The Thesis Committee for Davin Lawrence
certifies that this is the approved version of the following thesis:

Resonating Visuals: Generative AI Strategies for Capturing the Essence of Music in Image Form

SUPERVISING COMMITTEE:

David Harwath, Supervisor

Eunsol Choi

Resonating Visuals: Generative AI Strategies for Capturing the Essence of Music in Image Form

by

Davin Lawrence

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Computer Science

The University of Texas at Austin

May 2024

Acknowledgments

The work presented in this thesis is part of work performed while employed by Vibe Video. The majority of the work in this document was performed by the author, unless otherwise noted.

Abstract

Resonating Visuals: Generative AI Strategies for Capturing the Essence of Music in Image Form

Davin Lawrence, M.S.Comp.Sci
The University of Texas at Austin, 2024

SUPERVISOR: David Harwath

This thesis studies methods of generating images which captures corresponding semantic information of music. The primary goal of each method is to generate an image given a song as the input. First, I describe a model-agnostic LLM interaction method which extracts information from lyrics to generate an image generation prompt. Second, I describe a model which produces image generation prompts directly from source music by presenting the problem as a music captioning task. To achieve this task, I create a synthetic dataset of music-image pairs first extracted from music videos before being captioned with a BLIP2 model. Finally, I further use this dataset to fine tune a Stable Diffusion model to produce images conditioned on text-audio pairs by introducing a CLAP encoder to the Stable Diffusion pipeline.

Table of Contents

Chapter 1: Introduction

Music is a uniquely human practice that spans across the globe and through time into antiquity. Nearly every culture has their own rich musical tradition informed by their shared cultural knowledge. In my own culture, music is used for celebration, for mourning, for filling gaps in conversation, and for so much more. My own life, and I assume the life of the reader, has been punctuated with important memories accentuated by favorite songs and pieces of music. Likewise, cultures across the globe have rich traditions in visual art. As with music, visual art serves many purposes, whether it to be to simply document important historical figures, or evoke deep emotion within the viewer.

Images and music both carry semantic information which relate information or evoke emotional responses in viewer or listener. Each modality has been used to bolster the other. For example, music is often used in film to add urgency to a tense situation or to evoke strong emotions during as a tragic event unfolds on screen. Terminology from one modality is borrowed to describe the other, such as when we describe the color of a piece of music or the tone of a photograph. In the case of music videos, images are added to the music to carry information about the personality of the performer or to add emotional weight to the music itself.

This work seeks to find the creative correspondence between the two modalities, specifically to create visual artifacts which correspond to a given sample of music. Music creation in the modern era is an increasingly solo endeavor, where an emerging musician must handle their own promotional material, marketing and promotion, scheduling, while still finding time to create and perform their music. The overarching goal of this work is to create tools which can lessen this burden on musicians. Digital music platforms are increasingly pushing for the inclusion of visual artwork to accompany music while streaming. The tools developed through this thesis aim

to assist musicians in creating such art, so they can focus on their primary mode of expression.

The last couple of years have been a whirlwind of advancement in many so-called generative models, with image and video generation enjoying rapid progress. Despite the quick jump in quality of generated images, these models still largely rely on natural language as their conditioning input. Conditioning on other modalities is largely unexplored, with music being perhaps the least explored. In this work, I explore three methods to bridge the gap between music and image generation. First, I describe the process which is currently used by Vibe Video[1] and their competitor MusixMatch[2]. Both of these are commercial systems which rely on a multi-turn LLM interaction to extract information from lyrics which is then synthesized into an image generation prompt. Second, I explore generating prompts for image generation models directly from musical inputs using a novel Encoder-Decoder model. Finally, I condition a Stable Diffusion (Rombach et al., 2022) model on musical inputs by combining text and audio embeddings. To study all three of these approaches, I collect two datasets of music videos as described in 3.

I begin this thesis with a literature review in Chapter 2. In the context of machine learning and AI research, the combination of music and imagery has been understudied. Prior work has largely focused on retrieval tasks, such as linking appropriate music to video (Suris et al., 2022), or discovering a common emotional space between images and music (Won et al., 2021; Stewart et al., 2023). Work has been performed for general audio in the context of Foley and sound effects for film (Yang et al., 2022; Yuan et al., 2023). Additionally, researchers have explored generating general audio from images (Iashin and Rahtu, 2021; Sheffer and Adi, 2022) as well as generating images from environmental sound (Wu et al., 2022). To my knowledge, application of music to produce imagery is lacking in the literature.

[1] `https://vibevideo.ai`
[2] `https://musixmatch.com`

The first approach I describe in chapter 4, represents existing work conditioning image prompts from lyrics. This work was developed by Ben Gillin at Vibe Video and represents the main motivation for the following methods. This work is centered around a multi-turn interaction with a LLM model to extract metadata about the lyrics and synthesize them into an image prompt which can be fed to an image generation model. My contributions to this work are two-fold. First, the majority of my contribution to the approach was engineering-focused, making small tweaks to the pipeline to make it more amenable to deployment. More importantly, I introduced quantitative analysis into the approach. Prior to my involvement, work on the strategy was done in an ad-hoc manner. In this section, I develop the analytical framework which will be applied to other methods. The lyric-to-prompt pipeline will act as a baseline for other work throughout this thesis.

The lyric-to-image pipeline is completely dependent on external APIs and lyrics, which is not a universal feature of music. In chapter 5, I address the shortcomings in the lyric-to-image pipeline by producing prompts directly from music. I introduce this as a modification to the music-captioning task, where the captions no longer describe the music, but rather images which represent a good accompaniment to the task. I train a family of models based on an Encoder-Decoder structure on this task and compare the results to the LLM interaction baseline, as well as each other. To train these models, I collect two datasets which are described in detail in chapter 3. Each of the models produce reasonable prompts for image generation, however, it is difficult to ascertain the exact correspondence to the underlying music.

Chapter 6 documents the work and challenges of producing images directly from music and music-text pairs. I develop a family of models which attempt to solve this problem using the datasets described above. First, I describe a model which fine-tunes an off-the-shelf Stable Diffusion model (Rombach et al., 2022) with audio embeddings extracted from the CLAP Audio model (Wu et al., 2023b). Second, I directly provide CLIP embeddings (Radford et al., 2021) from the Wav2CLIP model (Wu et al., 2022) in a training-free approach.

Chapter 2: Literature Review

The work performed in service of this thesis draws on the ideas from several subfields within AI research. The most notable of these fields are audio and music information retrieval, generative models, and multimodal information retrieval. In this section I will highlight notable works which are landmark moments in their field as well as other works which are directly related to the work performed here.

2.1 Audio and Music Information Retrieval

Information retrieval as applied to audio has several nuances as compared to information retrieval on other modalities such as images and text. As a modality, audio has the notion of different *types* of audio, commonly categorized as speech, music, and environmental sound. Each of these sub-modalities carries its own type of information, and as a result, its own family of approaches. Of the three, speech is perhaps the most studied with a rich history of Automatic Speech Recognition (ASR) and rapid advances in text-to-speech (TTS) systems. While music information retrieval (MIR, see 2.1.2) has benefited from research in these fields, MIR contains additional challenges to researchers due to the fundamental differences between speech and music. Speech can be represented at the phonetic level, requiring representations with periodicity on the order of milliseconds. Tasks in MIR, on the other hand, struggle to extract meaningful representations from such frames, as musical structures can vary over several time spans with several time-dependencies (Dieleman et al., 2018). In the task of music segmentation, for example, what constitutes a verse and chorus can span the time-frame of several seconds or even minutes in extreme cases. Moreover, repetitions of the same structure often have slight variations within them, further muddying the task. While the two fields are related, they often necessitate subtle differences in their approaches.

2.1.1 Audio Classification

Audio tagging is a classification task which aims to match an audio with a set of labels. Audio data is typically multi-label as there could be several unrelated, overlapping sounds in a single example (Wang et al., 2021). High quality, separated audio data requires highly-controlled environments and can be expensive to produce. Moreover, source-separated audio does not represent real-world situation in which models are deployed. Work in this field is typically guided by AudioSet (Gemmeke et al., 2017), a massive ontological dataset with roughly two million examples with 527 classes. ESC-50 (Piczak, 2015) and UrbanSound (Salamon et al., 2014) are two environmental sound datasets which assessing in sound event detection. FSD50k (Fonseca et al., 2022) is a more recent dataset constructed from the FreeSound library[1] while adhering to the AudioSet ontology.

Progress in audio classification has largely followed the overarching trends in other fields such as computer vision. Early efforts used shallow fully connected layers (Gemmeke et al., 2017) before finding early successes in CNN architectures (Hershey et al., 2017; Kong et al., 2020b). Further improvements were made with the introduction of attention modules and transformer architectures (Gong et al., 2021a; Koutini et al., 2022). Gong et al. (2021b) studied training techniques and showed pretraining and label aggregation can boost accuracy while minimizing the final network size. More recent work has continued to minimize model size using distillation from transformers to CNN-based Residual Nets(Schmid et al., 2023) or from large transformer ensembles to minimized, efficient Vision Transformers (Dinkel et al., 2023).

2.1.2 Music Information Retrieval

MIR comprises several sub-tasks such as genre classification (Lee et al., 2020; Zeng et al., 2021; Zhao et al., 2022), pitch estimation (Kim et al., 2018), beat de-

[1] `https://freesound.org/`

tection (Hockman et al., 2012; Mauch and Dixon, 2012; Nieto et al., 2019; Heydari et al., 2021; Chen et al., 2023a), and music transcription (Humphrey and Bello, 2012; Thickstun et al., 2017; Hennequin et al., 2019; Li et al., 2019; Kelz et al., 2019; Salamon et al., 2021). Most related to my work is Lyric extraction and alignment. ASR systems trained with music inputs perform very well on this task. Despite being evaluated originally on speech, casual observation shows Whisper (Radford et al., 2022) works well on lyrics, with alignment being further improved by variants such as DistillWhisper (Gandhi et al., 2023) and WhisperX (Bain et al., 2023). LyricWhiz (Zhuo et al., 2023) further improves on Whisper by prompting ChatGPT to produce the most likely correct lyrics from multiple whisper inferences on the same audio. Gao et al. (2022) first identify and condition on the genre of music before producing lyrics to assist the model with difficult genres of music such with distorted vocals such as heavy metal or fast lyrics as in hip hop.

2.1.3 Learning useful representations

Audio is a high-dimensional data source which contains perceptual information, such as pitch and timbre, as well as meaningful information through speech and music. Working directly with raw audio waveforms is computationally inefficient at best and intractable for some tasks at worst. Representation learning seeks compact, low-dimensional representations which can be used on a variety of downstream tasks. Historically, speech and music systems have relied on spectral features to extract information from source audio. In contrast, learned representations have driven recent progress in automatic speech recognition (ASR), audio tagging and classification, and audio generation. Early learned representations focused on dilated CNNs (van den Oord et al., 2016) and RNNs (Mehri et al., 2017; Kalchbrenner et al., 2018), before giving way to self-supervised models such as Wave2Vec (Baevski et al., 2020) and HuBERT (Hsu et al., 2021). Both of these models learn strong representations in pretraining, but are exclusively trained on speech data. MusicBERT (Zeng et al., 2021) and more recently MERT (Li et al., 2023b) have extended HuBERT to music

data to address difficulties of longer temporal correspondence in music data. Neural Codec models (Zeghidour et al., 2021; Défossez et al., 2022) have shown to produce powerful representations using residual vector quantization, and have impressive applications in generative audio (Copet et al., 2023; Vyas et al.). Generative models have also been shown to produce useful representations (Castellon et al., 2021)

Multimodal and contrastive approaches have shown to produce useful features for a variety of tasks. COALA (Favory et al., 2020) and COLA (Saeed et al., 2021) both use contrastive learning to produce speech representations. COALA additionally addresses speech and text alignment by simultaneously learning text representations. MuLan (Huang et al., 2022) learns music representations from contrastive learning on music and noisy text such as playlist titles, lyrics, and video comment. CLaMP (Wu et al., 2023a) employs a similar approach but utilizes a RoBERTa (Liu et al., 2019) language model to denoise and combine several text descriptions. CLAP (Wu et al., 2023b) uses finds representations of audio and text, training on both labels from AudioSet and captions from AudioCaps (Kim et al., 2019) and (Drossos et al., 2019). MusicCaps (Doh et al., 2023) generates natural language descriptions from tagged music and uses the synthetic dataset to train language model to produce music descriptions.

