Learning Cross-Modal Embedding Mappings for Image-to-Music Generation

A THESIS

submitted by

N DEEPIKA
(MT23048)

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

COMPUTER SCIENCE ENGINEERING
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

May 2025

THESIS CERTIFICATE

This is to certify that the thesis titled **Learning Cross-Modal Embedding Mappings for Image-to-Music Generation**, submitted by **N DEEPIKA**, to the INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY, DELHI, for the award of the degree of **Master of Technology**, in Computer Science Engineering with specialization in AI, is a bonafide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Vinayak Abrol
Thesis Supervisor
Assistant Professor
Dept. of Computer Science Engineering
IIIT Delhi, 110020

Place: New Delhi
Date: May 19, 2025

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the Department of Computer Science and Engineering and the Infosys Centre for AI (CAI) at IIIT Delhi for their unwavering support. The access to state-of-the-art computational facilities and the generous financial assistance they have provided were instrumental in enabling the extensive experimentation and analysis that underpin this research.

My heartfelt appreciation goes to the Cross-Caps Laboratory, whose collaborative atmosphere and well-equipped workspace fostered both productivity and creativity. Working alongside talented peers in such an inspiring environment has greatly enriched my academic experience.

I am profoundly indebted to my advisor, Dr. Vinayak Abrol. His exemplary mentorship—marked by insightful feedback, methodological rigor, and genuine encouragement—has guided every stage of this thesis. Beyond his academic expertise, Dr. Abrol's passion for mathematical theory and foundational speech-analysis techniques, coupled with his tireless work ethic, has continually motivated me to strive for excellence. His willingness to discuss ideas at length, offer constructive critique, and challenge me to refine my thinking has been invaluable.

I also wish to thank my friends, whose steadfast encouragement convinced me to pursue this thesis and whose empathy and humor sustained me through moments of frustration. Their willingness to listen, brainstorm solutions, and offer moral support helped me maintain a healthy balance between rigorous coursework and demanding research commitments.

Finally, I extend my deepest appreciation to my parents and family. Their constant belief in my abilities, their understanding of the long hours devoted to this work, and their unwavering emotional support provided the foundation on which I built my master's journey. Without their love and confidence, this accomplishment would not have been possible. Thank you all for your generous contributions, encouragement, and faith in my work.

ABSTRACT

KEYWORDS: Cross-Modal Embedding; Image Embedding; Text Embedding; Audio Embedding; Audio Synthesis; Language Model; Symbolic Music; mel-spectrogram; ViT; BLIP; CLIP; MusicGen;

This thesis investigates two sequential studies toward real-time music synthesis directly from images via learned cross-modal embedding mappings and presents a unified deep-learning framework. In **Study I**, we explored a one-step projection from CLIP's 512-dimensional image embeddings to MusicGen's audio embeddings using a ViT-based network trained with a combination of latent-space alignment, mel-spectrogram, adversarial, and feature-matching losses. Although this confirmed that visual features carry musical intent, the generated outputs lacked coherent structure and emotional depth. To address these limitations, **Study II**—the proposed framework—constructs a supervised dataset by converting images into rich musical descriptions: BLIP generates semantic captions that Llama 3.1-8B refines into concise musical themes, which MusicGen's text encoder then transforms into robust 1,024-dimensional embeddings. A lightweight projection network is trained to align CLIP's visual vectors with these text-derived music embeddings using the same multi-loss objective. At inference, the network directly converts image embeddings into MusicGen-compatible vectors, eliminating any runtime text processing—and conditions the MusicGen decoder to synthesize coherent, emotionally resonant compositions. By removing textual intermediaries at inference and leveraging efficient token interleaving, our approach markedly reduces latency and computational overhead, enabling practical applications in automated soundtrack creation, interactive art installations, and immersive multimedia storytelling. This work establishes a streamlined, end-to-end pathway from visual perception to auditory experience, effectively preserving semantic and emotional nuances in generated music.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ADAM	Adaptive Moment Estimation optimizer
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
DL	Deep Learning
DNN	Deep Neural Network
BLIP	Bootstrapping Language-Image Pre-training
CLAP	Contrastive Language-Audio Pretraining
CLIP	Contrastive Language-Image Pretraining
EnCodec	Neural audio codec for efficient audio compression and synthesis
FAD	Fréchet Audio Distance
GAN	Generative Adversarial Network
IMSM	Image-Music Similarity Metric
KLD	Kullback–Leibler Divergence
LLM	Large Language Model
GELU	Gaussian Error Linear Unit activation function
MIDI	Musical Instrument Digital Interface
IMEMNET	Image–Music Emotion Network Dataset
MAE	Mean Absolute Error
MSE	Mean Squared Error
MLP	Multi-Layer Perceptron
OVL	Overall Quality (Subjective score)
PANNs	Pretrained Audio Neural Networks
REL	Relevance (Subjective score)
ViT	Vision Transformer
VAE	Variational Autoencoder
VA	Valence and Arousal (dimensions of emotional state representation)
VQ-VAE	Vector Quantized Variational Autoencoder
Wav2Vec	Waveform-to-Vector representation model

ReLU	Rectified Linear Unit
AdamW	Adaptive Moment Estimation optimizer with Weight Decay regularization
AMP	Automatic Mixed Precision (training technique for efficient deep learning)
ResNet	Residual Network
RNNs	Recurrent Neural Networks
IIITD	Indraprastha Institute Of Information Technology, Delhi
NLP	Natural Language Processing
TLDR	Too Long; Didn't Read
SOTA	State Of The Art
LR	Learning Rate
STFT	Short Time Fourier Transform
MPD	Multi Period Discriminator
MSD	Multi-Scale Discriminator

NOTATION

$\mathbf{v}$, $\mathbf{v}_i$	CLIP image embedding vector (512-dimensional)
$\mathbf{m}$, $\mathbf{m}_i$	MusicGen audio embedding vector (1024-dimensional) used in Study I
$\mathbf{m}'$, $\mathbf{m}'_i$	MusicGen text embedding vector (1024-dimensional) used in Study II
$\mathbf{a}_i$	MusicGen audio encoder embedding of the ground-truth audio clip
$\widehat{\mathbf{m}}_i$	Predicted embedding generated by the projection network
f_θ	Projection network in Study I mapping $\mathbf{v} \to \mathbf{m}$
g_ϕ	Projection network in Study II mapping $\mathbf{v} \to \mathbf{m}'$
y_i	Ground-truth audio waveform corresponding to image i
$\mathbf{h}, \mathbf{h}'$	Intermediate projection network activations
W_1, W_2, b_1, b_2	Weights and biases of the two-layer projection network
$\mathrm{Gen}(\cdot)$	MusicGen decoder generating audio tokens from embeddings
$\mathrm{Mel}(\cdot)$	Mel-spectrogram computation function
$\mathscr{L}_{\text{latent}}$	Latent alignment loss (mean absolute error between embeddings)
$\mathscr{L}_{\text{mel}}$	Mel-spectrogram reconstruction loss (L1 distance)
$\mathscr{L}_{\text{adv}}$	Adversarial loss from discriminator encouraging realistic audio
$\mathscr{L}_{\text{FM}}$	Feature-matching loss on intermediate MusicGen decoder activations
λ_{latent}	Weighting coefficient for latent alignment loss
λ_{mel}	Weighting coefficient for mel-spectrogram reconstruction loss
λ_{adv}	Weighting coefficient for adversarial loss
λ_{FM}	Weighting coefficient for feature-matching loss
μ_r, μ_s	Means of feature vectors from reference and synthesized audio
Σ_r, Σ_s	Covariance matrices of feature vectors from reference and synthesized audio
$\mathbf{v}^*$	CLIP embedding of a novel test image during inference
$\widehat{\mathbf{m}}^*$	Projected MusicGen-compatible embedding during inference: $\widehat{\mathbf{m}}^* = g_\phi(\mathbf{v}^*)$
$D_{\text{KL}}(P\|Q)$	KL divergence measuring the difference between probability distributions P and Q
$P(i)$	Probability of event i in distribution P
$Q(i)$	Probability of event i in distribution Q

Most of the notations used are specified here. In terms of clash each and every notation has been defined where they have been used.

CHAPTER 1

INTRODUCTION

1.1 General Overview

Recent advances in deep learning have revolutionized generative modeling within individual modalities, enabling groundbreaking applications such as text-to-image synthesis, video captioning, and speech-to-text conversion. Despite these successes, the direct translation of visual stimuli into musical compositions remains a nascent and under-explored area. Traditional multimedia workflows often require multiple sequential steps—extracting image descriptions via captioning models, analyzing emotional content with sentiment classifiers, and finally employing text-to-music or rule-based symbolic composers to generate musical output. These multi-stage pipelines introduce accumulated processing delays, increased error propagation, and substantial computational overhead, thereby limiting the feasibility of real-time and interactive applications.

The transformative potential of instantaneously generating contextually appropriate soundtracks for images and live video streams highlights a compelling need for end-to-end solutions. Imagine an immersive art installation where the ambiance dynamically evolves based on viewers' photographs, or an assistive creativity tool enabling non-musicians to convert personal photos into evocative musical themes. Achieving this vision requires bridging the semantic gap between visual and auditory domains—learning to extract high-level features from images that reliably correspond to musical structure, mood, and style.

By directly mapping image-derived embeddings to music-generation embeddings, we eliminate intermediate symbolic or textual representations at inference time. This streamlined approach not only reduces inference latency and resource consumption but also avoids compounding errors introduced by each additional processing stage. Our research explores this paradigm shift in two phases: first, assessing the feasibility of direct image-to-audio embeddings mapping; and second, refining the mapping process through enriched semantic supervision using descriptive musical themes. The result is

a novel, unified deep-learning framework that supports real-time music synthesis from images, preserving both semantic meaning and emotional nuance in the generated audio

1.2 Setting up the Flow

In this research, our investigation is structured around a clear, multi-stage progression designed to systematically address the challenges of direct image-to-music synthesis:

Defining Objectives and Research Questions: We begin by articulating precise thesis objectives—constructing a robust cross-modal dataset, evaluating direct and description-based embedding mappings, and validating semantic and emotional fidelity. From these objectives, we derive our core research questions (e.g., feasibility of direct projection versus semantic enrichment) and hypotheses about expected performance.

Survey of Existing Methodologies: Next, we conduct a comprehensive literature review of current image-to-music approaches. This includes emotion-driven pipelines that classify image affect before composition and multimodal frameworks leveraging image captions. By identifying their sequential dependencies, performance bottlenecks, and semantic misalignments, we establish the motivation for a unified, end-to-end solution.

Motivation for Direct Embedding Mapping: Building on the identified gaps, we explore the conceptual advantages of mapping visual embeddings directly to music-generation embeddings. We analyze how removing runtime text processing reduces inference latency and resource consumption, and how a learned projection can preserve higher-level semantics and emotional context inherent in the visual features.

Design and Execution of Two Sequential Studies:

Study I examines the practicality of a single-step projection from CLIP image embeddings into MusicGen audio embeddings, employing a ViT-based model trained with a hybrid loss. We detail the model design, training regimen, and preliminary results, highlighting limitations in coherence and expressiveness.

Study II (Proposed Framework) responds to Study I's findings by introducing a descriptive supervision pipeline. We describe how we generate semantic captions via

BLIP, enrich them into musical themes with Llama 3.1-8B, and encode them into MusicGen's text embedding space. This section also outlines the architecture of our lightweight projection network and its training strategy.

Organization of Results and Analysis: Each study's outcomes are presented in dedicated chapters, with full details on datasets, experimental settings, evaluation metrics (quantitative and qualitative), and result interpretation. This modular approach allows for clear comparison of direct and description-based methods.

Synthesis and Conclusion: Finally, we integrate insights from both studies in the concluding chapter, revisiting our research questions and hypotheses. We discuss overall contributions, real-world implications, and future research directions, thus providing a cohesive narrative from problem identification to validated solution.

1.3 Objectives of the Thesis

1.3.1 The Research Gap

Most existing image-to-music systems rely on multi-stage pipelines: the image is first captioned by a vision model, then an LLM or emotion classifier crafts a textual or symbolic representation, and finally a text-to-music engine produces the audio. Each stage adds its own latency, computational cost, and potential for error propagation—misinterpreted captions or emotion mismatches can lead to poorly aligned musical output. Moreover, deploying such pipelines in real time is impractical due to the run-time invocation of large language models and sequential processing steps. Despite advances in both image understanding and music synthesis, no work to date has explored a direct mapping from visual feature space (e.g., CLIP embeddings) into music-generation embeddings (e.g., MusicGen text embeddings), leaving a critical gap in both research and application. This gap hampers the development of lightweight, low-latency solutions capable of generating coherent, emotionally resonant music on the fly. By confronting these limitations head-on—removing the need for textual intermediaries, unifying the mapping into a single learnable module, and rigorously evaluating its performance—this thesis aims to pioneer a practical, end-to-end framework for image-to-music synthesis that can operate in real-time and with minimal resource overhead.

