luo**, **Y. Xie**, **Y. Qin**, **T. Luo**, **Y. Li**, **S. Liu**, *et al.* (2023*c*). Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*.

467. **Zhang, Z.**, **L. Wang**, and **J. Yang**, Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023*d*.

468. **Zhao, S.**, **Y. Ma**, **Y. Gu**, **J. Yang**, **T. Xing**, **P. Xu**, **R. Hu**, **H. Chai**, and **K. Keutzer**, An end-to-end visual-audio attention network for emotion recognition in user-generated videos. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34. 2020.

469. **Zheng, L.**, **W.-L. Chiang**, **Y. Sheng**, **S. Zhuang**, **Z. Wu**, **Y. Zhuang**, **Z. Lin**, **Z. Li**, **D. Li**, **E. Xing**, **H. Zhang**, **J. E. Gonzalez**, and **I. Stoica**, Judging LLM-as-a-judge with MT-bench and chatbot arena. *In Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2023. URL https://openreview.net/forum?id=uccHPGDlao.

470. **Zheng, L.**, **W.-L. Chiang**, **Y. Sheng**, **S. Zhuang**, **Z. Wu**, **Y. Zhuang**, **Z. Lin**, **Z. Li**, **D. Li**, **E. Xing**, *et al.* (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, **36**.

471. **Zhu, D.**, **J. Chen**, **X. Shen**, **X. Li**, and **M. Elhoseiny** (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

472. **Ziegler, D. M.**, **N. Stiennon**, **J. Wu**, **T. B. Brown**, **A. Radford**, **D. Amodei**, **P. Christiano**, and **G. Irving** (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.