Contrastive learning has also been applied to music-image and music-video pairs. Wav2CLIP (Wu et al., 2022) learns by freezing CLIP (Radford et al., 2021) image encoders and using a contrastive loss on an audio encoder. Won et al. (2021) sought to learn an emotional embedding space motivated to pairing music to stories from text. Emo-CLIM (Stewart et al., 2023) follows a similar approach by audio representation from CLIP's vision encoder. Rather than use direct text descriptions, they group music-image pairs based on emotional descriptions to create emotional embeddings. (Suris et al., 2022) defines a multimodal retrieval task to pair appropriate music to a given video. Avramidis et al. (2023) learns audio representations using accompanying music videos as contextual information.

2.2 Generative Models

The past few years have witnessed rapid development of generative models with the development of diffusion models (Ho et al., 2020; Dhariwal and Nichol, 2021). Diffusion models learn to perform a denoising process by predicting the amount of noise added to an image. Most related to my work is Stable Diffusion (Rombach et al., 2022), an open-sourced image generation model from StabilityAI. Rombach's key contribution was several computational enhancements to the diffusion problem, notably moving the diffusion process to the latent space of variational autoencoder. Further work studies the effects the ability of editing existing images through cross attention control (Hertz et al., 2022) or by directly guiding the diffusion process (Brack et al., 2023). Similar techniques have given rise to video generation (Yang et al., 2022; Chen et al., 2023b; Liu et al., 2023b). Blattmann et al. (2023) extended Stable Diffusion to video by adding temporal convolutions to track consistency between video frames for videos ranging from three to five seconds. OpenAI recently announced their SORA model (Brooks et al., 2024) with claims of video generation lasting up to an hour. The authors formulate video generation as being akin to a language modeling task, allowing for long term dependencies in generated videos.

2.2.1 Audio and Music Generation

Audio generation has largely followed the approaches in the vision field, with advances in text-to-speech (Wang et al., 2023a), audio editing (Peng et al., 2024), environmental and Foley sound creation (Borsos et al., 2022; Yuan et al., 2023), and music generation (Copet et al., 2023). As in classification tasks, audio has built in difficulties when performing a reconstruction task. A common preprocessing for audio models is to perform a Fourier transform as a first step in feature extraction. A side effect of this transform is the removal of phase information of the original audio, which is crucial for realistic audio synthesis. Early neural approaches invoked generative adversarial networks (Goodfellow et al., 2014; Donahue et al., 2019; Kumar

et al., 2019), with the most notable example being HiFi-GAN (Kong et al., 2020a), which has shown to excel at the spectrogram inversion task. The development of neural audio compression models such as SoundStream (Zeghidour et al., 2021) and EnCodec (Défossez et al., 2022) produce waveforms directly using a residual VQ-VAE (van den Oord et al., 2018). The residual codebook entries from the latent space can be used as tokens for discrete language modelling. Neural codec language models have enabled long form audio generation across speech, audio, and environmental sound (Kreuk et al., 2022; Agostinelli et al., 2023; Rubenstein et al., 2023; Vyas et al.). Diffusion has also been introduced to audio diffusion through spectrogram diffusion (Hawthorne et al., 2022; Huang et al., 2023b; Liu et al., 2023a; Huang et al., 2023a; Chen et al., 2023a), which uses diffusion models to generate spectrograms which can subsequently run through a spectrogram inversion model such as HiFi-GAN.

Combining audio and visual is a nascent area of research, with approaches using either modality as a conditioning signal for the other. Iashin and Rahtu (2021) and Sheffer and Adi (2022) both condition generated audio on images using combinations of VQ-VAE and Transformer models. Bigioi et al. (2023) conditions video outputs on speech commands to enable vocal editing of generated video. A common approach is to modify an existing video by conditioning the original on an input sound (Lee et al., 2023a; Jeong et al., 2023), allowing different environments and effects to be added to the source video. Liu et al. (2023b) study the problem of simultaneously generating audio and video in a unified approach.

Chapter 3: Datasets and Data Collection

Both image and caption generation from music relies on paired music-visual data, making music videos a natural fit to the problem. Music videos can be described as a flow of images chosen by human creators chosen to accompany music. In many cases, can be used to highlight the emotional content of the song, with choices being made across content, coloring, and mood to match the song. While this holds true in many cases, there are also music videos which arbitrarily create images given music, making music videos a noisy data source. In this chapter, I report the data used throughout this thesis, using Gebru et al. (2021) as guidance.

Throughout the following chapters, I use two datasets to develop the various models or combinations thereof. The first dataset is YouTube8m-MusicVideo (YT8m-MV) as described in Suris et al. (2022). This is a subset of YouTube8m (Abu-El-Haija et al., 2016), a large video dataset consisting of 500K hours of multi-label video. The primary motivation of YouTube8m is as a benchmark and training data for video classification tasks, collected, maintained and released by Google. By default, the dataset provides frame-level features and labels of both the video and audio data. To recreate YouTub8m-MusicVideo, I search YouTube8m for videos classified as music videos, and save the video ID and download the associated video from YouTube. As the dataset lives on YouTube, the final YT8m-MV dataset should have high overlap with Suris et al. (2022), but may also be missing videos which have been removed for various reasons.

While YT8m-MV is a large dataset, it skews heavily to videos of performers, people lip-syncing popular music, people performing covers, or generally low-quality data. Examples of low-quality include low-resolution video, low-fidelity audio, and video with artifacts introduced from antiquated media formats or methods attempting to bypass content filters. To fill this gap, I create the AnimatedMusicVideo (AMV) dataset for this work. This dataset seeks higher-quality data by focusing on animated

music videos under the assumption that animations are often created with more intention and artistic interpretation of the source music. This assumption holds for most of the dataset, but there are still low-quality examples with video game footage being paired to music being a prime example. Removal of the majority of such low-quality examples is possible through automated filtering, but full removal of all low-quality examples would take significant effort. Videos for this dataset were collected in November 2023 by searching and collating playlists of animated music videos on YouTube. Before processing, videos are deduplicated by video ID as well as title and artist matching. At the end of collection and preprocessing described below in section 3.1, AMV's train split contains 31K examples, representing 86 hours of music and YT8m-MV contains 72K examples representing approximately 200 hours of music.

After collection, both datasets are processed by dividing the video into 10-second clips and sampling randomly from that list. As YT8m-MV is a much larger dataset, I randomly sample four clips from each video, while sampling eight clips from each video for the AMV dataset. For each clip, a frame is chosen at random to characterize the video data and the 10-seconds of audio is summed to mono and normalized. This processing step provides music-image pairs with which to train the models throughout this work. The video id and time span is also saved with each example, so the original video is recoverable at evaluation time. Both datasets are divided into train and evaluation splits before the chunking process to ensure each model never sees any information from evaluation data. It is probable that the two datasets contain some overlap, but with most training they are kept separate.

It is important to note that the data collected and used throughout this process contains the creative work of scores of individuals as well as copyrighted material. Moreover, much of the visual data in YT8m-MV contains images of people, some of which may be overrepresented in the dataset. As a lone researcher, it would be impossible to seek informed consent from all the individuals who have contributed to the content in this dataset. As a result, the neither the created datasets nor models derived from the data will be distributed in part or as a whole as a result of this

BLIP2 Prompts	"Describe this image", "Does this image have a particular style?", "Give a short generation prompt for this image"
Mistral-7B Prompt	Your task is to infer a a short image generation prompt for Stable Diffusion given a description of music and an accompanying image. Here are some examples of Stable Diffusion prompts: `PROMPT_EXAMPLES` Describe a Stable Diffusion prompt using the following descriptions in less than 70 words. Do not prepend with any additional information, just respond with the prompt. Do not use words such as "create", "image:" or "generate" at the beginning of the prompt. Do not use anyone's proper name or refer to any artists in the generated prompt. Only create a prompt for the image, do not make a prompt for the music. Here are the descriptions:

Table 3.1: Prompts from BLIP2 and Mistral-7B used in the captioning pipeline for the AMV dataset. YT8m-MV prompts are similar, but lack language which refers to descriptions of music. `PROMPT_EXAMPLES` is replaced with 5 examples drawn at random the DiffusionDB dataset.

work. Both models and metadata from the datasets can be provided upon request to researchers for verification of results.

3.1 Music Video Caption Generation

In chapter 5, I describe an approach which modifies the music captioning task to provide captions of images which may accompany the music. To my knowledge, there is no such dataset of music and image caption pairs. To cover this gap, I create synthetic captions for both YT8m-MV and AMV. While the process is largely the same for both datasets, there are slight differences in the processes for creating captions for each dataset. Since AMV is a much smaller dataset than YT8m-MV, I was able to iterate on the synthetic captioning task, as well as inject more information into the captioning process. Despite the differences in the approach, the final synthetic captions do not significantly differ in terms of quality.

Common to both approaches is an image captioning and summarization task performed to create synthetic captions. For the initial captions, I prompt an instruct-

Dataset	CLIPScore (↑)	B-CLIPScore (↑)
YouTube8m-MusicVideo	0.2717	–
AnimatedMusicVideo	0.2863	0.1612

Table 3.2: CLIPScores for the captioning process in either dataset. The captioning process results in a low CLIPScore, suggesting low correspondence between generated captions on the source image. B-CLIPScore reports the CLIPScore between the intermediate inferences generated by BLIP2 in captioning pipeline.

fine-tuned BLIP2 model (Li et al., 2023a) to extract information about the image. The extracted data is fed through InstructMistral-7B (Jiang et al., 2023), with instructions to use the information to create an image prompt to generate a similar image. The process for AMV is supplemented with musical information derived from LP-MusicCaps (Doh et al., 2023) to provide descriptions of the music with the goal of increasing correspondence between music and the final image. I show both families of prompts in table 3.1. Not including this information in the YT8m-MV is primarily due to computational and time constraints.

Table 3.2 shows the CLIPScore (Hessel et al., 2022) between the generated captions and the ground truth images. This metric is bound by zero and one, with one meaning perfect correspondence between the source caption and the image. The lower score suggests a significantly less than ideal amount of semantic overlap between the captions and the source images. Interestingly, I expected AMV to perform lower than YT8M-MV on this metric, as the generated caption is prompted with both music and image information, but this is not the case. Moreover, the CLIPScore has a wide variance over the captions, with the minimum and maximum scores on YT8m-MV being 0.7570 and 0.4012, respectively. Interestingly, the CLIPScore between the BLIP2 inferences and ground truth images averages 0.1612, which is consistent with the captioning ability reported in the BLIP2 paper.