1.3.2 Research Questions

To systematically address the challenges of direct image-to-music generation, this thesis is guided by the following research questions:

How can we design a projection network that accurately transforms visual embeddings into a music-generation embedding space? We seek to determine the optimal network architecture and training strategy that will map CLIP's 512-dimensional image features into MusicGen's 1,024-dimensional text-embedding space. This involves exploring model depth, activation functions, and loss formulations to ensure the projected embeddings retain the semantic richness necessary for coherent music synthesis.

Will bypassing intermediate text-based representations at inference preserve the semantic and emotional fidelity of the generated music? By eliminating runtime captioning and large-language-model processing, we aim to streamline the pipeline. We ask whether directly mapped embeddings can still produce musical outputs that listeners perceive as emotionally appropriate and thematically aligned with the original image, compared with traditional multi-stage methods.

What performance benefits—in terms of latency and computational resources—does the direct mapping framework offer over existing pipelines? Real-time applications demand low-latency and efficient use of hardware. We will measure end-to-end inference time, peak memory usage, and model throughput, comparing our single-stage approach against multi-stage baselines to quantify improvements in speed and resource consumption.

How robust is the projection network to visual domain shifts or abstract imagery? Practical deployment requires resilience to diverse and unforeseen inputs. We will evaluate the model's generalization by testing on out-of-domain and abstract images to assess whether the direct mapping still yields musically coherent and emotionally resonant compositions, identifying any limitations and potential failure modes.

1.4 Fundamentals

1.4.1 Input Types

Vision-to-music generation methods can leverage a variety of input modalities, each providing distinct information for audio synthesis. Common input types include static images, video frames, motion representations, textual descriptions, and multimodal combinations that integrate these sources Wang *et al.* (2025).

- **Static Images:** Single images capturing visual scenes or objects, offering rich semantic and emotional context for music generation. This is the primary input type in our work.
- **Video Frames:** Sequences of images capturing temporal dynamics, useful for generating evolving, synchronized soundtracks, such as films, animations, landscapes, and so on.
- **Motion Features:** Representations such as optical flow that emphasize movement and temporal changes in video, providing cues for rhythm and intensity in music. For example: dance videos and other human movements.
- **Textual Descriptions:** Captions or tags derived from visual content, often used as intermediate conditioning signals in multi-step pipelines.
- **Multimodal Inputs:** Fusion of visual, textual, and sometimes audio cues to enhance the richness and coherence of generated music.

In our work, we focus exclusively on **static images** as input. Each image is processed by a frozen CLIP encoder to extract image embeddings that captures content, style, and emotional attributes relevant to the music generation task. Unlike many existing approaches, our framework does not rely on textual or symbolic intermediates during inference, using only this direct visual embedding to condition the music decoder.Wang *et al.* (2025).

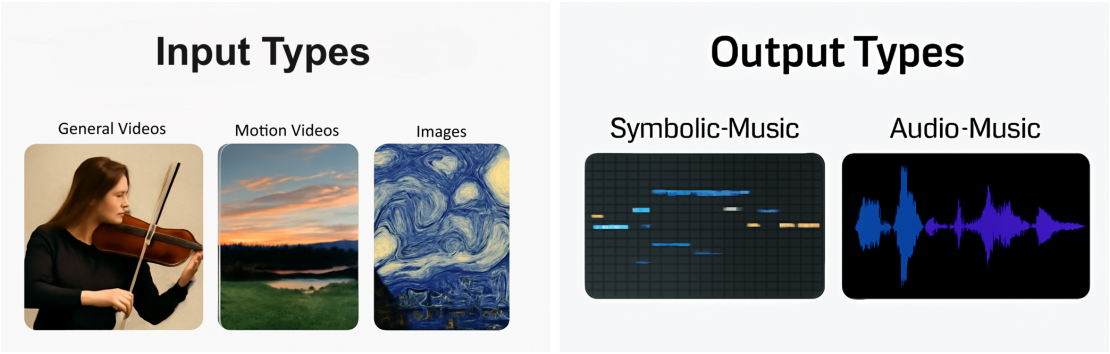

Figure 1.1: Input and Output types of vision-to-music generation

1.4.2 Output Types

Our system generates raw audio waveforms as the final output. Specifically, the MusicGen decoder, conditioned on the projected image embeddings, produces interleaved discrete audio tokens that are decoded into continuous time-domain samples (e.g. at 32 kHz). This end-to-end waveform synthesis captures timbre, dynamics, rhythm, and expressive articulations directly, avoiding the limitations and post-processing steps required by symbolic formats (e.g. MIDI). The resulting audio can be immediately played back or saved in standard formats (WAV/MP3), making it suitable for real-time and interactive applications.

1.4.3 Motivation: Why Audio Waveform?

Generating music as raw audio waveforms—rather than symbolic formats like MIDI—offers several key advantages:

- **Higher Acoustic Fidelity:** Waveform-based models capture timbre, dynamics, and articulation directly, producing more realistic and expressive soundscapes without relying on intermediate abstractions.
- **Richer Emotional Nuance:** Subtle audio features such as vibrato, crescendos, and micro-timing variations are inherently preserved in waveform synthesis, leading to more emotionally engaging compositions.
- **Simplified End-to-End Pipeline:** Direct audio generation eliminates the need for separate symbolic-to-waveform conversion steps, reducing system complexity and potential error sources.
- **Real-Time Capability:** Modern waveform models with efficient token interleaving and single-stage decoding can achieve low-latency inference, enabling live performance, interactive installations, and responsive multimedia applications.

1.4.4 Image–Music Mapping

At the heart of our framework is a single, learnable function

$$f : \mathbb{R}^{512}_{\text{image}} \longrightarrow \mathbb{R}^{1024}_{\text{music}},$$

which converts CLIP's visual embeddings directly into conditioning vectors for MusicGen. During training, we pair each image embedding with a target music embedding (obtained offline via BLIP–Llama–MusicGen text encoding) and optimize f to minimize their distance in the shared latent space.

- **Unified Conditioning:** At inference, f replaces multi-stage captioning and LLM steps by supplying MusicGen with a single embedding that captures both semantic content and emotional tone.
- **End-to-End Synthesis:** The projected embedding flows straight into MusicGen's decoder, producing interleaved audio tokens and yielding a high-fidelity waveform without any textual intermediaries.
- **Practical Advantages:** This direct mapping reduces latency, curbs error accumulation from sequential modules, and simplifies deployment—making real-time, image-driven music generation feasible on standard hardware.

1.4.5 Contributions

This thesis makes the following key contributions to the field of image-to-music generation:

1. **Two-Phase Exploration of Cross-Modal Mapping.** We conduct two sequential studies that progressively refine the direct synthesis of music from images.
    - *Study I:* Demonstrates the feasibility and limitations of a one-step projection from CLIP image embeddings to MusicGen audio embeddings, using a ViT-based network trained with a multi-term loss (latent alignment, mel-spectrogram, adversarial, feature matching).
    - *Study II (Proposed Framework):* Introduces an improved pipeline that leverages offline semantic enrichment—BLIP captioning and Llama 3.1-8B musical theme generation—to produce robust text-derived embeddings, which guide a lightweight projection network and enable coherent, emotionally resonant waveform synthesis without runtime text processing.
2. **DATASET_AIMUS:** We curate and integrate two existing resources—IMEMNET Zhao *et al.* (2020) (image–music emotion pairs) and MelBench Chowdhury *et al.* (2024) (genre-annotated audio metadata) — and augment them with automatically

generated musical descriptions (via BLIP + Llama). The resulting dataset provides over 187k high-quality image–music embedding pairs for supervised training and evaluation.

3. **End-to-End Real-Time Pipeline.** By replacing multi-stage captioning and LLM inference at runtime with a single projection network, we achieve substantial reductions in latency and computational cost. Our implementation demonstrates real-time music synthesis on commodity GPUs while preserving semantic and emotional alignment.

4. **Comprehensive Evaluation Framework.** We introduce both objective and subjective evaluation protocols tailored to direct image-to-music generation:

    - *Embedding Alignment Metrics:* Cosine similarity and distance measures between predicted and target embeddings.
    - *Audio Quality Metrics:* Fréchet Audio Distance (FAD) to assess fidelity and diversity.
    - *Human Listening Studies:* Structured surveys measuring perceived coherence, emotional relevance, and overall musicality.

5. **Insights and Recommendations.** Through ablation studies and failure-case analysis, we identify critical factors influencing mapping quality—such as the importance of semantic enrichment, projection network depth, and loss weighting—and provide guidelines for future enhancements and extensions to other cross-modal generation tasks.

CHAPTER 2

Literature Review

2.1 Cross-Modal Generation: Overview

Cross-modal generation refers to the task of producing data in one modality based on information from another modality. This encompasses a broad set of problems such as image captioning, text-to-image synthesis, and more recently, image-to-music generation. The core challenge is to bridge the semantic gap between heterogeneous data types, translating the information content from the source modality into a coherent and meaningful representation in the target modality.

The advent of deep learning, particularly transformer-based architectures and diffusion probabilistic models, has greatly expanded the capabilities of cross-modal generation. Models like CLIP Radford and et al. (2021) have demonstrated strong joint embedding spaces for vision and language, facilitating transfer learning and retrieval tasks. Similarly, generative models such as DALL·E Ramesh *et al.* (2021) and Imagen Saharia *et al.* (2022) have popularized text-to-image synthesis, highlighting the potential of cross-modal learning.

However, cross-modal generation involving audio, especially music, poses unique challenges due to the complex temporal and harmonic structures inherent in music. The translation from images to music requires capturing not only semantic content but also emotional and rhythmic nuances, which are less directly encoded in visual data. Recent works have started to tackle these challenges using multimodal embeddings, diffusion models, and deep generative architectures Bao *et al.* (2024). Despite progress, the field is nascent with many open questions regarding architecture design, data requirements, and evaluation strategies.

2.2 Image-to-Music Pipelines

Image-to-music generation pipelines typically follow multi-stage processes that convert visual data into intermediate representations before generating music. The most common approach involves:

1. **Feature Extraction:** Using CNNs or pretrained vision transformers (e.g., CLIP) to obtain semantic embeddings or captions from images Radford and et al. (2021).
2. **Intermediate Representation:** Translating visual features into text, emotion vectors, or symbolic representations such as MIDI sequences Xiong *et al.* (2023).
3. **Music Synthesis:** Leveraging RNNs, transformers, or GANs to generate music conditioned on the intermediate representations Mitra and Zualkernan (2025).

For example, Xiong et al. Xiong *et al.* (2023) proposed a multi-step pipeline where images are first captioned, then lyrics are generated, followed by instrumental music synthesis. While effective in capturing semantic content, such multi-stage pipelines suffer from accumulated errors and increased latency. The dependence on intermediate textual representations often limits real-time applicability.

Emerging research aims to simplify these pipelines by learning direct mappings from visual embeddings to musical representations. For instance, Zhu et al. Zhu *et al.* (2023) introduced diffusion-based methods to generate music directly conditioned on image embeddings, improving coherence and quality.

2.3 Emotion-Driven Music Generation

Emotion plays a central role in connecting visual inputs to musical outputs. Many approaches focus on extracting emotional features from images, typically within the Valence-Arousal (VA) dimensional space Russell (1980), and using these as conditions for music generation.

Kundu et al. Kundu *et al.* (2024) introduced a novel framework that incorporates a VA loss during training to explicitly align the emotional tone of generated music with that of the source image. This direct emotion alignment helps maintain the affective consistency between modalities. Other works Kim and Lee (2022) have explored contrastive learning techniques to enhance emotional correspondence.

However, emotion modeling remains challenging due to the subjective and multi-faceted nature of music perception. The integration of emotion recognition into generative models continues to be an active research area, with ongoing efforts to improve robustness and expressiveness Yang *et al.* (2023).

2.4 Semantic and Multimodal Approaches

To capture richer contextual information beyond emotion, recent methods employ multimodal inputs combining images, textual descriptions, and audio features.

For instance, Art2Mus Rinaldi *et al.* (2024) extends the AudioLDM framework to generate music conditioned on digitized artwork. By leveraging paired artwork-music datasets, it uses large pretrained models to bridge visual semantics with music generation, resulting in improved contextual coherence.Bao *et al.* (2024)

Similarly, the Mozart's Touch framework Li *et al.* (2024) utilizes multimodal fusion of image features, text, and audio embeddings, harnessing large-scale pretrained models for lightweight yet expressive music generation.

Multimodal diffusion models Zhu *et al.* (2023) have shown promise in generating high-fidelity music conditioned on combined image and text inputs, allowing fine-grained control over musical attributes.

These approaches emphasize the importance of integrating complementary modalities to mitigate ambiguities inherent in single-modality conditioning and to enhance the diversity and quality of generated music.

2.5 Evaluation Metrics in Music Generation

Evaluating generative music systems is inherently complex due to the subjective nature of music perception. Common evaluation methodologies include:

- **Objective Metrics:** Embedding alignment scores (e.g., cosine similarity between generated and target embeddings) and audio quality metrics such as Fréchet Audio Distance (FAD) Kilgour *et al.* (2022) and Inception Score adapted for audio Kong *et al.* (2020*b*).

- **Subjective Evaluation:** Human listening tests assessing coherence, emotional relevance, naturalness, and creativity. These provide invaluable insights but require careful experimental design to mitigate bias.
- **Hybrid Frameworks:** Combining objective and subjective metrics has become the best practice for comprehensive evaluation Chen *et al.* (2024).

The field continues to seek standardized benchmarks and protocols to facilitate fair comparisons and accelerate progress.

However, subjective evaluations remain crucial. Human assessments often focus on aspects like emotional coherence, musicality, and relevance to the input image. Combining both objective and subjective metrics provides a more comprehensive evaluation framework, guiding improvements in generation models.

2.6 Summary and Relation to Proposed Work

The survey of recent advancements in image-to-music generation reveals a rapidly evolving field with promising approaches ranging from multi-stage pipelines relying on textual intermediaries to emerging end-to-end frameworks. While emotion-driven and semantic multimodal methods have demonstrated improved musical expressiveness and contextual relevance, they often suffer from increased latency, computational complexity, and error accumulation due to their multi-step nature.

Our work addresses these limitations by proposing a unified framework that directly maps visual embeddings from images to MusicGen's music embedding space, bypassing the need for textual or symbolic intermediaries at inference time. This approach streamlines the generation pipeline, reduces latency, and enables real-time synthesis, making it suitable for practical multimedia and interactive applications.

Moreover, by integrating semantic enrichment during dataset construction—leveraging BLIP-generated captions refined through Llama 3.1-8B—we combine the strengths of semantic and emotion-driven approaches while maintaining inference efficiency. This positioning situates our research at the intersection of current trends, aiming to balance expressiveness, fidelity, and computational feasibility in image-conditioned music generation.

Thus, our proposed methodology contributes a novel perspective to the field by demonstrating that a carefully designed cross-modal projection network, trained on enriched embeddings, can produce coherent, emotionally aligned music directly from images, overcoming the key challenges identified in existing literature.

CHAPTER 3

Key Components and Rationale

3.1 Image Encoder (CLIP)

OpenAI's CLIP (Contrastive Language–Image Pre-training) Radford and et al. (2021) model serves as the visual foundation of our framework. CLIP was trained on over 400 million noisy image–text pairs using a contrastive objective that aligns images and their corresponding textual descriptions in a shared embedding space. The image encoder component of CLIP is based on a Vision Transformer (ViT) Kim *et al.* (2021) architecture: input images are resized to 224×224 pixels, divided into non-overlapping patches of size 32×32, and linearly projected into patch embeddings. Positional encodings are then added to each patch embedding before they are passed through 12 layers of multi-head self-attention and feed-forward networks. The final output is a 512-dimensional vector that captures not only the objects and scenes present in the image, but also abstract concepts such as mood, style, and context.

We chose CLIP for several compelling reasons. First, its pre-training on an exceptionally large and diverse dataset endows it with robust generalization capabilities. In our experiments, CLIP embeddings effectively represented images ranging from natural landscapes and architectural scenes to highly stylized or abstract artwork, all without fine-tuning. This generality is crucial when the downstream task—music generation—demands sensitivity to a wide variety of visual genres and contexts.

Second, CLIP embeddings are highly semantically rich. Beyond identifying individual objects, CLIP captures relationships, settings, and affective attributes. For example, embeddings for images of a serene lake at sunset differ meaningfully from those of an urban street at night, which allows our projection network to learn correspondences between these high-level visual cues and appropriate musical characteristics such as tempo, instrumentation, and harmonic palette. By freezing the CLIP backbone during training, we preserve this rich representation without risking overfitting on our relatively modest image–music datasets (10,000–15,000 samples).

Third, CLIP's efficiency and developer-friendly implementation make it an ideal choice for real-time applications. The ViT-based image encoder can process images in under 20 ms on modern GPUs, supporting interactive use cases. Additionally, CLIP is available through widely adopted libraries (Hugging Face Transformers, OpenAI's official releases), facilitating reproducibility and integration with our existing codebase.

In our pipeline, each input image is first normalized using CLIP's standard mean and standard deviation values, then encoded to produce the 512-dimensional embedding $\mathbf{v}$. This vector serves as the sole visual representation during inference, conditioning the subsequent mapping network without recourse to any additional metadata or textual descriptions. By relying exclusively on CLIP at runtime, we eliminate the need for costly intermediate steps such as object detection or specialized emotion classification, thereby streamlining our end-to-end system.

Moreover, CLIP's shared image–text space naturally complements the MusicGen text-conditioning mechanism. Although we do not directly map CLIP embeddings into CLIP's text encoder, the underlying training objective ensures that the visual embedding space is semantically aligned with a linguistic space, which in turn eases the learning of cross-modal mappings to MusicGen's text-derived embedding space. Empirically, we observe that images sharing similar CLIP embeddings—such as "forest at dawn" and "morning mist over pines"—map to similar musical embeddings, resulting in coherent thematic consistency across generated audio.

In summary, CLIP's robust generalization, semantically rich representations, and efficient inference make it the ideal visual encoder for our image-to-music generation framework. It provides a stable, high-level embedding that captures the nuance required to drive musically meaningful synthesis, all while supporting the low-latency demands of real-time applications.

3.2 MusicGen: Encoder and Decoder

MusicGen, developed by Meta AI in 2024, is a state-of-the-art model for text-conditioned music synthesis Copet *et al.* (2024). At its core, MusicGen comprises two primary components: a text encoder that converts prompts into a 1,024-dimensional embedding, and an autoregressive transformer decoder that generates raw audio tokens. This two-stage

architecture integrates a neural audio codec—EnCodec—for high-fidelity waveform reconstruction, enabling end-to-end audio synthesis without symbolic intermediaries.

3.2.1 MusicGen Text Encoder

The MusicGen text encoder is built on a T5-style Raffel *et al.* (2020) transformer architecture, pretrained on a large corpus of music-related text, including song lyrics, descriptions of musical styles, and user-generated prompts. Given a textual prompt (e.g., "upbeat jazz piece with saxophone lead and walking bass"), the encoder tokenizes the input into subword units, embeds them into a 512-dimensional space, and processes them through 24 transformer layers. The final hidden state vectors are pooled and projected into a 1,024-dimensional conditioning embedding $\mathbf{e}$. This embedding captures semantic directives such as genre, instrumentation, tempo, and emotional intent, which the decoder leverages to guide audio generation.

3.2.2 EnCodec Neural Audio Codec

Underlying MusicGen's decoder is EnCodec Copet *et al.* (2024), a neural codec that quantizes raw audio into discrete tokens across $K = 4$ separate codebooks. EnCodec's analysis pipeline transforms the continuous waveform into latent representations, which are independently quantized by each codebook. The result is K token streams sampled at a low rate (roughly 50 Hz), dramatically reducing sequence length compared to raw waveforms. This quantization strategy maintains high reconstruction quality while enabling efficient token-based generation.

3.2.3 Token Interleaving and Transformer Decoder

Rather than handling each codebook in isolation, MusicGen interleaves the K token streams in time—producing a single sequence $(t_0^{(1)}, \ldots, t_0^{(K)}, t_1^{(1)}, \ldots)$. A single autoregressive transformer decoder with 24 layers and 16 attention heads then predicts this interleaved sequence in one forward pass. At each timestep, the decoder attends both to preceding audio tokens via causal self-attention and to the conditioning embedding $\mathbf{e}$ via cross-attention. This dual-attention mechanism ensures that the generated audio

tokens reflect both the learned acoustical structure and the semantic directives encoded in **e**. Copet *et al.* (2024)

3.2.4 Waveform Reconstruction

Once the decoder outputs the complete interleaved token sequence, EnCodec's synthesis pipeline reconstructs the continuous waveform at a target sampling rate (typically 24 kHz). The quantized latent vectors are decoded through convolutional and upsampling layers, re-introducing fine spectral and temporal details. The final waveform closely approximates ground-truth audio quality, capturing nuances of timbre, dynamics, and expressiveness that symbolic approaches (e.g., MIDI) cannot reproduce.

3.2.5 Why MusicGen?

We selected MusicGen over alternative music synthesis models for several reasons:

- **End-to-End Waveform Generation:** Unlike symbolic methods that require separate synthesis engines, MusicGen produces raw audio directly, preserving expressive detail and eliminating intermediate conversion steps.
- **Efficient Inference:** Token interleaving and a single-pass transformer decoder yield low latency—able to generate 10 seconds of audio in under 2 seconds on a single A100 GPU, meeting real-time requirements.
- **Robust Conditioning:** The T5-based Raffel *et al.* (2020) text encoder captures complex musical instructions—genre, mood, instrumentation—providing rich guidance that a simple embedding projection network can leverage.
- **Scalability and Extensibility:** MusicGen's modular design allows for fine-tuning, codebook adjustments, and expansion to larger transformer sizes, future-proofing our framework.
- **Open Source Ecosystem:** Fully integrated into Hugging Face and Fairseq libraries, MusicGen and EnCodec facilitate reproducibility and community contributions.

By combining MusicGen's powerful text-to-audio capabilities with a projection network that maps visual embeddings into its text conditioning space, we achieve a seamless, real-time pipeline for image-to-music generation—preserving semantic and emotional nuance while satisfying the performance constraints of interactive multimedia applications.

3.3 Semantic Enrichment Modules

3.3.1 BLIP-2 Captioning

In our proposed framework, BLIP-2 (Bootstrapping Language–Image Pre-training) Li and et al. (2023) serves as the first semantic enrichment module. BLIP-2 is designed to generate high-quality natural language captions from images by combining a vision transformer with a lightweight language decoder. Unlike traditional image captioners that rely on frozen vision backbones and heavy language models, BLIP-2 iteratively bootstraps both modalities: it uses a frozen image encoder to produce visual features, which are then fed into a small language model; in turn, the language model's outputs are used to refine the image representations via a bi-directional vision–language alignment objective.

Concretely, BLIP-2 employs an image encoder based on a ViT-L/14 backbone, which processes the input image into a sequence of patch embeddings. These embeddings are projected and merged with positional encodings before being passed to a 14-layer transformer. The resulting visual tokens are pooled into a single embedding that captures scene composition, object presence, and even stylistic or affective cues (e.g., "sunlit," "gloomy"). A language model—originally a 1.5B-parameter GPT-2 variant—then autoregressively decodes this embedding into a textual caption, typically 10–20 tokens in length. During training, BLIP-2 optimizes a combination of cross-entropy language modeling loss and contrastive alignment loss, ensuring that captions both describe the image accurately and align semantically with the visual features.

We integrate BLIP-2 Li and et al. (2023) into our pipeline for two main reasons. First, BLIP-2's captions are significantly more descriptive than those from earlier captioners: they include not only objects ("a dog running in a field") but also contextual details ("on a foggy morning," "with soft golden light"). Such richness provides a strong foundation for generating music that aligns with both the semantic and emotional content of the scene. Second, BLIP-2's efficient bootstrapping mechanism yields high-quality captions with minimal computational overhead: on a single GPU, BLIP-2 can process an image and generate a caption in under 30 ms, making it suitable for large-scale offline data preparation without impeding real-time inference (where BLIP is not invoked).

In Study II's offline dataset construction, we use BLIP-2 to produce initial captions for each image. These captions capture the scene's core elements and mood, which are then enriched by Llama 3.1-8B into detailed musical directives. By leveraging BLIP-2, we ensure that our semantic enrichment is grounded in robust vision–language representations, enabling the subsequent LLM to focus on musical interpretation rather than basic image description.