Figure 3.1 displays a few samples drawn from the AMV dataset with their associated source video links and captions. This tells a slightly different story than the pure CLIPScore. While the CLIPScore may suggest a low correspondence, the images in the figure have somewhat appropriate captions for the scenes. However,

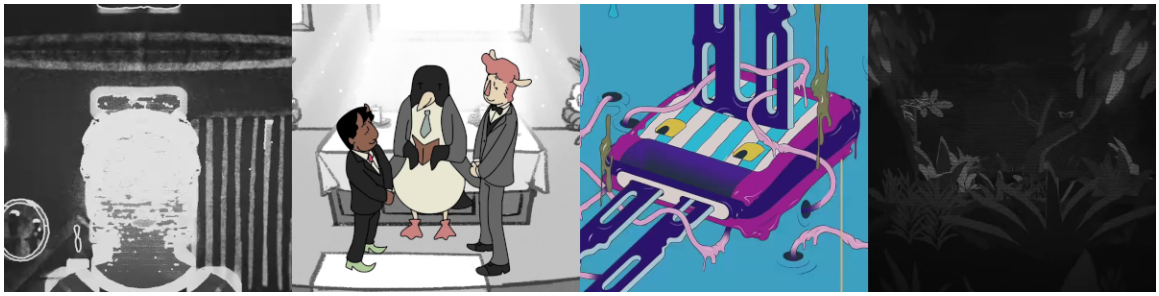


Figure 3.1: Randomly drawn samples from the AnimatedMusicVideo dataset. Each item has been center-cropped to fit on the screen. From left-to-write the associated YouTube video and caption are as follows: **(1)** `https://youtube.com/watch?v=YX4glq6llMg` Generate a vintage living room scene in black and white. Three chairs arranged centrally, two sofas flanking either side, an armchair by the fireplace, and a television visible on a dresser. **(2)** `https://youtube.com/watch?v=MneRtx7x2vs` A whimsical scene of two formally dressed individuals exchanging vows amidst vibrant floral arrangements. **(3)** `https://youtube.com/watch?v=VJm7IPrBmLY` Cartoon scene: A purple and pink suitcase floats among various objects in vibrant waters teeming with seagulls and an inflatable boat. Synth accordion plays harmoniously alongside a cheerful retro dance beat. **(4)** `https://youtube.com/watch?v=YX4glq6llMg` Generate an image of a mystical nighttime forest, filled with towering trees and lush greenery. Shadows dance among the foliage as unseen creatures rustle in the underbrush. Birds fly silhouetted against the moonless sky, their calls adding to the enchanting ambiance

as in the first and third examples, images can be abstract, leading the model to compensate through hallucinations. More examples from both datasets are presented in Appendix A. In general, the captions do a decent job of capturing the main concepts from the presented images. In some cases, the summarization model may misinterpret background details, as in the fourth example, or it may pull details from the music caps inference when present as in the third example. Another common failure case for the captioning process is creating a prompt with multiple options for the prompt, where the model outputs prompt A, followed by the word 'or' and prompt B.

Chapter 4: Image prompts from lyrics: multi-turn LLM interaction for image generation from music

In this chapter, I will discuss the currently deployed approach of image generation as performed by Vibe Video. The overall approach described in this chapter was largely developed by Ben Gillin for Vibe Video. My contributions to this work are primarily refinement, analysis, and engineering efforts to deploy the pipeline as an app which is currently available for interaction at `https://app.vibevideo.ai`. As such, this chapter will remain brief, but acts as a baseline for other approaches described in this paper.

The primary benefit of a multi-turn interaction approach is that is largely model-agnostic. While the currently deployed pipeline utilizes APIs provided by OpenAI and StabilityAI, the approach is easily adaptable to other models and APIs. As an example, we have witnessed wildly different outputs from the same prompts when provided to Stable Diffusion (Rombach et al., 2022) versus Midjourney[1]. Moreover, outputs at each stage of the pipeline can be provided and curated by the user, lending itself to a more interactive creative process. Inferences can be modified to better match user preference and hallucinations can be corrected before moving into subsequent stages in the pipeline. The LLM interaction also lends itself to other modalities. The approach can easily be adapted to create images for stories, poetry, or any other long form text. Despite these benefits, this approach solely relies on lyrics from music, which ignores the contextualization which comes from the surrounding music. This approach, therefore, cannot be used on instrumental music which lacks lyrics. This limitation and others are further discussed in section 4.3.1.

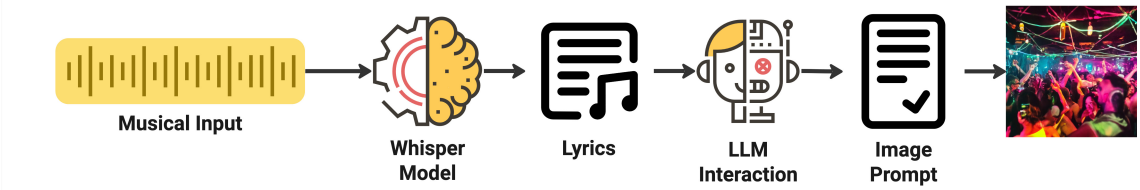


Figure 4.1: Flow of the multi-turn interaction. Lyrics are first extracted by a Whisper model from a musical input. The lyrics are used in a multi turn interaction with a large language model to synthesize a prompt for an image generation model, which in turn produces an image capturing the vibe of the music.

4.1 Multi-turn LLM interaction

The multi-turn interaction begins with the lyrics of a song. The lyrics can either be provided directly by the end user, but if the user provides audio, we extract the lyrics using Whisper (Radford et al., 2022)(Gandhi et al., 2023), which provides other useful information such as approximate time-stamping of the lyrics. The lyrics are then fed into a LLM with instructions to extract global information about the lyrics. The data extracted from the extractive prompts are injected into the final prompt to ensure similar style across multiple image generations. This approach could be further extended with cross-attention control (Hertz et al., 2022) to control stability between successive images, though this would remove the model-agnostic guarantees of the interaction. The final outcome of the multi-interaction is an image generation prompt which can be provided to any diffusion model to generate an image or video artifact.

4.1.1 Extractive Prompts

The first step of the interaction process is to extract global semantic information from the lyrics. For the rest of this chapter, these prompts will be referred to as extractive prompts and outputs from the process will be referred to as image prompts

[1]https://midjourney.com

Prompt Goal	System Prompt	User Prompt
Thematic Content	Suggest a theme and style for this song in a short paragraph. The theme should be a general idea or concept that the song conveys, and the style should be a visual style that would best represent the song.	$SONG
Visual Vibe	Given the theme and song, suggest a vibe which best captures the feeling of the song.	Theme: $THEME Song: $SONG
Color Palette	Infer a color pallete that would work well with the theme, style, and vibe. Make your responses as concise as possible.	Theme: $THEME Vibe: $VIBE Song: $SONG
Visual Metaphor	Develop a visual metaphor based on the provided song, art-style, and theme in as few tokens as possible.	Theme: $THEME Vibe: $VIBE Song: $SONG
Sentiment	Infer the main sentiment and emotions of the provided text in as few tokens as possible	$SONG

Table 4.1: Global prompts used in the current VibeVideo pipeline. Each word prepended with $ is replaced with the appropriate information. For example, $SONG is replaced with the lyrics and/or inferences from MusicCaps, $THEME is replaced with the inference from the Thematic Content prompt, and so on. When used in an interaction based approach, the system prompts are moved into the user prompt and the LLM is instructed to refer to the context window rather than providing the results each turn.

to differentiate the two flavors of prompting in this work. The extractive prompts in 4.1 are used to extract this information in one of two ways. The first method prompts the LLM in parallel with the prompts as system prompts and lyrics/prior information provided as user prompts. Once all extractive prompts receive an associated inference, they are synthesized into an image prompt. The second method prompts the LLM in a conversational manner consistent with the expected use of LLM chat agents. Some of these prompts, such as emotion extraction, are a zero-shot sentiment analysis task provided to the LLM in use. Others, such as notable themes, are more open-ended to the LLM. When available, more open-ended prompts are paired with a higher temperature to encourage more "creative" generations from the LLM. In each

prompt, the LLM is instructed to give an appropriate length for the prompt. For example, the art-style prompt requires the model to only generate the name of the style rather than a description of an art style. In this case, the model is asked to generate n art styles, which are then combined and ranked to provide multiple options to the user or to downstream prompts.

The prompts currently deployed in the system are found in 4.1. While the majority of these stand alone, there is some notion of hierarchy within the prompts. For example, the predicted sentiment output typically changes if the outcome from emotional content is included in the context. The same is true with art style, thematic content, and visual metaphor. In our current pipeline, the output of art style is provided as an input for the color palette, which assists the LLM produce an appropriate color palette for the outcome. When using a conversational approach, the LLM is instructed to refer to previous inferences in the context window to interpret these relations.

4.1.2 Fusion into Image Prompt and Animation Generation

Once all the extractive prompts have been generated, they are infused into a single prompt by providing a system message as instructions and the extractive prompts as the initial user message in a LLM interaction. The system message is integral to controlling the output of the LLM. In our current production system, we find that ChatGPT tends to produce longer outputs than necessary than our task requires. As of time of writing, the ChatGPT API does not provide a length penalty parameter and the max tokens parameter is unreliable as the model will stop mid-sentence. The current system prompt is as follows:

> Generate a prompt for an AI image generator inspired by the provided information with less than 75 output tokens. Be creative and do not mention lyrics in your response. Do not use phrases such as "Create an image" or "Generate an Image." Do not describe the metaphor or symbolism of

Figure 4.2: Images generated from the LyricToImage pipeline. For each image in the set, lyrics transcribed from source audio is provided to a LLM. The LLM is prompted to derive information relating to possible themes, art styles, sentiments, etc from the lyrics. The LLM is then asked to synthesize an image generation prompt using the information. The generated prompt is then fed into Stable Diffusion 2.1 for image generation.

> choices, simply describe what the generated image should look like. Do not use the following words: ...

The word exclusion list has been truncated for brevity, but is necessary to prevent generation of prompts which may trigger NSFW filters in image generation models. Once the final prompt has been generated, it can be used with any current image or video generation model to produce a visual artifact which has overlap with the input music.

4.2 Evaluation

While this approach was developed ad-hoc iteration over time primarily using qualitative evaluation, in this section I will describe evaluations I applied to this

approach. These evaluations are my primary contribution to the lyric to prompt approach. Proper evaluations grant a baseline for other approaches discussed in chapters 5 and 6 while spurring development of generating prompts from lyrics by giving objective measurements between iterations. As this is a creative task, quantitative evaluations can be tricky to clearly define. For example, is a high measure of similarity between lyrics and prompts preferable, or would it be better to optimize towards a lower score? Such questions suggest quantitative measurements are largely suggestive and open to many interpretations.

The multi-turn interaction from lyrics to prompt with the LLM in L2I is similar to a summarization and synthesis task. To measure the effectiveness of LLM, I report ROUGE (Lin) and BERT (Zhang et al., 2020) scores between the lyrics and generated extractive prompts, as well as the final prompt given to an image generator. These scores are open to interpretation when applied to questions of quality of the final output. To measure the extent of how much a generated output looks like a reasonable image generation prompt, I report the Jaccard distance between the body of outputs and a sampling of the DiffusionDB dataset (Wang et al., 2023b), a collection of over 2 million human-generated prompts for Stable Diffusion.

I should note that these evaluations are a slight abuse of the ROUGE and BERT scores, and they should not be interpreted as a measure of quality of these models. While the task is similar to a summarization, it is not necessarily preferable for the model to simply summarize lyrics. A certain amount of creativity in each output can result in more interesting final outcomes. The measurements provided by both BERT and ROUGE are instead used to measure similarity between different stages of the interaction pipeline. Furthermore, they are useful when comparing the performance of different models to get a sense of the tendencies of each model in the pipeline. Finally, they could also be used to determine the effect of tweaking individual extractive prompts, though I leave that to future work and do not perform that analysis here.

LLM	Theme	Vibe	Color	Metaphor	Sentiment	Lyrics
GPT4 Conversational	0.855	0.858	0.866	0.885	0.848	0.776
GPT4 Parallel	0.856	0.859	0.867	0.887	0.846	0.776

Table 4.2: BERT F1 scores on conversation-based flow and parallel prompting. There is no significant change in inferences between the two approaches.

4.3 Analysis

To understand the interaction between the final image prompt and extractive prompts, I report the BERT scores in table 4.2 between the final prompt and the extractive prompts and lyrics. Additionally, I report the difference between prompting GPT4 in a parallel vs conversational manner. The two approaches are measured, as it has been easier to prompt other models using a conversational manner rather than building unique interactions for each prompt. As seen in the table, there is no significant difference between prompting GPT4 in a conversational vs parallel manner, suggesting the two approaches can be considered identical. The conversational approach has the benefit of using a LLM's context window as a store of information, whereas the parallel approach must be provided with dependent information for each prompt. The parallel approach has the main benefit that multiple independent extractive prompts can be provided simultaneously, speeding up latency for the user. In general, the final image prompt shows the lowest dependency on the input lyrics, and a higher dependency on the extractive prompts. This shows the model is placing a higher importance on information extracted from the lyrics rather than the lyrics itself. The highest dependency is on the visual metaphor prompt.