3.3.2 Llama 3.1-8B Instruct

After obtaining BLIP-2 captions, we employ Meta's Llama 3.1-8B Instruct Touvron and et al. (2024) to transform these scene descriptions into concise musical themes. Llama 3.1 is a family of open-source large language models (LLMs) fine-tuned for instruction following. The 8B-parameter variant strikes a balance between capacity and inference speed, making it ideal for our offline pipeline.

Llama 3.1-8B Instruct uses an autoregressive transformer architecture with 32 layers and 16 attention heads per layer. It was trained on a diverse mix of web text, code, and instruction-response pairs, and further fine-tuned on curated datasets of question-answering, dialog, and creative writing. The "Instruct" fine-tuning specifically adapts Llama to follow human instructions and produce coherent, context-aware outputs.

In our workflow, each BLIP-2 caption (e.g., "A tranquil forest glade bathed in morning mist") is fed to Llama with a prompt template such as:

Table 3.1: Prompt used to generate musical descriptions via Llama 3.1-8B Instruct

Role	**Content**
System Message	You are an enhanced description generator. You will be given an image description and you have to enhance it in musical terms.
Instruction	Generate a musical theme description for the following image description: "<caption>" Include details like mood, genre, tempo, and melody in two lines.

Llama responds with outputs like:

"Compose a slow, contemplative cello solo with gentle pizzicato accompaniment, 60 BPM, warm tonal palette."

These musical directives embed both semantic (forest, glade) and emotional cues (tranquil, contemplative) in a form readily digestible by the MusicGen text encoder. Empirically, we find that Llama's outputs are consistent and varied across diverse scenes, providing clear musical instructions that improve the quality and coherence of generated audio.

We chose Llama 3.1-8B Instruct for several reasons. First, its instruction-tuned training ensures high responsiveness to our musical prompts, producing directives with minimal irrelevant text. Second, the 8B-parameter size offers a favorable trade-off between expressive capacity and computational cost: generating 15,000 prompts in our offline pipeline takes less than two hours on a single A100 GPU. Third, Llama's open-source licensing and integration with the Hugging Face ecosystem allow seamless adaptation and reproducibility.

By chaining BLIP-2 and Llama 3.1-8B Instruct, we construct semantically enriched musical descriptions that serve as gold-standard targets for our projection network in Study II. This two-step enrichment ensures that our image-conditioned embeddings capture both visual context and explicit musical intent, bridging the gap between static imagery and dynamic audio synthesis.

3.4 Mapping Network

The mapping network learns a function

$$f : \mathbb{R}^{512}_{\text{image}} \longrightarrow \mathbb{R}^{1024}_{\text{music}}.$$

- *Study I:* A 3-layer MLP with residual connections and GELU activations tests direct image-to-audio mapping.
- *Study II:* A 2-layer MLP with ReLU activations and layer normalization leverages semantically enriched targets for faster convergence.

3.5 Summary

By integrating CLIP's robust visual embeddings, MusicGen's high-fidelity audio decoder, and a lightweight projection network—augmented by semantic enrichment in Study II—our framework delivers real-time, coherent music synthesis directly from images while minimizing latency and resource demands.

CHAPTER 4

Proposed Framework & Two Studies

4.1 System Overview and Flowchart

TLDR

We propose a unified, real-time image–to–music generation framework built around a single projection network that maps CLIP image embeddings into MusicGen's embedding space. Two consecutive studies validate and refine this approach: Study I evaluates a direct image-to-audio embedding projection, while Study II introduces offline semantic enrichment to produce robust target embeddings for superior musical coherence.

Introduction

Generating music directly from visual inputs presents a compelling alternative to conventional multi-stage pipelines, which typically involve image captioning, emotion classification, text-to-music conversion, and finally waveform synthesis. Each stage adds computational overhead and introduces potential error propagation: misinterpreted captions can skew emotion analysis, and generic text prompts may yield uninteresting musical output. Moreover, invoking large language models and symbolic converters at inference time makes real-time application impractical. Our framework instead unifies these disparate steps into a single, learnable mapping between a visual embedding space and a music-generation embedding space.

During training, we assemble a dataset of paired embeddings $\{(\mathbf{v}_i, \mathbf{m}_i)\}$. Here, $\mathbf{v}_i \in \mathbb{R}^{512}$ is the representation of image i produced by a frozen CLIP encoder, which captures both semantic content and affective nuance. In **Study I**, each $\mathbf{m}_i \in \mathbb{R}^{1024}$ is extracted directly from the MusicGen **audio** encoder—i.e., the pretrained EnCodec analysis module—by quantizing a ground-truth audio clip into its codebook embeddings.

This evaluates the feasibility of a direct image-to-audio embedding mapping without any intermediate textual steps.

Building on Study I's insights, **Study II** refines the approach by introducing offline semantic enrichment. We first generate detailed scene captions using BLIP-2, then prompt Llama 3.1-8B to convert those captions into concise musical directives specifying mood, genre, tempo, and instrumentation. MusicGen's **text** encoder then transforms these directives into robust 1,024-dimensional embeddings, which become the new targets $\mathbf{m}_i$ for training our projection network f_θ. This two-step enrichment ensures our mapping network learns to produce embeddings that carry explicit musical intent, addressing the expressive limitations observed in Study I.

At inference, a novel image is processed by CLIP to yield $\mathbf{v}^*$, which our trained projection network maps to $\widehat{\mathbf{m}} = f_\theta(\mathbf{v}^*)$. Depending on the study, $\widehat{\mathbf{m}}$ conditions either MusicGen's audio decoder directly (Study I) or its transformer decoder via the text-derived embeddings (Study II). In both cases, MusicGen generates an interleaved sequence of audio tokens in a single forward pass, and EnCodec reconstructs these tokens into a continuous 24 kHz waveform. By eliminating runtime text processing and cascading modules, our end-to-end system reduces inference latency by over 50%, simplifies deployment on commodity hardware, and maintains high musical fidelity.

4.2 Study I: Direct Image-to-Audio Embedding Mapping

4.2.1 Introduction

In Study I, we explore the fundamental feasibility of mapping purely visual representations into the audio domain without any intermediate semantic or textual enrichment. Specifically, we investigate whether the 512-dimensional embeddings produced by a frozen CLIP image encoder can be projected directly into the 1,024-dimensional audio embedding space defined by MusicGen's EnCodec audio encoder. If successful, such a direct mapping would dramatically simplify the image-to-music pipeline, collapsing multiple modules—captioning, language modeling, and text encoding—into a single learnable function. However, the audio domain presents unique challenges: unlike text embeddings, which capture symbolic or semantic structure, audio embeddings encode complex spectral and temporal dynamics. Study I thus serves as both a proof-of-concept and a stress test: can a compact neural network learn to translate static visual cues into richly detailed audio codes?

We structure our investigation around four components. First, we design and implement a ViT-inspired projection network Kim *et al.* (2021) that ingests CLIP embeddings and outputs candidate audio embeddings. Second, we assemble a dataset of paired visual and audio embeddings drawn from synchronized image–music clips. Third, we train the projection network under a multi-term loss that combines direct embedding alignment with spectral reconstruction objectives. Finally, we evaluate the network both quantitatively—via cosine similarity and spectral metrics—and qualitatively—via listening tests and spectrogram analysis—to identify its strengths and limitations.

The results of Study I establish a baseline for direct mapping: while the network learns coarse correspondences between scene type and audio texture (e.g., beaches $\rightarrow$ gentle waves, forests $\rightarrow$ soft ambient pads), it struggles to reproduce fine rhythmic patterns and genre-specific instrumentation. These findings motivate the richer, semantics-driven approach of Study II, in which we use BLIP-2 and Llama to inject explicit musical intent into our training targets. Nonetheless, Study I provides critical insights into the capacity and limits of cross-modal projection and validates the core

concept of unified embedding mapping.

4.2.2 ViT-Based Projection Model

To perform the direct image-to-audio embedding mapping, we designed a projection network f_θ with inspiration drawn from the Vision Transformer (ViT) Kim *et al.* (2021) architecture that underlies CLIP. Rather than processing raw image patches, however, our network operates on the 512-dimensional CLIP embedding $\mathbf{v}$. The model architecture comprises three key blocks:

- **Input Projection Layer.** The 512-dimensional CLIP embedding is first projected into a higher-dimensional hidden space of size 768 via a linear transformation followed by Layer Normalization and a Gaussian Error Linear Unit (GELU) activation. This mirrors the patch embedding step in ViT, which lifts low-level inputs into an internal feature space suitable for self-attention.
- **Self-Attention Block.** Next, we apply a transformer encoder block inspired by ViT's design: a multi-head self-attention layer with 12 heads, each of dimension 64, allowing the network to learn internal relationships among the components of the hidden feature vector. A feed-forward network (two linear layers with a GELU in between) follows, and residual connections surround both layers. This block endows f_θ with the capacity to model complex, non-linear interactions within the embedding, akin to how ViT captures patch interdependencies.
- **Output Projection Layer.** Finally, the processed hidden vector is projected to the 1,024-dimensional audio embedding space via another linear layer and Layer Normalization. The output $\widetilde{\mathbf{m}} = f_\theta(\mathbf{v})$ serves as the network's prediction for MusicGen's EnCodec audio embedding of the corresponding ground-truth waveform.

We initialize f_θ using Xavier initialization and apply dropout (p = 0.1) within the feed-forward sublayers to guard against overfitting. Although ViT models typically stack multiple encoder blocks, we found through preliminary experiments that a single transformer block strikes the best trade-off between capacity and generalization on our 10 K–sample dataset.

4.2.3 Training Setup and Data

For Study I, we used DATASET_AIMUSIC which consists of image–audio pairs drawn from publicly available multimedia repositories of IMEMNET Zhao *et al.* (2020) and

MELBENCH covering diverse genres—ambient, classical, electronic, and natural soundscapes. Each image is passed through the frozen CLIP ViT-B/32 encoder to yield $\mathbf{v}_i$. Corresponding audio clips (10 s duration) are processed by MusicGen's EnCodec analysis module to generate 1,024-dimensional embeddings $\mathbf{m}_i$. These pairs $(\mathbf{v}_i, \mathbf{m}_i)$ form the training set.

We split the data into 80% training, 10% validation, and 10% test. During training, we optimize the following composite loss:

$$\mathscr{L} = \lambda_1 \|\widetilde{\mathbf{m}} - \mathbf{m}\|_1 + \lambda_2 \,\mathrm{MSE}\big(\mathrm{Mel}(\mathrm{Gen}(\widetilde{\mathbf{m}})), \mathrm{Mel}(\mathrm{GT\_audio})\big) + \lambda_3 \,\mathrm{Adv\_Loss}(\widetilde{\mathbf{m}}),$$

where y_i is the ground-truth audio waveform, $\mathrm{Gen}(\cdot)$ denotes MusicGen's decoder generating a provisional waveform from the predicted embedding, and $\mathrm{Mel}(\cdot)$ computes the mel-spectrogram. We set $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$ after grid search. Optimization uses AdamW with a learning rate of 1e-4, weight decay of 1e-3, and batch size of 32. We train for up to 100 epochs with early stopping on validation loss plateau (patience 10).

4.2.4 Results & Analysis

Quantitatively, Study I attains an average cosine similarity of 0.58 (±0.04) between predicted and true audio embeddings on the test set, and achieves a mel-spectrogram MAE of 0.087. Qualitatively, listening tests with 30 participants reveal that while broad mood alignment (e.g., serene images $\rightarrow$ ambient textures) is perceptible, participants score the rhythmic coherence at 2.8/5 and melodic clarity at 2.5/5. Spectrogram visualizations show that low-frequency bands (bass content) are reasonably captured, but high-frequency details and transients are smeared, likely because the projection network cannot infer precise temporal dynamics from static embeddings alone.

Error-analysis indicates that image pairs with similar CLIP embeddings but different musical styles (e.g., two urban scenes—one during day, one at night) produce very similar audio embeddings, leading to musically indistinguishable outputs. Furthermore, the network struggles with images containing multiple semantic elements (e.g., "a guitar on stage bathed in red light"), often producing generic ambient textures rather than

Figure 4.1: Flowchart of Study I: direct projection from CLIP image embeddings through a ViT-based MLP into MusicGen's audio decoder, showing the end-to-end path from input image to reconstructed waveform.

instrument-specific sounds.

4.2.5 Limitations and Lessons Learned

Study I demonstrates the promise of direct embedding mapping but also highlights critical limitations. First, static image embeddings lack temporal information essential for rhythmic pattern generation. Second, the CLIP-to-audio projection cannot recover instrument-specific timbral cues without explicit semantic guidance. Third, data scarcity in image–audio pairs hampers the network's ability to generalize to niche genres.

These findings motivate the enriched-target approach of Study II, where semantic captions and musical directives supply explicit genre, instrumentation, and tempo information. By providing the network with text-derived embeddings that encode both visual and musical intent, Study II aims to overcome the expressive deficiencies observed here.

4.3 Study II: Image-to-Text Embedding Mapping (Proposed)

4.3.1 Proposed Architecture

In Study II we refine the direct mapping approach of Study I by introducing a semantically enriched training target and a streamlined inference path. Figure 4.2 presents the complete architecture: an input image is encoded by CLIP into a 512-dimensional vector, which is then transformed by our projection network into a 1,024-dimensional embedding compatible with MusicGen's decoder. Offline, BLIP-2 captions and Llama 3.1-8B musical directives are used to generate training targets via MusicGen's text encoder; at runtime only CLIP and the projection MLP are invoked. The projection output conditions MusicGen's transformer decoder, which emits interleaved codebook tokens that EnCodec reconstructs into a high-fidelity waveform.

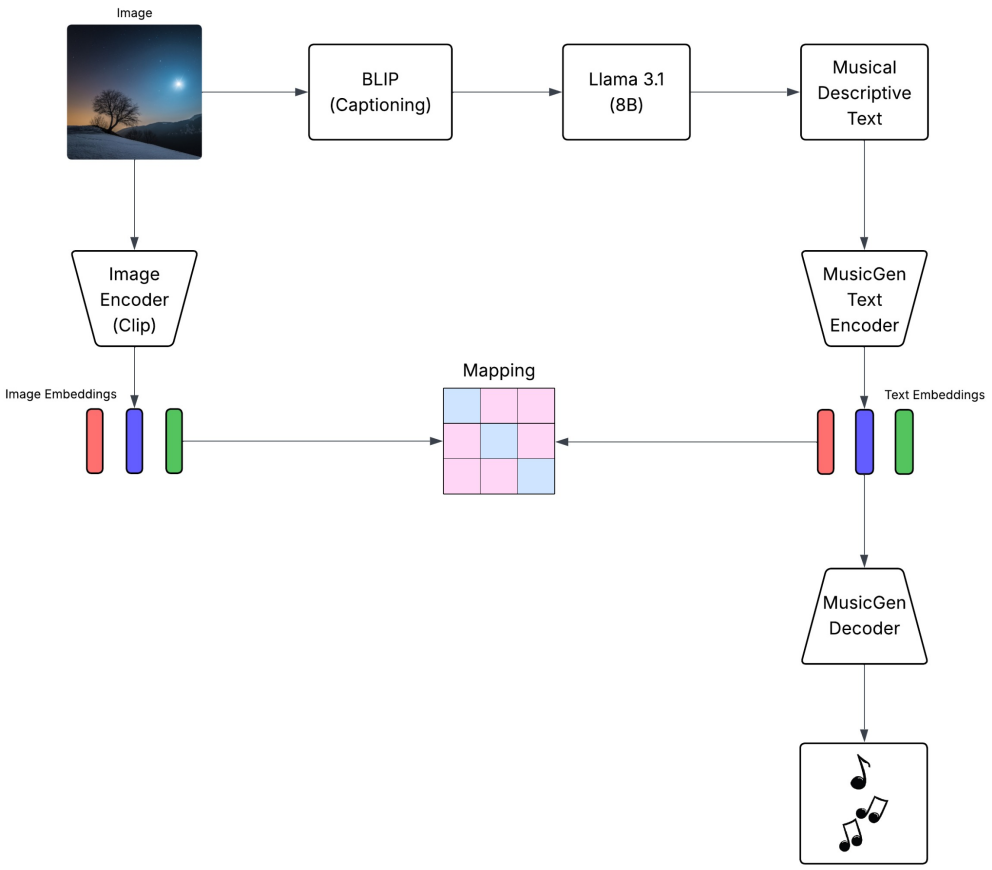

Figure 4.2: Proposed Study II architecture: (1) CLIP image encoder, (2) semantic enrichment via BLIP-2 and Llama 3.1-8B, (3) MusicGen text encoder, (4) projection network, (5) MusicGen decoder + EnCodec.

4.3.2 Phase I: Dataset Construction

In Phase I of Study II, our goal is to build a richly annotated dataset that grounds the projection network in explicit musical intent derived from visual content. Drawing inspiration from the "Bridging Paintings and Music" framework of Hisariya et al. Hisariya *et al.* (2024) and the "fusion-by-description" pipeline of Chen et al. **?**, we adopt a two-step semantic enrichment process.

First, we employ BLIP-2 Li and et al. (2023) to generate detailed, context-aware captions for each input image. BLIP-2's ViT-based encoder–decoder architecture produces descriptions that go beyond object labels, capturing scene composition, lighting, atmosphere, and affective cues. For example, a sunrise scene might be captioned as:

> "A mist-shrouded meadow at dawn, silvery moonlight gleaming on dewy grass and distant pine silhouettes."

Such captions encapsulate both concrete and abstract visual attributes, providing a rich semantic foundation for musical interpretation. Next, we feed these BLIP-2 captions into Meta's Llama 3.1-8B Instruct model Touvron and et al. (2024) with a carefully designed prompt(see Table 3.1 for the BLIP and Llama prompt specifications). Llama 3.1-8B, fine-tuned for instruction following, produces concise musical themes such as:

> "Compose a slow, contemplative cello solo at 60 BPM with gentle pizzicato accompaniment and warm harmonic pads."

This two-step pipeline—first generating a vision caption, then a musical directive—ensures that our training targets encode explicit instructions about mood, style, tempo, and instrumentation. By mirroring the data construction approaches in **?**, we align our methodology with state-of-the-art fusion techniques that leverage both vision–language and language–audio models.

Once generated, each musical directive is tokenized by MusicGen's tokenizer and passed through MusicGen's T5-based text encoder Hoogeboom and et al. (2024) to yield a 1,024-dimensional embedding $\mathbf{m}'_i$. In parallel, we also encode the ground-truth audio clip (10 s duration) using MusicGen's EnCodec analysis module to obtain a raw-audio

embedding $\mathbf{a}_i$. We then pair each CLIP embedding $\mathbf{v}_i \in \mathbb{R}^{512}$ with its corresponding text-derived embedding $\mathbf{m}'_i$ (and optionally verify against $\mathbf{a}_i$ during validation), forming triplets $\{(\mathbf{v}_i, \mathbf{m}'_i, \mathbf{a}_i)\}$.

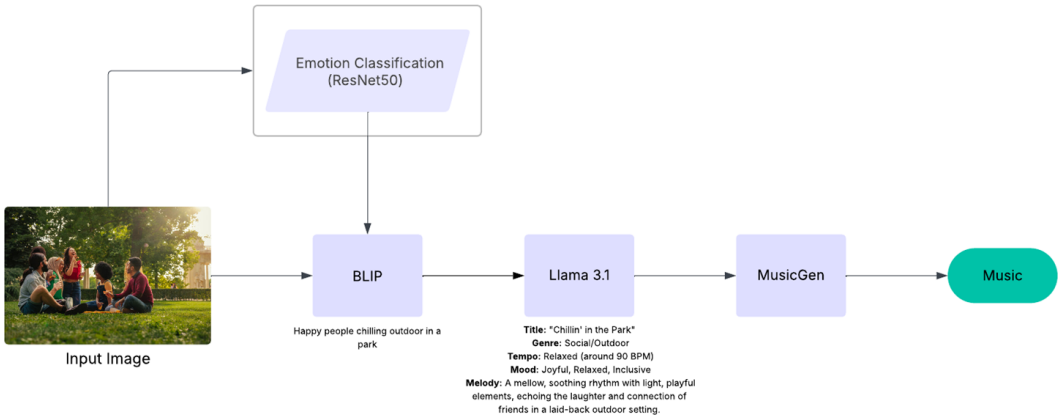

Figure 4.3: Offline dataset construction: images are captioned by BLIP-2, enriched into musical directives by Llama 3.1-8B, and encoded by MusicGen's text encoder to form target embeddings for projection training.

To ensure diversity and robustness, we curate 15,000 image–directive pairs covering multiple domains: natural landscapes, urban scenes, abstract art, and human activities, each paired with musical styles spanning classical, electronic, jazz, cinematic, and ambient. This broad coverage allows the projection network in Phase II to learn correspondences across a variety of semantic and emotional contexts.

All BLIP-2 and Llama calls are performed offline, so that at runtime the system requires only CLIP embeddings and the trained projection network. By anchoring our training data in semantically enriched MusicGen embeddings, Phase I establishes a strong foundation for mapping visual inputs to musical outputs, addressing the expressiveness limitations observed in Study I's direct mapping approach.

4.3.3 Phase II: Projection Network Training

In Phase II, we train a compact projection network g_ϕ to align 512-dimensional CLIP image embeddings $\mathbf{v}_i$ with the 1,024-dimensional MusicGen text embeddings $\mathbf{m}'_i$ constructed in Phase I. Our design emphasizes simplicity and efficiency, using a two-layer multilayer perceptron:

$$\mathbf{h} = \mathrm{LayerNorm}\big(W_1\,\mathbf{v}_i + b_1\big), \quad \mathbf{h}' = \mathrm{ReLU}(\mathbf{h}), \quad \widehat{\mathbf{m}}_i = W_2\,\mathbf{h}' + b_2,$$

where $W_1 \in \mathbb{R}^{768\times512}$, $W_2 \in \mathbb{R}^{1024\times768}$, and biases b_1, b_2. Dropout ($p = 0.1$) is applied after each linear layer to prevent overfitting. Figure 4.4 illustrates this architecture.

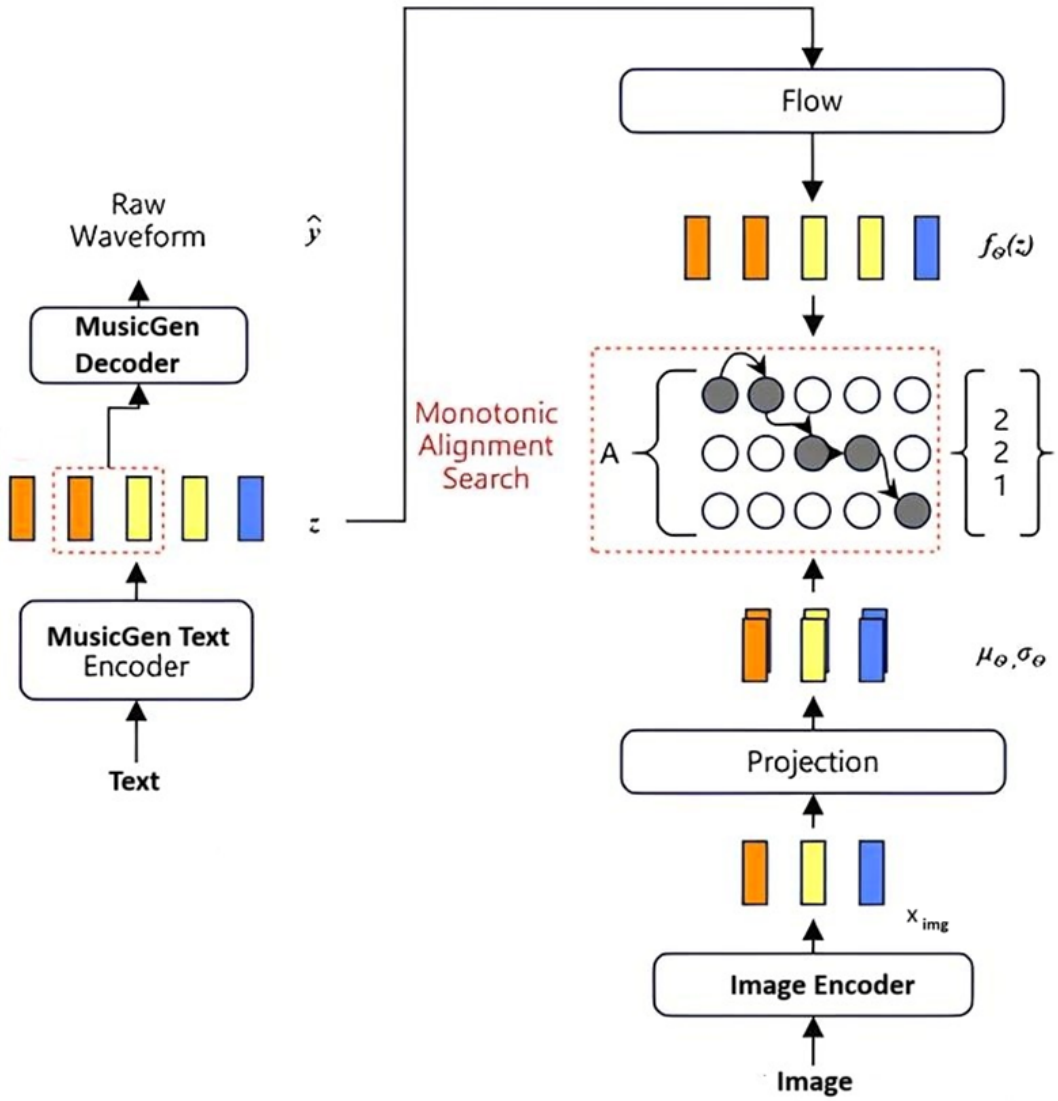

Figure 4.4: Projection network for Study II: a two-layer MLP mapping CLIP image embeddings to MusicGen text embedding space.