Another point of interest to this approach is the performance of different language models on the L2I task. For this comparison, I run pipeline using GPT4, LLaMa-2-7B (Touvron et al., 2023), and Mistral-7b (Jiang et al., 2023) as the LLM. Table 4.3 reports results between different language models. In general, GPT4 tends to focus on the visual metaphor more than other extractive prompts. This makes sense as the visual metaphor is typically the culmination of the previous extractive prompts. However, Mistral shows a tendency to spread its focus out, drawing on in-

LLM	Theme	Vibe	Color	Metaphor	Sentiment	Lyrics	Mcaps
ROUGE-L							
GPT4-L	0.144	0.151	0.201	0.291	0.099	0.046	–
GPT4-M+L	0.132	0.139	0.192	0.261	0.095	0.042	0.108
Mistral+L	0.182	0.185	0.169	0.220	0.098	0.048	–
Mistral+M+L	0.184	0.191	0.171	0.221	0.093	0.049	0.120
LLaMA+L	0.224	0.229	0.185	0.253	0.099	0.079	–
LLaMA+M+L	0.220	0.233	0.181	0.257	0.100	0.079	0.138
BERT F1							
GPT4-L	0.856	0.859	0.867	0.887	0.846	0.776	–
GPT4-M+L	0.851	0.856	0.865	0.881	0.845	0.774	0.835
Mistral+L	0.872	0.871	0.864	0.881	0.851	0.777	–
Mistral+M+L	0.872	0.871	0.862	0.879	0.848	0.775	0.840
LLaMA+L	0.865	0.862	0.837	0.870	0.833	0.774	–
LLaMA+M+L	0.864	0.862	0.836	0.870	0.834	0.774	0.832

Table 4.3: ROUGE-L (top) and BERT (bottom) scores measuring final prompt dependence on global prompts, input lyrics, and MusicCaps inferences. Results are shown for prompts derived using lyrics (+L) or using MusicCaps inferences and lyrics (+M+L). The ROUGE-L metric provides insight in to how much language is repeated from each input source. The BERT metric measures the similarity between the prompt and input source.

formation across the extractive prompts. This is reflected through a lower ROUGE-L and BERT score for the visual metaphor, with increases in the scores for theme and vibe. LLaMA-2 is more likely to repeat language than the other models as seen in the ROUGE-L score for the theme, vibe, and visual metaphor. As a result, its focus is shared more evenly across the extractive prompts. Across the board, there is a lower similarity between the lyrics and final prompts, suggesting the models are relying more on the extractive prompts rather than the input lyrics.

In table 4.3, I show the pairwise BERT scores between the extractive prompts, lyrics, and music caps inference if it was used for production of the output prompt. This provides an intuition on the inter-prompt dependencies. There is no significant movement in the BERT scores with the introduction of the MusicCaps across the board. However, MusicCaps interactions have a slightly lowered dependence on the lyric score as shown with a consistent decrease in the lyric column. As prompts are dependent on each other through the prompting flow, this chart is as expected. For

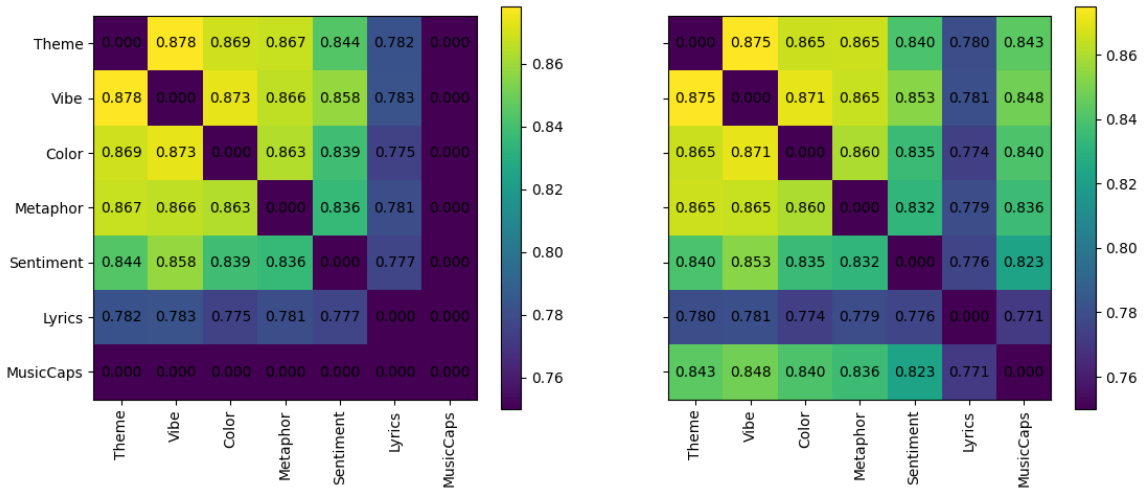


Figure 4.3: Pairwise BERT Score for all global prompts. For better scaling, the diagonal and unused features have been pushed to 0 rather than keeping them at 1. On the left is inferences generated without use of the MusicCaps inferences. On the right shows scores when MusicCaps is incorprated into the LLM interaction.

example, when the model is prompted for a color palette, it is given both the theme and the vibe as context, where the sentiment prompt is only given the lyrics. It is important to note that using BERT Score in this manner is an abuse of the metric.

4.3.1 Drawbacks and Limitations

While the current multi-turn LLM interaction creates interesting results, the method is far from perfect and contains drawbacks and Limitations which motivate the remainder of this thesis. Perhaps the most glaring limitation is the reliance on lyrics for guidance towards visual generation. Lyrics by themselves can contain high emotional content, but do not complete the song. The emotional content of lyrics themselves can change depending on the framing which instrumental music provides. There are many songs and entire genres of music which do not contain lyrics for which this process will fail.

The reliance on external APIs for generation is a double-edged sword for any

kind of service. On one hand, swapping out better performing products for old products is a quick process, necessitating only a simple library switch and fine-tuning of system-level prompts. However, service APIs can change in cost and quality without warning. While the current trend is for inference costs to decrease, this is often accompanied by rate limits and higher costs for higher rates of inference. The distribution of outputs from a model can change from weight updates, which are often opaque to the end user. Notably, studies suggest ChatGPT's performance on tasks changes over time (Chen et al., 2023b), which can affect the stability of outputs in our pipeline. Such issues have a downstream effect on any core product which utilizes these services.

Once the final prompt has been generated, it is fed directly into an image generation model to provide a visual artifact which matches the semantic qualities of the lyrics. The major challenge in this process is creating image prompts which are an appropriate length for the image model. For example, due to reliance on the CLIP (Radford et al., 2021) text encoder, Stable Diffusion can only utilize image prompts of 77 token or less. Outputs from the multi-turn process often go over this limit causing much of the prompt to be truncated. Longer prompts may therefore contain important information which may never be considered by the image generator. To some extent, simple engineering tricks can mitigate this issue, but condensing information into as few tokens as possible still results into computational efficiencies at scale.

In this chapter, I describe a method using interaction with a LLM to extract information from lyrics and use that information to synthesize a prompt for an image generation model. This approach benefits from being model agnostic and training free. Analysis shows that the pipeline typically results in images which are related to the source lyrics. However, the multi-turn approach can be slow and dependent on external resources, which can be undesirable for reasons above. In the next chapter, I address these issues by bypassing the multi-turn aspect and casting the problem as a music captioning task.

Chapter 5: Image Prompts from music: music captions as image descriptions

Many of the limitations of the LLM interaction approach described in chapter 4 can be avoided by finding a single model which can produce appropriate descriptions of images directly from music. Such a model would provide lower latency and avoid risks of quality drift of external APIs. In this chapter, I build a model which given an audio input produces a description of an image to accompany the image. While the desired outcome is an image description, this work can still be considered a music captioning task (Won et al., 2021; Doh et al., 2023). Prior work has been in finding appropriate music to match images (Suris et al., 2022) or to retrieve related images and music (Wu et al., 2022; Stewart et al., 2023). This work differs as it produces a description of the image as a prompt for image generation which can be used to find existing footage or to generate images.

5.1 Music Captioning

The audio captioning problem is a simple formulation: given an audio source, generate a caption which describes the audio. This task has benefitted from the release of the AudioCaps and Clotho datasets (Kim et al., 2019; Drossos et al., 2019), with provide many audio-natural language pairs. AudioCaps is much larger and is generated using captions synthesized from AudioSet (Gemmeke et al., 2017) tags, while Clotho contains higher-quality human generated captions. A similar approach for music is MusicCaps (Doh et al., 2023), which synthesizes music captions by prompting a LLM with tags retrieved from the Million Song Dataset (Bertin-Mahieux et al., 2011) before being fine-tuned on a high quality internal dataset of human annotations. While general audio captioning has a variety of approaches (Mei et al., 2021, 2022; Deshmukh et al., 2023), models solely focused on music remain limited.

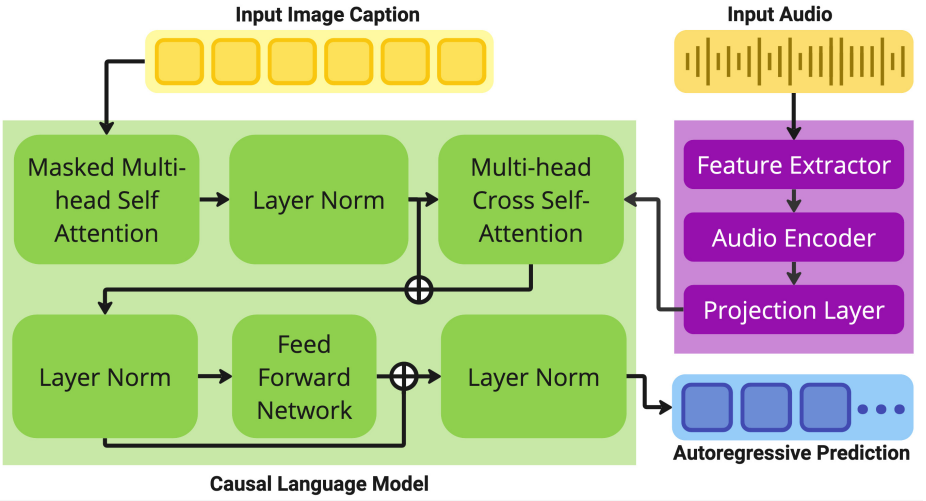


Figure 5.1: Generalized architecture of Music 2 Prompt models. An off-the-shelf audio encoder is used to create embeddings to be fed into a cross attention layer of a causal language model.

To my knowledge, there exists no dataset of music with accompanying images and image descriptions. As a result, I build the two datasets of music videos as described in 3 primarily for this and the music-to-image task described in chapter 6. These datasets rely on synthetic captions, which inherently limits the effectiveness of the model. As seen in 3.2, the CLIP score of these datasets leaves much to be desired. Moreover, any biases introduced by the models in the captioning process will be amplified by this model.