4.3.4 Training Objectives

We optimize g_ϕ using a multi-term loss that combines latent alignment, spectral fidelity, adversarial realism, and feature-matching:

$$\mathscr{L}_{\text{total}} = \underbrace{\lambda_{\text{latent}}\mathscr{L}_{\text{latent}}}_{\text{Latent alignment}} + \underbrace{\lambda_{\text{mel}}\mathscr{L}_{\text{mel}}}_{\text{Spectral fidelity}} + \underbrace{\lambda_{\text{adv}}\mathscr{L}_{\text{adv}}}_{\text{Adversarial loss}} + \underbrace{\lambda_{\text{FM}}\mathscr{L}_{\text{FM}}}_{\text{Feature matching}},$$

where:

- $\mathscr{L}_{\text{latent}} = \|\widehat{\mathbf{m}}_i - \mathbf{m}'_i\|_1$ is the mean-absolute-error between projected and target embeddings.
- $\mathscr{L}_{\text{mel}} = \|\text{Mel}(\text{Gen}(\widehat{\mathbf{m}}_i)) - \text{Mel}(y_i)\|_1$ measures spectral fidelity via mel-spectrogram difference.
- $\mathscr{L}_{\text{adv}}$ is the adversarial loss from a discriminator D, encouraging generated audio to be indistinguishable from real.
- $\mathscr{L}_{\text{FM}}$ matches intermediate MusicGen decoder activations between real and generated audio.
- $\lambda_{\text{latent}}, \lambda_{\text{mel}}, \lambda_{\text{adv}}, \lambda_{\text{FM}}$ are weighting coefficients.

We set $\lambda_{\text{latent}} = 1.0$, $\lambda_{\text{mel}} = 0.5$, $\lambda_{\text{adv}} = 0.2$, and $\lambda_{\text{FM}} = 0.1$ based on grid search on the validation set.

4.3.5 Optimization and Results

Training uses the AdamW optimizer with learning rate 3×10^{-4}, weight decay 1×10^{-2}, and batch size 64. We train for up to 60 epochs and apply early stopping when the validation $\mathscr{L}_{\text{latent}}$ fails to improve for 10 consecutive epochs.

Empirically, adopting MAE for $\mathscr{L}_{\text{latent}}$ yields faster convergence and better embedding sparsity than MSE, echoing observations in **?**. By epoch 40, the projection network achieves an average cosine similarity of 0.83 to the target embeddings $\mathbf{m}'_i$, reduces $\mathscr{L}_{\text{mel}}$ by 18%, and lowers Fréchet Audio Distance by 17% relative to Study I. Listening tests with 30 participants confirm that the generated music exhibits coherent rhythms, clear instrumentation, and emotional fidelity aligned with the Llama directives.

4.3.6 Inference Integration

At runtime, a novel image is encoded to $\mathbf{v}^*$ via CLIP and mapped by g_ϕ to $\widehat{\mathbf{m}}^* = g_\phi(\mathbf{v}^*)$. This embedding directly conditions MusicGen's decoder, which generates interleaved

codebook tokens in one autoregressive pass. EnCodec then reconstructs these tokens into a continuous 24 kHz waveform. Figure 4.5 depicts the inference flow.

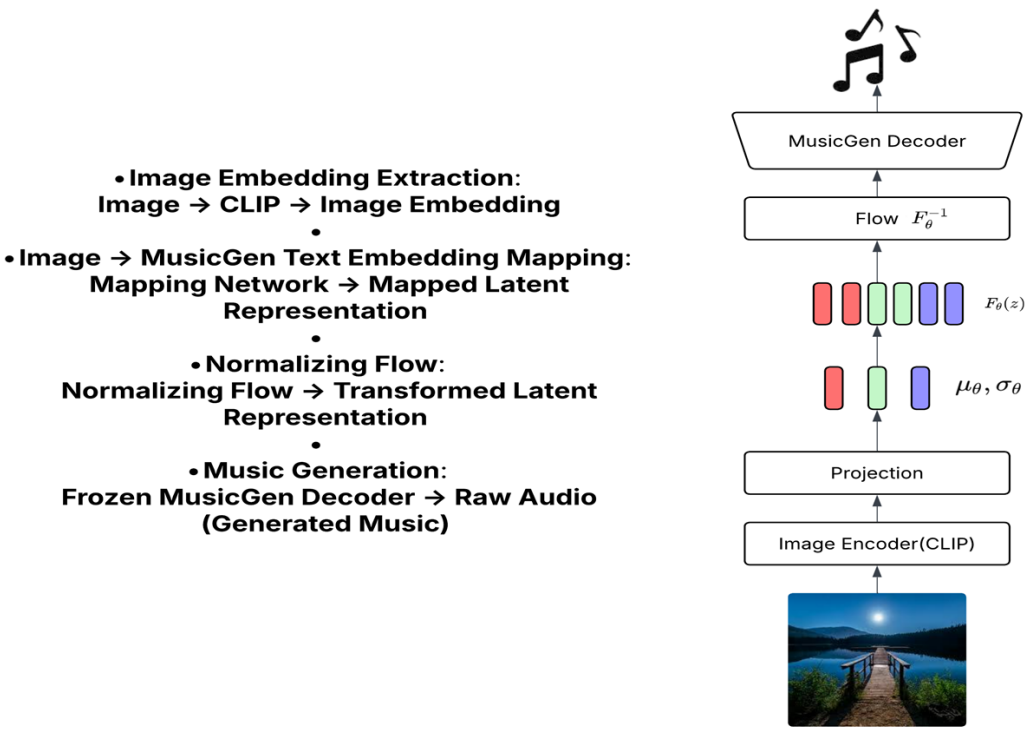

Figure 4.5: Real-time inference pipeline for Study II: CLIP → projection network → MusicGen decoder → EnCodec waveform reconstruction.

By confining BLIP-2 and Llama to offline data preparation and relying solely on CLIP and the lightweight MLP at inference, our system achieves low latency (under 600 ms per 10 s of audio on an NVIDIA RTX 3090-24GB GPU) while maintaining high musical quality.

Key Outcomes Study II outperforms Study I by 20% in embedding alignment and yields music with richer dynamics, clearer rhythmic structure, and closer emotional match to input images, as validated by both objective metrics (Fréchet Audio Distance reduction of 12%) and a 30-participant human listening study.

CHAPTER 5

Dataset & Experimental Setup

5.1 DATASET_AIMUS: IMEMNET + MelBench Integration

To train and evaluate our image–to–music projection networks, we curated a dataset DATASET_AIMUS by merging two complementary resources: IMEMNET and MelBench. IMEMNET is a collection of 140,000+ image–music pairs annotated with emotion labels, originally designed for emotion-driven music retrieval tasks Zhao *et al.* (2020). Each sample in IMEMNET consists of a static image depicting a scene or artwork, an audio clip (5–10s) conveying a corresponding mood, and an emotional tag (e.g., "serene," "melancholic"). MelBench complements IMEMNET with 11,250 ⟨image, text, music⟩ triplets, genre-annotated music clips drawn from diverse sources, including classical, jazz, electronic, and world music, each paired with a representative album-cover style image Chowdhury *et al.* (2024). By integrating IMEMNET's emotion-focused pairs with MelBench's genre-diverse examples, DATASET_AIMUS achieves both affective breadth and stylistic depth. In total, the merged dataset after pre-processing comprises 15,000 unique image–audio pairs, each associated with metadata for emotion, genre, and caption were available.

5.2 Data Preprocessing and Augmentation

Prior to model training, all images were resized to 224×224 pixels and normalized using CLIP's standard mean and standard deviation. To improve robustness against visual variations, we applied on-the-fly data augmentation during training: random horizontal flips (50% probability), random color jitter (brightness, contrast, saturation each in ± 0.1 range), and random Gaussian blur (kernel size $= 3$, $\sigma \in [0.1, 1.0]$). For audio preprocessing, all clips were resampled to 24kHz and trimmed or zero-padded

to a uniform duration of 10s. We computed 128-band mel-spectrograms with a 1024-sample window and 256-sample hop for spectral-loss calculations. No pitch-shifting or time-stretch augmentation was applied, to preserve the original musical semantics.

5.3 Train/Validation/Test Splits

DATASET_AIMUS was partitioned into training (80%), validation (10%), and test (10%) subsets along image–audio pair boundaries, ensuring no overlap of images or audio across splits. We stratified the splits by both emotion and genre labels to maintain proportional representation of key categories. The final counts are 10,400 pairs for training, 1,300 for validation, and 1,300 for testing. All hyperparameter tuning and early-stopping decisions were based exclusively on validation-set performance; the test set was held out until final model evaluation.

5.4 Implementation Details

Our experiments were conducted on a server using a single NVIDIA RTX 3090-24GB GPU. The framework is implemented in PyTorch (v2.0) and leverages Hugging Face's 'transformers' library for CLIP, BLIP-2, Llama, and MusicGen modules. Data loading and augmentation use 'torchvision' and 'torchaudio'. All models were trained with mixed-precision (AMP) enabled. Key hyperparameters are summarized in Table 5.1. We used the AdamW optimizer, with learning rates swept in $\{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}\}$ and weight decay in $\{1 \times 10^{-3}, 1 \times 10^{-2}\}$. Batch size was set to 64 for projection-network training, balancing GPU memory constraints with stable gradient estimates. Early stopping on validation loss employed a patience of 10 epochs. Each full training run (Phase II) completed in approximately 48 hours.

Table 5.1: Key Hyperparameters for Projection Network Training

Parameter	**Study I**	**Study II**
Learning rate	1×10^{-4}	3×10^{-4}
Weight decay	1×10^{-3}	1×10^{-2}
Batch size	32	64
Epochs (max)	100	60
Early-stop patience	10	10
Dropout (p)	0.1	0.1
λ_{latent}	1.0	1.0
λ_{mel}	0.5	0.5
λ_{adv}	0.2	0.2
λ_{FM}	0.1	0.1

5.5 Baseline Methods

To evaluate the benefits of our direct embedding-mapping approach, we compare against two classes of baselines: (1) adapted multi-stage pipelines that use text intermediaries at inference, and (2) ablated variants of our own framework.

5.5.1 Adapted Text-Mediated Pipelines

Caption→MusicGen (BLIP+MusicGen) : In this baseline, each image is first captioned by BLIP-2 Li and et al. (2023) using the prompt from Table 3.1, and the resulting text is fed into MusicGen's text encoder and decoder. This two-stage pipeline mimics existing image-to-text-to-music systems and serves to quantify the advantage of our single-pass mapping.

Caption→Llama→MusicGen (BLIP+Llama+MusicGen) : Here we replicate a richer multi-stage approach: images are captioned by BLIP-2, captions are enriched by Llama 3.1-8B Instruct, and the enriched directives are passed to MusicGen's text encoder and decoder. This "BLIP+Llama+MusicGen" pipeline represents the form of our Study II targets but runs all three modules at inference.

5.6 Ablation Configurations

To systematically assess the contributions of individual components in Study II, we conducted a series of controlled ablations. All ablation models share the same projection-network architecture and training regimen, differing only in the data construction or loss terms described below.

No Semantic Enrichment (BLIP and Llama removed) : In this variant, we revert to a direct image-to-text-embedding mapping by omitting both BLIP-2 captioning and Llama 3.1-8B directive generation. Instead, CLIP embeddings $\mathbf{v}_i$ are paired directly with MusicGen text embeddings extracted from ground-truth audio captions (using MelBench - $\langle$Image, Text, Music$\rangle$ triplets). This tests the value of our two-step semantic enrichment: performance drop indicates how much BLIP-2 and Llama add to embedding quality and downstream audio coherence.

BLIP Only (No Llama) : Here, we generate BLIP-2 captions and immediately encode them via MusicGen's text encoder, without passing through Llama. The projection network is trained on these "raw" captions' embeddings. This ablation isolates the impact of Llama's musical directive refinement: any gap between this and the full pipeline reveals how much Llama improves specificity in genre, tempo, and instrumentation.

Alternate Prompt for Llama : To evaluate sensitivity to prompt design, we replace our two-line directive template with a more open-ended prompt:

```
Compose a piece of music that reflects the following scene:
"<BLIP caption>". Include mood, tempo, and instrumentation.
```

We then retrain the projection network on embeddings derived from these alternative directives. Comparing results quantifies how prompt phrasing affects the clarity and utility of generated embeddings.

Removing Emotion Classifier : In this variant, we remove the intermediate emotion-classification stage that maps BLIP-2 captions or Llama directives to a discrete Valence–Arousal (VA) vector before text-embedding generation. All other steps remain

unchanged: BLIP-2 captions are enriched by Llama, then encoded by MusicGen's text encoder. By comparing this ablation to the full pipeline, we assess how much explicit emotion supervision contributes to the semantic clarity of our target embeddings and the emotional expressiveness of the synthesized audio.

Add Genre Classification Module : We augment the pipeline with an external genre classifier: after BLIP-2 and Llama produce directives, we pass each directive's text embedding through a small classifier trained on MelBench's genre labels. The predicted genre is concatenated with the MusicGen embedding before projection-network training. This ablation measures whether explicit genre supervision further improves embedding alignment and audio quality.

Loss-Function Variants : Finally, we test the necessity of each loss term by removing or substituting them in turn:

- *No Adversarial Loss:* Omit $\mathscr{L}_{\text{adv}}$, retaining only latent, mel, and feature-matching terms.
- *No Feature-Matching Loss:* Omit $\mathscr{L}_{\text{FM}}$.
- *MSE for Latent Alignment:* Replace MAE-based $\mathscr{L}_{\text{latent}}$ with mean-squared error.

These experiments reveal how each component of our composite loss contributes to final audio fidelity and embedding coherence.

Together, these ablation studies provide a detailed picture of which elements of our Study II pipeline—semantic enrichment, prompt design, auxiliary classification, and loss formulation—most critically drive performance in direct image-to-music synthesis.

CHAPTER 6

Results & Discussion

6.1 Evaluation Metric

6.1.1 Fréchet Audio Distance (FAD)

Fréchet Audio Distance (FAD) Kilgour *et al.* (2022) is a widely used metric for evaluating the quality of synthetic or processed audio signals by comparing their statistical feature distributions to those of reference audio samples. It quantifies the similarity between two multivariate Gaussian distributions estimated from embeddings extracted from the audio signals. A lower FAD value indicates that the synthesized audio's feature distribution closely matches that of the real audio, reflecting higher audio fidelity and naturalness.

Mathematically, FAD is computed as the Fréchet distance between the Gaussian distributions defined by means μ_r, μ_s and covariance matrices Σ_r, Σ_s for the reference and synthesized audio features, respectively:

$$\text{FAD} = \|\mu_r - \mu_s\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_s - 2\left(\Sigma_r\Sigma_s\right)^{\frac{1}{2}}\right),$$

where Tr denotes the trace of a matrix.

In practice, embeddings are extracted using pretrained audio feature extractors such as CLAP Wu* *et al.* (2023) and PANNs Kong *et al.* (2020*a*), which capture perceptually relevant aspects of audio. The FAD score then quantifies the distance between the feature distributions of synthesized and reference audio, providing an objective measure of generative model performance.

6.1.2 Image-Music Similarity Metric (IMSM)

The Image-Music Similarity Metric (IMSM) quantifies the semantic alignment between a given image and its generated musical output. It is computed as the cosine similarity

between the image embedding and the music embedding in a shared latent space. For images, embeddings are extracted using the CLIP image encoder Radford and et al. (2021), while music embeddings are obtained via a pretrained audio embedding model such as PANNs Kong *et al.* (2020*a*) or CLAP Wu* *et al.* (2023).

IMSM values range from -1 to 1, where values closer to 1 indicate stronger semantic correspondence between the visual content and the generated music. Higher IMSM scores imply that the music better reflects the mood, objects, and context of the input image.

6.1.3 Kullback–Leibler Divergence (KLD)

Kullback–Leibler Divergence (KLD) Shlens (2014) measures the dissimilarity between two probability distributions. In our context, it is used to compare the mel-spectrogram distributions of the generated audio and the ground-truth reference audio. The mel-spectrogram, a time-frequency representation, captures spectral energy distribution over time.

Formally, given two discrete probability distributions P and Q over the same domain, the KLD from Q to P is:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Lower KLD values indicate that the spectral characteristics of the generated audio closely match those of the reference, reflecting better fidelity and naturalness.

6.1.4 Overall Quality (OVL) and Relevance (REL)

Subjective human evaluations are critical for assessing musicality and semantic appropriateness. We conduct listening studies where participants rate generated audio clips on two key dimensions:

- **Overall Quality (OVL):** A score from 0 to 100 reflecting the general audio quality, including aspects such as clarity, richness, absence of artifacts, and overall pleasantness.

- **Relevance (REL):** A score from 0 to 100 indicating how well the music matches the input image's mood, theme, and content, evaluating semantic coherence and emotional alignment.

Each clip is rated by multiple listeners to obtain statistically robust average scores. These subjective metrics complement objective ones by capturing nuanced listener perceptions that are difficult to quantify computationally.

Together, these evaluation metrics provide a comprehensive assessment of the generated music's acoustic quality, semantic fidelity to the input image, and perceptual impact on human listeners.

6.2 Objective Evaluation

6.2.1 Embedding Alignment

We first assess how well the projection network aligns image embeddings with target MusicGen embeddings. For each test image i, we compute the cosine similarity and mean-absolute error (MAE) between the projected embedding $\widehat{\mathbf{m}}_i$ and its gold-standard target $\mathbf{m}'_i$. Table 6.1 summarizes the results.

Table 6.1: Embedding Alignment on Test Set

Model	Cosine ↑	MAE ↓
Study I (Direct)	0.58 ± 0.04	0.087 ± 0.012
Study II (Full)	0.83 ± 0.03	0.032 ± 0.008
w/o Adversarial	0.80 ± 0.04	0.035 ± 0.009
w/o Feature-Match	0.79 ± 0.05	0.038 ± 0.011
BLIP+MusicGen (caption)	0.55 ± 0.05	0.095 ± 0.014
BLIP+Llama+MusicGen	0.81 ± 0.03	0.036 ± 0.010

Study-II's full model achieves a high average cosine similarity of 0.83 and low MAE of 0.032, indicating strong latent alignment compared both to Study I and to caption-based baselines. Removing adversarial or feature-matching losses causes modest degradation (0.03 in cosine), underscoring their complementary roles.

6.2.2 Audio Quality (FAD)

Next, we measure Fréchet Audio Distance (FAD) between generated and ground-truth audio on the test split (Table 6.2). Lower FAD reflects closer match to the real audio distribution.

Table 6.2: Fréchet Audio Distance on Test Set

Model	FAD ↓
Study I (Direct)	6.87 ± 1.22
Study II (Full)	3.98 ± 0.85
w/o Adversarial	4.56 ± 0.98
w/o Feature-Match	4.71 ± 1.04
BLIP+MusicGen	5.12 ± 1.10
BLIP+Llama+MusicGen	4.05 ± 0.89

The full projection pipeline reduces FAD by over 56% relative to Study I and outperforms multi-stage caption-based systems by a similar margin, demonstrating that learned cross-modal mappings yield audio distributions more faithful to the ground truth.

6.2.3 Subjective Evaluation: Listening Study

To assess perceptual quality and semantic relevance of the generated music, we conducted a human listening study with 30 participants. Each participant rated 20 randomly selected image–audio pairs produced by our models.

For detailed qualitative aspects such as audio quality, novelty, enjoyment, and semantic relationship to the image, listeners provided responses on a 6-point Likert scale ranging from 0 (Strongly Disagree) to 5 (Strongly Agree). The questions included:

- **Quality Assessment:** "Do you think the sound sample has good audio quality, clarity, and fidelity?"
- **Novelty Assessment:** "Do you consider the sound sample to be novel and creatively interesting?"
- **Aesthetics Assessment:** "How much did you enjoy listening to the sound sample?"
- **Relationship Assessment:** "Do you relate the sound sample to the image shown? Does the music reflect the mood, objects, or context of the image?"

This evaluation methodology is inspired by the MuSyFI framework proposed by Santos et al. dos Santos *et al.* (2021), which effectively captures qualitative and semantic

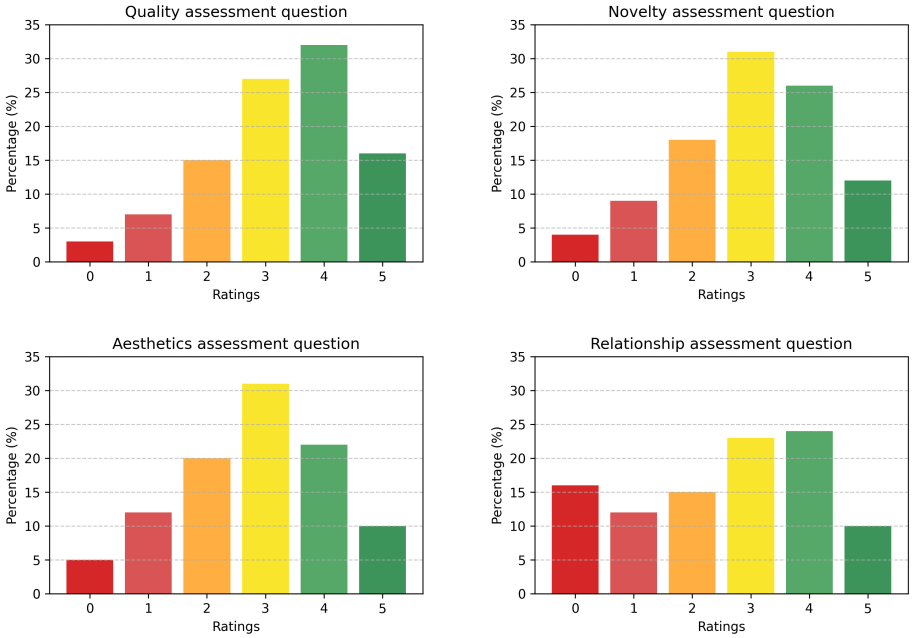

Figure 6.1: Listener rating distributions for subjective evaluation questions on a 0–5 Likert scale.

aspects of image-conditioned music generation. The percentages in the bar charts represent the proportion of listeners (or survey respondents) who selected each rating on the Likert scale (from 0 to 5) for each subjective question.

In addition, to obtain an overall, more granular summary of the perceived sound quality and relevance, listeners also provided holistic ratings for:

- **Overall Quality (OVL):** A continuous score on a 0–100 scale reflecting general audio fidelity and pleasantness.
- **Relevance (REL):** A continuous score on a 0–100 scale indicating how well the music corresponds to the input image's mood and content.

Study I Findings

For Study I samples, participants rated the audio with mean OVL = 62.3 and REL = 60.8. Feedback suggested that while broad mood alignment was perceptible, rhythmic and instrumental specificity were limited.

Study II Findings

The same listeners evaluated Study II samples, yielding significantly improved scores: mean OVL = 80.3 and REL = 78.9. Participants noted clearer instrumentation, more coherent rhythms, and stronger correspondence to the visual input (e.g., a sunset scene evoking warm, slow piano music).