With these datasets, I train an encoder-decoder network with cross attention to produce prompts given a clip of music on the music2prompt (M2P) task. Given a clip of music, x, I first encode the audio with an off-the-shelf audio model, $\mathcal{E}$ to produce a representation, $\mathcal{E}(x) = z$. This representation is fed into an initialized cross-attention layer in a text decoder model, $\mathcal{D}$, which then autoregressively predicts a string of tokens, $\hat{y}$ using a cross-entropy loss

$$\mathcal{L} = -\sum_{t=1}^{T} \log p_\theta(y_t | y_{1:t-1}, x) \tag{5.1}$$

For the encoder, I choose MERT (Li et al., 2023b) and Audio Spectrogram Trans-

former (AST) (Gong et al., 2021a) as the audio encoders. The former is HuBERT-like audio encoder trained specifically on music and shows good generalization across multiple MIR tasks. Like HuBERT (Hsu et al., 2021), MERT learns through self supervision while iteratively clustering similar latent representations into clusters which are assigned labels. I report results on both the 95M parameter and 330M parameter model variants. In addition to the difference in their parameter sizes, the 330M model was trained on 160k hours of music, whereas the 95M model variant was trained on a 1k hour subset of the full dataset. AST is trained on general purpose audio and is a common choice as an audio encoder. Following the implementation of MusicCaps (Doh et al., 2023), I introduce cross attention to RoBERTa (Liu et al., 2019), casting it in the role of the decoder in an encoder-decoder set up.

5.2 Evaluation

To maintained comparability with the lyric-to-prompt task, I report many of the metrics as described in section 4.2. For each inference in the evaluation set, I report ROUGE and BERT scores between ground truth and generated captions. Where available, I further report the ROUGE and BERT score pairwise between the generated captions and intermediate information generated by MusicCaps and BLIP2 in the caption synthesis pipeline. These metrics provide insight into what aspects of the captioning pipeline are most important for the final output to inform further development of the M2P approach. To compare the performance of the model in producing outputs which appear to be image prompts, I also report the Jaccard distance between outputs and DiffusionDB. As the final goal of this approach is to create images which correspond to the music, I show examples in 5.2.

As in the previous evaluations, these metrics are open to wide interpretations and may not directly indicate quality of outputs. Instead, they provide a good insight into relative performance of the different approaches. In some regards, the Lyric2Prompt baseline is not directly comparable to the Music2Prompt model. Lyrics

Figure 5.2: Example images generated from prompts created by the RoMERTa-95M models. The top two rows are from the model variant trained exclusively on AnimatedMusicVideo, while the bottom rows are trained on YT8m-MV. The distribution of the underlying datasets is apparent in the examples. the AMV variant generates exclusively animated images whereas the YT8m-MV variant his a high propensity for prompting images of people, following the high occurrence of people in the dataset.

Model	ROUGE-1	ROUGE-L	BERT-F1	J-Sim	CLIP
RoMERTa-95M-AMV	0.216	0.178	**0.866**	**0.089**	**19.2**
RoMERTa-95M-YT8m-MV	**0.232**	0.184	0.864	0.076	18.2
RoMERTa-330M-AMV	0.209	0.177	0.864	0.075	18.2
RoMERTa-330M-YT8m-MV	**0.232**	**0.188**	0.865	0.074	17.7
ASTRoBERTa-AMV	0.214	0.176	0.867	0.087	18.6
ASTRoBERTa-YT8m-MV	0.181	0.151	0.860	0.0788	14.7

Table 5.1: Scores for the music to prompt task. The model names denote the model formation and the dataset used for training and evaluation. All models have relatively similar performance on the metrics shown. A consistently low J-Sim score suggests generated prompts do not look similar to human made prompts, which is consistent with qualitative analysis.

represent global information which can generally be considered to stay consistent throughout a song. The M2P models are provided 10-second clips as their input, and therefore rely on local information to generate the prompts. Moreover, lyrics may convey a different mood or timbre than the music to increase dramatic effect, which could inject more noise into the process.

5.3 Results

Table 5.1 scores for each of the models. Across all MERT models, the scores do not differ in any significant way. The results for AST are less consistent with a major drop across all metrics between the two datasets, though the results on ASTRoBERTA-AMV are consistent with the MERT models. Here, the reported CLIP score is significantly lower that the score reported in Table 3.2, suggesting the generated prompt's dependence on the image has decreased with the introduction of an audio encoder. The ROUGE Scores show a low correspondence between the predicted and source captions, though it is not insignificant. The examples in table 5.3 show high disparity between the source and generated captions, suggesting the ROUGE-1 score is mainly capturing connecting words and articles. Despite the disparities, the models report a strong BERTScore, suggesting a tenuous relation between the source truth and predictions.

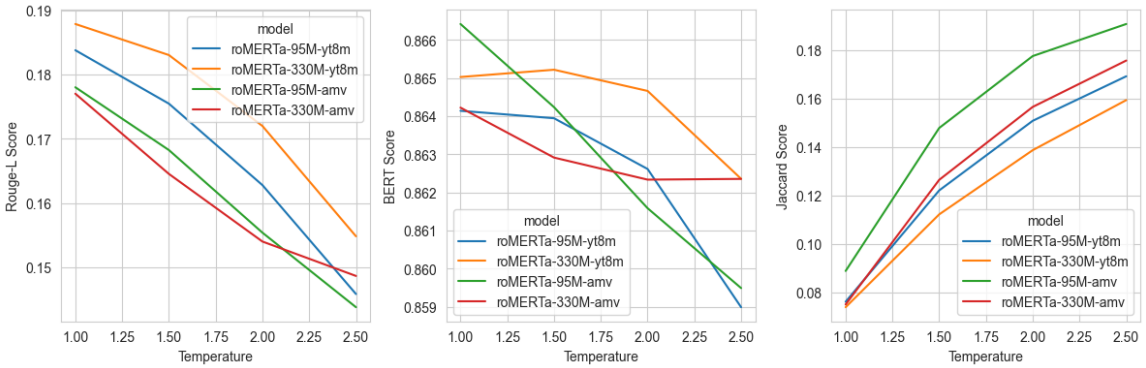


Figure 5.3: Effect of temperature on M2P inferences. As is the case with language models, increasing the temperature increases the diversity of the output. In this case, increasing temperature shifts the outputs towards the distribution of human-like Stable Diffusion prompts without significant drops in the BERT score. Moreover, larger MERT models appear to be more robust to changes in temperature than the smaller models, which drop in BERT F1 at a faster rate.

The examples generated in table 5.3 were generated with temperature, $\tau = 1.0$, resulting in repetitive language. For example, the word whimsical shows up many times in the generated examples; anime-style and cartoon are common prefix descriptors of any characters described within a prompt. This is even more pronounced in the model variants trained on the YT8m-MV dataset, reflecting the prevalence of human performers in the dataset images. I report the effects of temperature in figure 5.3. Rouge-L scores have an inverse relationship with the increase in temperature, suggesting increased diversity in the generated prompts, which is expected. Larger parameter models appear to be more robust to changes in temperature in terms of similarity to the ground truth, supported by the BERT score remaining fairly consistent for the 330M models, while falling for the 95M models. Finally, all models enjoy a significantly higher Jaccard score as temperature increases, though the distance between DiffusionDB and model outputs remains large. Future work could consider pretraining the RoBERTa model on the DiffusionDB dataset to encourage a higher Jaccard similarity score.

Cartoon characters gathered in a trio, two on the left and one on the right.
Generate an abstract image with a dark background adorned with intricate red line patterns, reminiscent of an R&B/pop scene with a masculine vocalist and a melodic keyboard melody. Incorporate subtle shapes and textures into the design, enhancing the sense of movement and depth.
A young girl in purple clothes sits cross-legged before a sunlit window, gazing thoughtfully at an unseen object. Bookshelves line the room, filled with colorful volumes and trinkets. A small dog naps peacefully nearby.
Anime-style woman with red hair and blue eyes, gazing directly at the viewer against a vibrant, abstract backdrop, surrounded by geometric shapes and vibrant colors.
Anime character in a chair, arms outstretched, wearing glasses and surrounded by abandoned books.

Table 5.2: Example caption outputs for the RoMERTa-95M-AMV model. More examples are provided in Appendix B

.

Qualitatively, the M2P generates reasonable prompts from the music inputs, following the data it was trained on, as seen in the examples in figure 5.2. However, the generated prompts result in low quality images across the board. This is partially due to the image generation model, however, many examples of high quality images exist from the same model exist online. As in the LLM interaction approach, the main benefit arises from the generated prompt being editable before being fed to an image generation model. Moreover, a natural language prompt is largely agnostic of the image generation model, and can be tailored to individual image generation models through human editing or further fine-tuning. Rather than producing a prompt, this model could easily be modified to produce image captions. This modification would allow these models to easily be rolled into the current Vibe interaction pipeline.

The M2P models are largely limited by the amount of data collected for their training. Other multimodal captioning models learn on data orders of magnitude larger than the datasets I have used throughout this process. For example. BLIP2 was trained on 129 million images (Li et al., 2023a) and MusicCaps was trained on over 4000 hours of music. As with all models in this thesis, the greatest boon to improving

the quality of captions produced from the various M2P models. Furthermore, higher-quality data is also necessary for better results. As seen in 3.2, the CLIPScore of the dataset is relatively low, suggesting the captioning process leaves much to be desired. Following the results from MusicCaps (Doh et al., 2023), a small set of higher quality prompts could also improve the final output of the model. In MusicCaps, the authors first trained on synthetic data before fine-tuning their model on high-quality human annotated music.

Training on music videos pulled from YouTube also opens the model up for generating prompts centered around copyrighted images or known people. Despite my best efforts remove such examples from the dataset, examples of known work persisted in the underlying datasets. As a result, when generating results for the models, references to Pokémon, video game characters, and musicians have occurred in generations. Such information is known to captioning models used to create the synthetic data used for training this class of models. Future work would need to improve the captioning process, as well as improve filtering known imagery prior to training time.

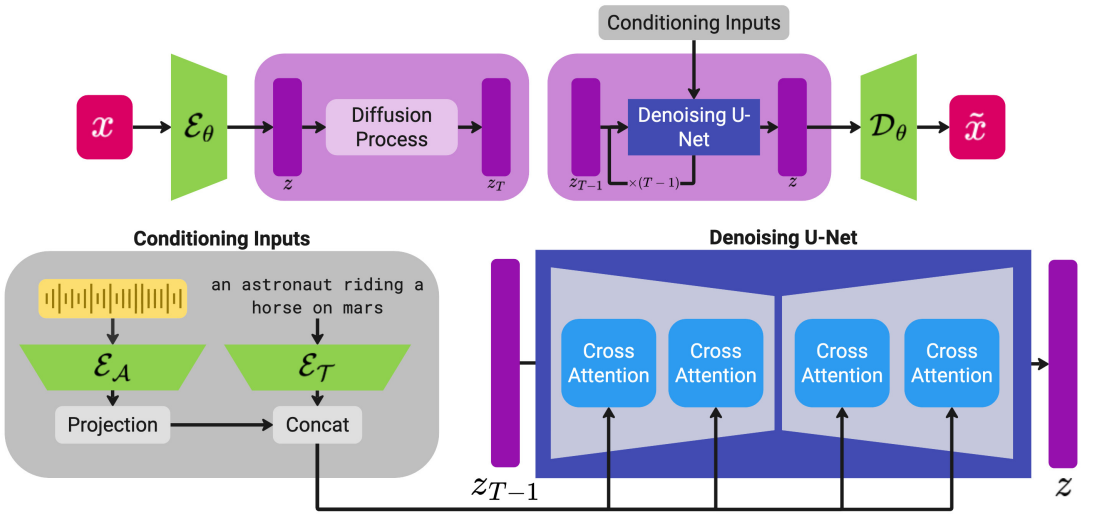
This chapter addresses the limitations presented in 4.3.1 by generating image prompts directly from music inputs. By conditioning a RoBERTa model on outputs from an audio encoder, the model learns image prompts from synthetic data, creating reasonable prompts given a musical input. While this process is highly data limited, it benefits from ease of implementation and directness of the approach. In the next chapter, I take the core idea of this chapter one step further and fine tune a Stable Diffusion model to directly generate an image conditioned on embeddings from an audio encoder.

Chapter 6: Music to image, conditioning text-to-image models with music

Both of the approaches outlined in chapters 4 and 5 rely on a multistage pipeline to go from a musical input to some visual representation. In this section, I consider closing the gap from music to image generation. The primary goal of this chapter is to find an end-to-end model which can output an image given audio, text, or an audio-text pair. Such an end-to-end model addresses many of the issues raised in sections 4.3.1 and 5.3. Most importantly, it removes the dependency on lyrics and conditions the output directly on the semantic information of the music. Second, this removes the dependency on external services, which can degrade or change over time without warning. An end-to-end model's quality can be assessed and tuned before released, and can be further fine-tuned as additional data is collected. This approach can be further enhanced as information gleaned from the lyrics in previous approaches can be introduced to this model via text. Ideally, a model trained in this approach would create images using only text prompts, only audio prompts, or a combination of both.

In this chapter I describe Music Controlled Imagery (MusCI), a family of models which attempt to condition image generation on both text and audio inputs. To my knowledge, there has been no significant work on this problem in literature. The closest approaches are work which conditions generated audio on video (Iashin and Rahtu, 2021), generating Foley sound from video (Yuan et al., 2023), modifying video with environmental sounds (Lee et al., 2023a), generating sound alongside video (Liu et al., 2023b), controlling video with speech (Lee et al., 2023a), or audio-reactivity (Lee et al., 2023b; Jeong et al., 2023).

Generating images from music has a number of significant challenges. First, high quality image generation is a function of both algorithms and data scaling. As in



language modeling, extremely large internet-scale datasets are used to train modern diffusion models, which puts such capabilities out of reach for most academic research labs. To condition generative models on a secondary modality may require a similar scaling of data in that modality. While video creates a large pool of music-image pairs to draw from, the quality of such data is unclear. While copyright-free music is available for training large generative music models, video is typically paired with copyrighted music, making development of such models difficult. A second difficulty is the diversity of the available data. In general, music videos skew heavily towards images of people playing instruments, with a long tail of truly unique inputs. Such a narrow grouping of training data could skew the outputs of a model towards such images. Third, the association between details or objects in images and music inputs is tenuous at best. While music is often described in terms of imagery, these are highly subjective and hard to quantify. For example, what snippet of music should be interpreted to be a picture of an astronaut riding a horse?

6.1 Extending Stable Diffusion Conditioning to Audio

MusCI seeks to condition a Stable Diffusion (Rombach et al., 2022) model on audio embeddings retrieved from a music representation model. Stable Diffusion first

Figure 6.1: Examples of the conditioned stable diffusion model. Top two rows arise from the model fine-tuned on the AMV dataset, while the bottom two rows are from the model fine-tuned on YT8m-MV.

trains a neural image compressor based on an autoencoder. Given an image x, the encoder, $\mathcal{E}$, compresses an image into some latent space $z = \mathcal{E}(x)$. The decoder, $\mathcal{D}$, is trained to reconstruct the image, $\hat{x} = \mathcal{D}(z)$. In contrast to earlier works, the diffusion process is then trained on the latent space with the following objective:

$$L_{LDM} := \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t)||_2^2\right] \tag{6.1}$$

where z_t is a noisy version of z and some timestep t, and $\epsilon_\theta(\cdot, t)$ is a time-conditional UNET (Ronneberger et al. (2015) as cited in Rombach et al. (2022)). To condition the final outcome on some other modality, they preprocess input y using an appropriate domain-specific encoder τ_θ to project the conditioning modality into the space of

a cross attention mechanism. Given the conditioning input, the training objective becomes

$$L_{LDM} := \mathbb{E}_{z,y,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))||_2^2\right] \tag{6.2}$$

where both τ_θ and ϵ_θ are jointly-optimized by this objective.

In this section, I describe MusCI-CLAP, which uses CLAP-music as the audio encoder to condition the image generation model. This model works on a simple general principle: given a music and text pair, produce an image which captures the qualities of both inputs. Formally, given a text, music pair, (y_T, y_M), I encode each with a modality-specific encoder, τ_* with a learnable linear projection layer to project the encodings into a common dimensionality. The two encodings are then concatenated into a single conditioning signal, $\mathrm{T}(y_T, y_M) = (\tau_T(y_T), \tau_M(y_M))$. The learning objective described in equation 6.2 is modified to give the general learning objective for these approaches:

$$L_{M2I} := \mathbb{E}_{z,(y_T,y_M),\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, \mathrm{T}(y_T, y_M))||_2^2\right] \tag{6.3}$$

For all three approaches, the primary difference is the audio encoder under study and the text encoder remains the CLIP Text tokenizer (Radford et al., 2021) as fine-tuned by the Stable Diffusion model.

MusCI-CLAP addresses the problem by taking the audio encoder of an off-the-shelf CLAP model (Wu et al., 2023b). The published version of CLAP is trained on general audio, where music represents a small fraction of the training set. Instead, I use weights released later which are trained entirely on music and text pairs. During training, I ensure the CLAP model remains unfrozen to help it adapt to the expected distribution of CLIP embeddings from the text encoder. I project the Audio embeddings into a shape of $\{N \times 10 \times d\}$ where d is the dimension of the cross embedding input of the image generation model. These audio embeddings are concatenated with the outputs of the CLIP text encoder before being fed into the cross embedding layer. As will be discussed in more detail in section 6.3.1, MusCI-CLAP is able to capture

some semantic information from the music and produces similar imagery given similar music inputs. However, the model suffers from an overall decline in quality of images and is unable to generate without text inputs.

6.2 Evaluation

For this task, I report a number of perceptual measurements commonly used in image generation models. I report the inception score (IS), which measures the diversity and quality of the generated images. A higher inception score indicates higher quality and diversity with the lowest possible value of 1. Fréchet Inception Distance (FID) also measures quality, but also provides a certain measure of realism. A lower FID shows that the generated images are close to realistic images, with high scores indicating unrealistic images. I also report the Kernel Inception Distance (KID), which attempts to improve over FID by incorporating kernels and unbiased estimators. Precision and recall is reported between the generated and ground truth representations. Precision measures the probability that a random generated image falls within the distribution of ground truth images, whereas recall describes the inverse relationship.

For all models, I sample prompts and audios from the evaluation split of AMV and YT8m-MV and generate 25,000 images. Since FID, KID and PRC are distributional measurements, they require a large amount of source images to be statistically robust. For these measure, I use images from the AMV and YT8m-MV train split. Since this introduces some bias in the metrics, I also report the measurements using the CIFAR-100 train split. Both KID and FID use a classification model to encode input images and measure their distance and are therefore sensitive to the feature extractor implementation. For all scores, I report using the torch-fidelity library (Obukhov et al., 2020) using `clip-vit-b-16` as the feature extractor.

Model	IS (↑)	FID (↓)	KID (↓)	Prec	Recall
SD-CLAP-AMV	1.069 ± 0.002	7.174	0.0256 ± 0.001	0.122	0.436
SD-CLAP-YT8M	1.084 ± 0.002	17.7936	0.066 ± 0.002	0.096	0.345
SD-2.1 (AMV)	1.006 ± 0.000	46.355	0.258 ± 0.004	0.002	0.004

Table 6.1: Perceptual quality metrics for M2I models with vanilla Stable Diffusion model as a baseline using the publicly available 2.1 checkpoint. This table reports the inception score (IS), Fréchet Inception Distance (FID), kernel inception distance (KID), precision and recall. All but the first are distributional measures which compare against a baseline set of images. For the baseline in this table, the ground truth frames from the videos are used.

Model	FID (↓)	KID (↓)	Prec	Recall
SD-CLAP-AMV	42.522	0.210 ± 0.002	0.004	0.000
SD-CLAP-YT8M	49.843	0.235 ± 0.003	0.006	0.000
SD-2.1	46.181	0.279 ± 0.004	0.001	0.003

Table 6.2: Perceptual metrics as in table 6.3, but using CIFAR100 as the base image distribution.

6.3 Results

I report perceptual metrics described in the previous section in tables 6.3 and 6.3. The former table reports against the base dataset. In these results, the inception scores suggest a small, yet insignificant, increase in perceptual quality. Both FID and KID suggest an increase in quality, but it is important to remember that these are distributional metrics which depend on the baseline. Comparing these scores between the two tables suggests that the quality has not greatly improved over the Stable Diffusion baseline, but rather the model fits closely with the distribution of training images. This is further supported by the precision and recall metrics which measure how likely an image was to arise from a second distribution. Both recall and precision seem to scale with the amount of data used to train the individual models, with a low score for the base model, which was trained on millions of images, and a high score for SD-CLAP-AMV which was trained on the least amount of images. In this context, a lower precision and recall are preferable as they are indicative of generalizability.

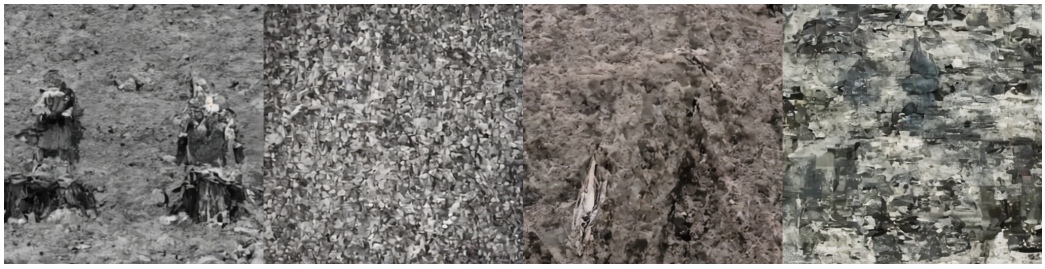


Figure 6.2: Example of images generated without any text conditioning. Before fine-tuning, Stable Diffusion has the ability to condition unconditionally, but loses this ability after being finetuned on audio inputs.

Table 6.3 tells a fuller story of the quality of the models trained in this work. Both FID and KID remain within a similar range of the stable diffusion baseline. This suggests the overall quality has not significantly changed from training. This suggests quality issues may not be entirely the fault of training, but rather from the choice of foundation model. Further work should study applying the fine-tuning procedure on other checkpoints or image models.

6.3.1 Qualitative Analysis

Examples of both models are provided in figure 6.1 with more examples provided in appendix C. In general, fine-tuning Stable Diffusion to condition on Audio Embeddings as well as text embeddings results in a loss of quality from the base Stable Diffusion model. When the model generates human beings, they commonly are presented in bizarre poses, with distorted facial features. Animated images tend to lose structure, with similar failure patterns seen the generated images. One such failure pattern is the characters appearing to have been abstracted to an extreme degree. Scenes of cities have buildings melting into each other, with the foreground lacking any distinction from the background. In short, the fine-tuning process seems to amplify the failure patterns of image generation models. Moreover, the model loses the ability to unconditionally generate images after only a few epochs of training as seen in figure 6.3.1.

Figure 6.3: Example showing consistency across prompts given the same source audio. The text prompts used to generate these images are a highly generic prompt: "an image of a spaceship flying through space." The provided audio is from different sections of the same song

I hypothesize the loss of quality in the MusCI models is due to two main problems. First, the audio embeddings produced by CLAP's audio encoder arise belong to a vastly different distribution than the CLIP Text encoder embeddings on which Stable Diffusion is trained. While a measurement of the exact distance between the two embeddings is beyond the scope of this work, it offers a reasonable explanation for image model's degeneration. Second, training is highly data limited, and I simply do not have access to the proper scale of data to account for the first issue. Stable Diffusion and other image generation models are trained on internet-scale amounts of information. While the exact number of training examples is not published, the number of training images is orders of magnitude larger than the datasets I have used for training.

Despite these failures, there is some evidence the conditioning audio has positive effect on the outcome. Figure 6.3.1 shows one such example. Both of the shown examples are generated using the same prompt and audio from the same song. The prompt itself is very generic: "An image of a spaceship flying through space." Generic prompts such as this one typically result in high variance in the style of the output image in the base Stable Diffusion model. These two images show a very similar

style, despite no stylistic information being provided by the prompt. This suggests the conditioning audio *may* be providing consistent style information to the image generation model.