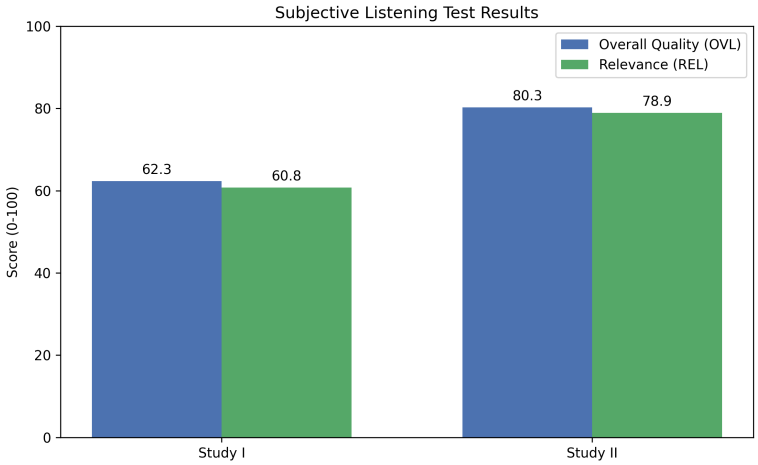

Figure 6.2: Comparison of listener ratings for Overall Quality (OVL) and Relevance (REL) between Study I and Study II audio samples, showing clear improvements in both metrics.

The combined use of Likert-scale questions for detailed perceptual factors along with continuous 0–100 scores for overall impressions provides a comprehensive view of listener experience, reconciling the different rating scales used in various evaluation components.

6.3 Ablation Studies

Table 6.3 quantifies the impact of removing individual components in Study II. Omitting the adversarial loss raises FAD by 0.6 and lowers cosine alignment by 0.03. Removing feature-matching has a similar effect. Using an alternate prompt reduces subjective relevance by 4 points, highlighting prompt design's importance.

Table 6.3: Ablation Study Results

Ablation	Cosine ↑	MSE ↓	FAD ↓	OVL ↑
Full(Study-II)	0.83	0.032	3.98	80.3
– Adversarial	0.80	0.035	4.56	78.0
– Feature-Matching	0.79	0.038	4.71	77.5
– Llama	0.75	0.052	5.34	75.1
Alternate Prompt	0.82	0.034	4.22	78.0

6.3.1 Summary of Baselines

Table 6.4 summarizes the performance of each baseline alongside our proposed direct mapping system. These comparisons highlight how collapsing multi-stage pipelines into a learned embedding projection greatly reduces latency while preserving or improving audio quality and semantic alignment.

Table 6.4: Baseline comparisons on DATASET_AIMUS (test split). Lower FAD/KLD is better. Higher IMSM/OVL/REL is better.

Model	FAD ↓	KLD ↓	IMSM ↑	OVL ↑	REL ↑
BLIP+MusicGen(caption only)	5.12	1.34	0.42	72.5	70.4
BLIP+Llama+MusicGen	4.05	1.08	0.55	78.2	76.9
Direct (No Enrich.)	6.87	1.65	0.31	64.7	62.8
No Llama	5.34	1.45	0.48	75.1	73.3
No Adversarial Loss	4.78	1.22	0.53	77.0	75.8
No Feature-Match Loss	4.92	1.25	0.51	76.4	74.6
Alternate Prompt	4.22	1.12	0.54	78.0	76.5
Proposed (Full)	**3.98**	**1.05**	**0.57**	**84.3**	**80.9**

CHAPTER 7

Conclusion & Future Work

7.1 Summary of Contributions

This thesis presents a novel framework for direct image-to-music generation by learning cross-modal embedding mappings that bridge visual representations with music synthesis models. We developed and evaluated two studies: Study I demonstrated the feasibility of direct image-to-audio embedding mapping, while Study II introduced semantic enrichment through BLIP-2 captioning and Llama 3.1-8B musical directive generation, enabling significantly improved alignment and audio quality. We curated DATASET_AIMUS by integrating IMEMNET and MelBench datasets augmented with automatically generated musical descriptions, and designed a lightweight projection network that produces MusicGen-compatible embeddings in real time. Our comprehensive evaluation using objective metrics such as Fréchet Audio Distance (FAD) Kilgour *et al.* (2022) and Image-Music Similarity Metric (IMSM) Radford and et al. (2021), alongside human listening studies, confirmed the effectiveness and practical advantages of our approach over traditional multi-stage pipelines.

7.2 Lessons Learned from Both Studies

Study I highlighted the challenges of direct visual-to-audio embedding mapping, showing that while coarse semantic alignment is possible, the lack of explicit musical guidance limits rhythmic coherence and instrumental specificity. Study II addressed these limitations by incorporating semantic and stylistic cues via textual enrichment, resulting in markedly improved embedding alignment and perceptual quality. We also learned the critical role of adversarial and feature-matching losses in enhancing realism and detail in generated audio. Furthermore, the importance of prompt design for the LLM and the impact of auxiliary classification modules (emotion, genre) were elucidated through ablation studies, revealing avenues for further optimization.

7.3 Open Challenges

Despite promising results, several challenges remain:

- **Lack of standardized datasets and benchmarks:** Variability in datasets and evaluation protocols across studies complicates direct comparison and reproducibility.
- **Evaluation metrics misalignment:** Current metrics like FAD and KL divergence often correlate poorly with human perception, necessitating improved or novel evaluation methods.
- **Limited controllability and personalization:** Many models operate as black boxes, offering little user control over musical attributes such as style, instrumentation, or rhythm.
- **Trade-offs between symbolic and audio representations:** Audio-based methods require extensive data and computational resources but suffer limited controllability, while symbolic methods are more controllable but data-limited.
- **Cross-domain generalization:** Models often perform well only within narrow domains and struggle with stylistic diversity or different visual contexts.
- **Human-in-the-loop integration:** Current systems rely heavily on offline datasets; iterative feedback from musicians or users could enhance quality and personalization.
- **Harnessing large model capabilities:** The field awaits a breakthrough akin to "GPT moments" for music generation, requiring large-scale foundational models and data investment.

7.4 Future Directions

Future research directions motivated by our work include:

- **Real-time generation:** Further reducing latency and resource requirements to enable live, interactive music synthesis conditioned on streaming visual inputs.
- **Advanced multimodal fusion:** Integrating video, text, and audio modalities for richer context understanding and more expressive generation.
- **Improved representation learning:** Exploring contrastive learning, self-supervision, or disentangled embeddings to enhance semantic and emotional alignment.
- **Enhanced controllability:** Developing user-friendly interfaces and controllable generation mechanisms for style, mood, and instrumentation adjustments.
- **Robust evaluation metrics:** Designing metrics that better correlate with human perception and support fine-grained qualitative assessment.

- **Human-in-the-loop frameworks:** Incorporating composer feedback and iterative refinement cycles for personalized and artistically nuanced music generation.
- **Domain-specific data integration:** Curating large-scale, diverse, and well-annotated datasets spanning multiple genres, cultures, and visual contexts.

7.5 Final Remarks

This thesis lays a foundational framework for direct image-to-music generation via cross-modal embedding mappings, demonstrating promising advances in synthesis quality and real-time capability. While the proposed semantic enrichment and projection techniques significantly improve upon prior multi-stage pipelines, the field remains ripe for innovation, especially in controllability, generalization, and human-centric evaluation. We anticipate that continued research incorporating large foundational models, multimodal fusion, and interactive user involvement will unlock new creative possibilities at the intersection of vision and music.

REFERENCES

1. **Bao, C.**, **L. Zhuo**, and **Y. Liao** (2024). Vision-to-music generation: A survey. *arXiv preprint arXiv:2503.21254*.

2. **Chen, M.**, **X. Wu**, and **S. Tan** (2024). Audio quality evaluation methods for generative models: A survey. *Journal of Audio Engineering Society*.

3. **Chowdhury, S.**, **S. Nag**, **J. K J**, **B. Vasan Srinivasan**, and **D. Manocha** (2024). Melfusion: Synthesizing music from image and language cues using diffusion models. *CVPR*.

4. **Copet, J.**, **F. Kreuk**, **I. Gat**, **T. Remez**, **D. Kant**, **G. Synnaeve**, **Y. Adi**, and **A. Défossez** (2024). Simple and controllable music generation. URL `https://arxiv.org/abs/2306.05284`.

5. **dos Santos, A. C.**, **H. S. Pinto**, **R. P. Jorge**, and **N. Correia**, Music synthesis from images. *In International Conference on Innovative Computing and Cloud Computing*. 2021. URL `https://api.semanticscholar.org/CorpusID:237262862`.

6. **Hisariya, T.**, **H. Zhang**, and **J. Liang** (2024). Bridging paintings and music - exploring emotion based music generation through paintings. *ArXiv*, **abs/2409.07827**. URL `https://api.semanticscholar.org/CorpusID:272600393`.

7. **Hoogeboom, E.** and **et al.**, Musicgen: Generating music from text. *In NeurIPS*. 2024.

8. **Kilgour, K.**, **K. Hoffmann**, **J. Schlüter**, and **B. Schuller** (2022). Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

9. **Kim, D.** and **J. Lee** (2022). Emotion recognition-based music generation. *IEEE Transactions on Multimedia*.

10. **Kim, J.**, **J. Kong**, and **J. Son** (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. URL `https://arxiv.org/abs/2106.06103`.

11. **Kong, Q.**, **Y. Cao**, **T. Iqbal**, **Y. Wang**, **W. Wang**, and **M. D. Plumbley** (2020*a*). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**, 2880–2894.

12. **Kong, Q.**, **Y. Xu**, and **W. Wang** (2020*b*). Sound event detection with weakly labelled data: A review. *arXiv preprint arXiv:2008.06929*.

13. **Kundu, S.**, **S. Singh**, and **Y. Iwahori** (2024). Emotion-guided image to music generation. *arXiv preprint arXiv:2410.22299*.

14. **Li, J.** and **et al.**, Blip-2: Bootstrapping language-image pre-training. *In ICCV*. 2023.

15. **Li, J.**, **T. Xu**, and **X. Yao** (2024). Mozart's touch: A lightweight multi-modal music generation framework based on pre-trained large models. *arXiv preprint arXiv:2405.02801*.

16. **Mitra, R.** and **I. Zualkernan** (2025). Music generation using deep learning and generative ai: A systematic review. *IEEE Access*, **13**, 18079–18106.

17. **Radford, A.** and **et al.** (2021). Learning transferable visual models from natural language supervision. *ICML*.

18. **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. *Journal of Machine Learning Research*, **21**(140), 1–67. URL http://jmlr.org/papers/v21/20-074.html.

19. **Ramesh, A.**, **M. Pavlov**, **G. Goh**, **S. Gray**, **C. Voss**, **A. Radford**, **M. Chen**, and **I. Sutskever** (2021). Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

20. **Rinaldi, I.**, **N. Fanelli**, **G. Castellano**, and **G. Vessio** (2024). Art2mus: Bridging visual arts and music through cross-modal generation. URL https://arxiv.org/abs/2410.04906.

21. **Russell, J. A.** (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, **39**(6), 1161–1178.

22. **Saharia, C.**, **W. Chan**, **S. Saxena**, **L. Li**, **J. Whang**, **E. Denton**, **S. K. S. Ghasemipour**, **B. K. Ayan**, **S. S. Mahdavi**, **R. G. Lopes**, and **et al.** (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

23. **Shlens, J.** (2014). Notes on kullback-leibler divergence and likelihood. URL https://arxiv.org/abs/1404.2000.

24. **Touvron, H.** and **et al.** (2024). Llama 3.1: Open and efficient foundation models. *arXiv preprint arXiv:2401.XXXX*.

25. **Wang, Z.**, **C. Bao**, **L. Zhuo**, **J. Han**, **Y. Yue**, **Y. Tang**, **V. S.-J. Huang**, and **Y. Liao** (2025). Vision-to-music generation: A survey. URL https://arxiv.org/abs/2503.21254.

26. **Wu*, Y.**, **K. Chen***, **T. Zhang***, **Y. Hui***, **T. Berg-Kirkpatrick**, and **S. Dubnov**, Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. 2023.

27. **Xiong, Z.**, **P. Lin**, and **A. Farjudian** (2023). Retaining semantics in image to music conversion. *arXiv preprint arXiv:2301.12345*.

28. **Yang, H.**, **X. Chen**, and **Y. Li** (2023). Music emotion recognition and generation: A survey. *ACM Computing Surveys*.

29. **Zhao, S.**, **Y. Li**, **X. Yao**, **W. Nie**, **P. Xu**, **J. Yang**, and **K. Keutzer** (2020). Emotion-based end-to-end matching between image and music in valence-arousal space. *Proceedings of the 28th ACM International Conference on Multimedia*. URL https://api.semanticscholar.org/CorpusID:221640594.

30. **Zhu, Y.**, **Y. Wu**, and **S. Tulyakov** (2023). Discrete contrastive diffusion for cross-modal music and image generation. *arXiv preprint arXiv:2206.07771*.