6.4 Limitations and Future work

The MusCI model is highly limited in its ability. As discussed in the previous section, their overall quality is much lower than the foundation Stable Diffusion model. The primary limitation for improvement is the scale of data available for training, though data scaling itself does not guarantee success of future models. Compared to the other approaches, this method has the least amount of interpretability. LLM interactions provide a trace back of decisions which can be modified and influenced by the end user. Likewise, the M2P model provides a lesser amount by providing a caption which can be approved, curated, or modified before being fed to an image model. In this approach, images are created without knowing the exact reason as to the 'decision made' by the model. While the distance from audio to image is greatly reduced in this method, it is much harder to directly control the output as in other methods.

Another major limitation comes from the lack of analysis of the correspondence between music and generated images. As part of the work in this thesis, I made attempts to create a model which applies contrastive learning to images and music from the two datasets used throughout this thesis. The resulting inferences of the trained models were unfortunately too noisy to provide a meaningful metric. However, work such as Suris et al. (2022) and Stewart et al. (2023) show that models can be trained in this manner. Such a model is essential to measuring the results of the MusCI model and would aid greatly in improving this model. Additionally, human evaluations and preferences for images generated would provide better insight into the usability and quality of this model. Regrettably, time constraints dictate that this model is left to future work.

Finally, other audio encoders and diffusion models should be studied in their ability to achieve this task. The choice of using CLAP as the model for this task may not be the best. Further work should explore approaches as introduced by Wu et al. (2022), which distills an audio encoder to produce CLIP-like embeddings for audio. Such a model would produce embeddings belonging to a similar distribution as the CLIP text encoder which the the original Stable Diffusion model was trained on. This approach has the potential to address the pitfalls of the current model. Moreover, choosing a different Stable Diffusion checkpoint or architecture would have further implications for this approach. For example Stable Cascade (Pernias et al., 2023) uses a novel architecture which greatly speeds up training and inference, which could speed up iteration. Stable Diffusion XL[1] scales up the original Stable Diffusion Model and generates images of much higher quality than the checkpoint considered in this chapter, which could assist in maintaining quality for inference.

[1]https://huggingface.co/docs/diffusers/en/using-diffusers/sdxl

Chapter 7: Conclusion

In this thesis, I presented three methods to generate images with some correspondence to provided music. Each of these rely on the assumption that images and music both share overlapping semantic information–whether this be emotion, mood, style, or some other unknowable overlap. Music and images are typically found juxtaposed together in movies, music videos, album artwork, and in many other contexts. Music provides context in movies, to telegraph how we should feel about a character or foreshadow impending doom for the hero. Music videos provide images to add context to the music, or expand the creative reach of both artists and musicians. Album artwork can telegraph information such as a musician's style, the genre of the music, or even what the album is about. In each of these cases, music and images are used to strengthen the impact of the other.

The main contributions of this work are

- Described and analyzed the Lyric2Prompt task which extracts information from lyrics to synthesize a related image prompt for an image generation model.
- Introduced the Music2Prompt task which modifies the music captioning task to generate text which describes related imagery to accompany the provided music.
- Introduced the Music2Image task which seeks to produce images directly from music-text pairs by fine-tuning a Stable Diffusion image model on musical inputs.

In chapter 4, I described a method to generate prompts from lyrics using a multi-turn LLM approach. This method is currently available for interaction at `https://app.vibevideo.ai`. This approach leverages the large-scale training and

knowledge of LLM to extract information about a song from the lyrics. The key assumption to this approach is that lyrics contain similar semantic information to the associated music. While this method enjoys strong results and is largely model-agnostic, there are several limitations which motivate the other two methods presented in this thesis. The primary limitation, of course, is not all music contains lyrics. In the case of instrumental music, there is no information to provide to the LLM. This is mitigated somewhat with the inclusion of MusicCaps (Doh et al., 2023) in the process, but the approach is highly lyric dependent.

To mitigate these issues, I present a method which generates prompts directly from musical inputs in chapter 5. This method is inspired from other audio captioning approaches, most notably LP-MusicCaps. Rather than produce captions which describe the music, the Music2Prompt method generates an image caption given music. This addresses the dependence on lyrics from the previous approach by circumventing the LLM interaction altogether. While the approach produces reasonable captions, it tends to repeat many ideas which may be common to the training set. Moreover, tracking correspondence between the prompts and source music is difficult as the caption does not directly relate to the music, but ideally relates to an image which should accompany the music.

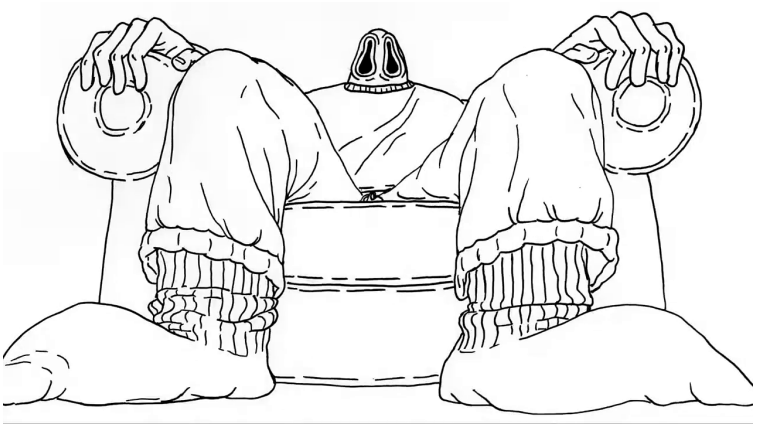
Finally, in chapter 6, I study conditioning image generation models on audio embeddings derived from music. Rather than deriving information from LLM interactions or generating intermediate prompts, this approach seeks to directly generate images given input music. Overall, this approach is the weakest of the three, as seen in the results and images provided by the model. Despite the shortcomings, this model is, to my knowledge, the first model which *generates* images conditioned on musical inputs. While there remains a large gap in the performance of this approach, it is a first step to conditioning image generation on modalities other than text.

This work can be improved by future work in several ways. A major theme throughout this work is the limitations imposed by music data. Future work could

address this by expanding both the quality and variety of music data used by model. Much of the data used in this thesis is pulled from YouTube and varies in terms of resolutions, audio fidelity, and content. While significant effort when in to filtering the datasets, several low-quality data pairs are likely present, which can affect training. In terms of variety, other music-image pairs can be found in album-art, social media posts, and music used paired with film. Moreover, while synthetic data was exclusively used in this work, the introduction of a high-quality data set for fine-tuning the model after training on synthetic data could likely improve the performance of each model.

More work is needed in cross modal retrieval between music and other modalities. LAION has released a variant of CLAP (Wu et al., 2023b) trained solely on music, which has been used throughout this paper. However, the corresponding published paper only describes the original version of the model, which was trained on mostly on speech and general audio. Thorough analysis on the CLAP music variant could perhaps elucidate problems with the models produced here. Cross modality retrieval between music and images is also in need of greater study. I attempted to create such a model as part of the thesis with little success. Other approaches for cross modal retrieval between images and music do not publish their models and are difficult to replicate. A major limitation of my work is the lack of a reliable measure between music and image correspondence.

Appendix A: Dataset Examples



- **Predicted Caption:** A cartoon figure sits in a cluttered room, strumming an acoustic drum set and plucking an e-bass, surrounded by scattered chairs and a footrest. In the background, a male singer emerges from the shadows, belting out melodies with added reverb and delay.

- **BLIP2 Inferences:** The image depicts a cartoon drawing of a man sitting on a chair, with his feet resting on the floor. There are multiple objects in the foreground and background, including an armchair, two chairs, one footrest, and more.

  Yes, this image has a cartoon style.

  a cartoon drawing of a person sitting in a chair

- **MusicCaps Inferences:** This song contains someone playing acoustic drums along with an e-bass. They both sound in the foreground. A male voice is singing in the high mids and with reverb and a little delay on his voice. Then the song breaks into the next part and an even higher pitch. This song may be playing as a live concert.

- **Predicted Captions:** Anime woman pondering over a mystical object beneath a vibrant night sky filled with twinkling stars and surrounding trees.

- **BLIP2 Inferences:** The image features a woman standing in front of the night sky, gazing up at the stars. She is wearing a dark green dress with long black hair and holding a star-shaped pendant hanging from her neck. There are several other objects scattered across the scene, such as two trees located to the left and one tree on the right.

  The image depicts a woman in an anime-style outfit, which is common for this type of artwork. This style may be due to the use of bright colors and stylized characters that are typically found in animated movies or TV shows.

  the girl is looking at something

- **MusicCaps Inferences:** This is a hardcore hip-hop music piece. There is a male vocal rapping in the lead. The melody in the beat is being played by the keyboard while the bass guitar is playing in the background. The rhythm is provided by a loud electronic drum beat. The atmosphere is aggressive. This piece could be used in the soundtrack of a crime movie or a TV series that takes place in Hungary.

- **Predicted Caption:** A man's face encircled by a whirl of vibrant colors. Pop art abstraction.

- **BLIP2 Inferences:** The artwork features a colorful, abstract image of a man's face in the middle of a swirling background. There are several objects visible on the left and right sides of the picture, including two people standing next to each other as well as an object that appears to be behind them.

  The image depicts a colorful, abstract painting in the style of pop art. Pop art is an artistic movement that emerged in the early 20th century and has become increasingly popular over time due to its vibrant colors, expressive nature, and ability to capture viewers' attention. In this case, the use of multiple brightly colored circles around the face adds visual interest while also contributing to the overall creative feel of the artwork.

  a colorful image of a man's face

- **MusicCaps Inference:** This is an electronic dance music piece. It is an instrumental piece. There is an ambient synth playing the melody. The rhythmic background consists of a slightly fast tempo electronic drum beat. The atmosphere is trippy. This piece could be playing in the background at a dance club.

- **Predicted Caption:** A whimsical Sesame Street-inspired character, clutching a cup amidst a scattered collection on a table.

- **BLIP2 Inferences:** the cartoon character is holding a cup in his hand. there are multiple cups visible, including one on the floor and another near the chair. The image depicts a cartoon character in the style of Sesame Street, an animated children's television series that has been running for more than 60 years. It features a blue and white furry creature wearing a light blue shirt holding a cup while sitting on a stool next to a table with other objects around it. There is no specific theme or message conveyed by this image; however, some viewers may interpret it as being humorous due to its comical nature.
  a cartoon character is holding a cup

- **MusicCaps Inference:** The low quality recording features a DJ scratching a hip hop song that consists of punchy kick and snare hits, shimmering hi hats, groovy bass and synth lead melody. There are crowd cheering noises in the background. It sounds energetic, aggressive and exciting.

- **Predicted Caption:** A man in the shadows, surrounded by piled crates and illuminated only by flickering streetlights.

- **BLIP2 Inferences:** The scene depicts a group of men standing in an alleyway, with several objects scattered around the room. There are multiple pieces of furniture and other items located throughout the area, including chairs, tables, bookshelves, and more. In addition, there is a snake lying on top of one of the stacks of boxes that can be seen from across the room.

  The image depicts a man standing in the dark with other people around him, which may suggest that it has a particular style or theme. For example, there is a sense of mystery and suspense due to the dim lighting as well as the presence of various objects such as books, boxes, furniture, etc. This could be indicative of an action-packed scene from a crime thriller movie like "The Dark Knight Rises."

  a man is standing next to some boxes

- **MusicCaps Inference:** This pop song features a male voice singing the main melody. This is accompanied by programmed percussion playing a simple beat. The kick is played on every count. Hand claps are played at every alternate

count. The bass plays the root notes of the chords. Synth chords are played in the background. This song can be played at a club.

Appendix B: Music2Prompt Examples

The following list comprises randomly sampled generations from the RoMERTa-95M-AMV model. All generations were generated with a temperature setting of 1.0

- Cartoon characters gathered in a trio, two on the left and one on the right.
- Generate an abstract image with a dark background adorned with intricate red line patterns, reminiscent of an R&B/pop scene with a masculine vocalist and a melodic keyboard melody. Incorporate subtle shapes and textures into the design, enhancing the sense of movement and depth.
- A young girl in purple clothes sits cross-legged before a sunlit window, gazing thoughtfully at an unseen object. Bookshelves line the room, filled with colorful volumes and trinkets. A small dog naps peacefully nearby.
- Anime-style woman with red hair and blue eyes, gazing directly at the viewer against a vibrant, abstract backdrop, surrounded by geometric shapes and vibrant colors.
- Anime character in a chair, arms outstretched, wearing glasses and surrounded by abandoned books.
- A whimsical scene of a cartoon character with blue hair and green eyes, dressed in a yellow shirt, amidst a backdrop of vibrant colors and playful details.
- A hauntingly lit cartoon scene with numerous masked figures huddled together, surrounded by looming monsters and ominous shadows amidst a backdrop of chaotic darkness.
- A cartoon character soaring through the sky, surrounded by billowing clouds and twinkling stars against a backdrop of twinkling stars. Create an image reflecting this whimsical scene with vibrant colors and playful details.

- A night sky filled with twinkling stars against a backdrop of absolute darkness.
- Generate an image of a futuristic scene with a dark backdrop illuminated by vibrant neon lights. In the foreground, hear the sound of an electric guitar melody playing in the background, accompanied by the rhythmic pulsing of electronic drums and the rhythmic thump of a bass guitar. Incorporate elements of
- A trippy urban scene with rap vocals and electronic drums. Create an image of a minimalistic green background with the empowering phrase "you can".
- A cartoon girl with green hair gazes introspectively through a kitchen window in an empty room, surrounded by scattered pots and pans, while two chairs and a table add depth to this whimsical setting. The sun sets behind her, casting warm hues upon the scene.
- A women dressed in casual attire sits in a shopping cart, observing her surroundings. Two figures stand by, creating an intriguing urban scene.
- A whimsical scene of a woman balancing on a floating boat amidst a whimsical underwater world.
- People dancing under vibrant red lights. A woman moves passionately in the foreground, her silhouette illuminated against the backdrop of swaying companions.
- Anime-style cartoon character with yellow hair and blue eyes, gazing intently at a mysterious object in a dimly lit room. Surroundings include two chairs, a table, and scattered objects.

Appendix C: Examples for MusCI models

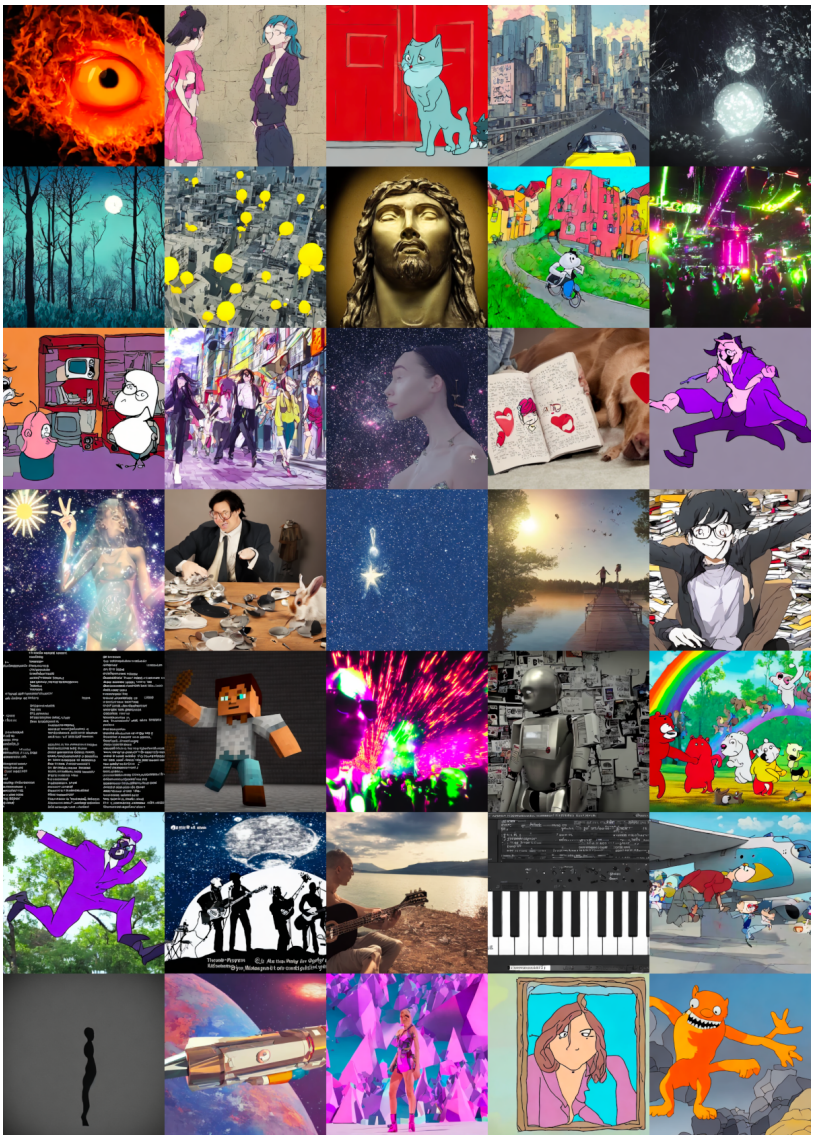


Figure C.1: AMV examples

Figure C.2: YT8m-MV examples

Works Cited

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark, September 2016.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating Music From Text, January 2023.

Kleanthis Avramidis, Shanti Stewart, and Shrikanth Narayanan. On the Role of Visual Context in Enriching Music Representations. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June 2023. doi: 10.1109/ICASSP49357.2023.10094915.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, October 2020.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio, July 2023.

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Dan Bigioi, Shubhajit Basak, Michał Stypułkowski, Maciej Zięba, Hugh Jordan, Rachel McDonnell, and Peter Corcoran. Speech Driven Video Editing via an Audio-Conditioned Diffusion Model, May 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets, November 2023.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A Language Modeling Approach to Audio Generation, September 2022.

Manuel Brack, Patrick Schramowski, Felix Friedrich, Dominik Hintersdorf, and Kristian Kersting. The Stable Artist: Steering Semantics in Diffusion Latent Space, May 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators, 2024.

Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval, July 2021.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies, August 2023a.

Lingjiao Chen, Matei Zaharia, and James Zou. How is ChatGPT's behavior changing over time?, October 2023b.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and Controllable Music Generation, June 2023.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High Fidelity Neural Audio Compression, October 2022.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An Audio Language Model for Audio Tasks, May 2023.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis, June 2021.

Sander Dieleman, Aäron van den Oord, and Karen Simonyan. The challenge of realistic music generation: Modelling raw audio at scale, June 2018.

Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang. CED: Consistent ensemble distillation for audio tagging, September 2023.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. LP-MusicCaps: LLM-Based Pseudo Music Captioning, July 2023.

Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial Audio Synthesis, February 2019.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An Audio Captioning Dataset, October 2019.

Xavier Favory, Konstantinos Drossos, Tuomas Virtanen, and Xavier Serra. COALA: Co-Aligned Autoencoders for Learning Semantically Enriched Audio Representations. June 2020.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An Open Dataset of Human-Labeled Sound Events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022. ISSN 2329-9304. doi: 10.1109/TASLP.2021.3133208.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. Distil-Whisper: Robust Knowledge Distillation via Large-Scale Pseudo Labelling, November 2023.

Xiaoxue Gao, Chitralekha Gupta, and Haizhou Li. Genre-Conditioned Acoustic Models for Automatic Lyrics Transcription of Polyphonic Music. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 791–795, May 2022. doi: 10.1109/ICASSP43922.2022.9747684.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets, December 2021.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, New Orleans, LA, March 2017. IEEE. ISBN 978-1-5090-4117-6. doi: 10.1109/ICASSP.2017.7952261.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer, July 2021a.

Yuan Gong, Yu-An Chung, and James Glass. PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3292–3306, 2021b. ISSN 2329-9290, 2329-9304. doi: 10.1109/TASLP.2021.3120633.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. 2014.

Curtis Hawthorne, Ian Simon, Adam Roberts, Neil Zeghidour, Josh Gardner, Ethan Manilow, and Jesse Engel. Multi-instrument Music Synthesis with Spectrogram Diffusion, December 2022.

Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: A Fast and State-of-the Art Music Source Separation Tool with Pre-Trained Models. 2019.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, March 2017. doi: 10.1109/ICASSP.2017.7952132.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross Attention Control, August 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning, March 2022.

Mojtaba Heydari, Frank Cwitkowitz, and Zhiyao Duan. BeatNet: CRNN and Particle Filtering for Online Joint Beat Downbeat and Meter Tracking, August 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, December 2020.

Jason A Hockman, Matthew E P Davies, and Ichiro Fujinaga. One in the Jungle: Downbeat Detection in Hardcore, Jungle, and Drum and Bass. 2012.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, June 2021.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. MuLan: A Joint Embedding of Music Audio and Natural Language, August 2022.

Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. Noise2Music: Text-conditioned Music Generation with Diffusion Models, March 2023a.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models, January 2023b.

Eric J. Humphrey and Juan P. Bello. Rethinking Automatic Chord Recognition with Convolutional Neural Networks. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 357–362, December 2012. doi: 10.1109/ICMLA.2012.220.

Vladimir Iashin and Esa Rahtu. Taming Visually Guided Sound Generation, October 2021.

Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The Power of Sound (TPoS): Audio Reactive Video Generation with Stable Diffusion, September 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel,

Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient Neural Audio Synthesis, June 2018.

Rainer Kelz, Sebastian Böck, and Gerhard Widmer. Deep Polyphonic ADSR Piano Note Transcription. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250, May 2019. doi: 10.1109/ICASSP.2019.8683582.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in The Wild. 2019.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. CREPE: A Convolutional Representation for Pitch Estimation, February 2018.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, October 2020a.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition, August 2020b.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient Training of Audio Transformers with Patchout. In *Interspeech 2022*, pages 2753–2757, September 2022. doi: 10.21437/Interspeech.2022-227.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually Guided Audio Generation, September 2022.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis, December 2019.

Jongpil Lee, Nicholas J Bryan, Justin Salamon, Zeyu Jin, and Juhan Nam. Metric Learning Vs Classification for Disentangled Music Representation Learning. October 2020.

Seung Hyun Lee, Sieun Kim, Innfarn Yoo, Feng Yang, Donghyeon Cho, Youngseo Kim, Huiwen Chang, Jinkyu Kim, and Sangpil Kim. Soundini: Sound-Guided Diffusion for Natural Video Editing, April 2023a.

Seungwoo Lee, Chaerin Kong, Donghyeon Jeon, and Nojun Kwak. AADiff: Audio-Aligned Video Synthesis with Text-to-Image Diffusion, May 2023b.

Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating A Multi-track Classical Musical Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications. *IEEE Transactions on Multimedia*, 21(2):522–535, February 2019. ISSN 1520-9210, 1941-0077. doi: 10.1109/TMM.2018.2856090.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, June 2023a.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Yike Guo, and Jie Fu. MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training, June 2023b.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models, February 2023a.

Jiawei Liu, Weining Wang, Sihan Chen, Xinxin Zhu, and Jing Liu. Sounding Video Generator: A Unified Framework for Text-guided Sounding Video Generation. *IEEE Transactions on Multimedia*, pages 1–13, 2023b. ISSN 1520-9210, 1941-0077. doi: 10.1109/TMM.2023.3262180.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019.

Matthias Mauch and Simon Dixon. A Corpus-Based Study of Rhythm Patterns. 2012.

Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. SampleRNN: An Unconditional End-to-End Neural Audio Generation Model, February 2017.

Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. Audio Captioning Transformer, July 2021.

Xinhao Mei, Xubo Liu, Mark D. Plumbley, and Wenwu Wang. Automated Audio Captioning: An Overview of Recent Progress and New Challenges. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):26, October 2022. ISSN 1687-4722. doi: 10.1186/s13636-022-00259-2.

Oriol Nieto, Matthew McCallum, Matthew E P Davies, Andrew Robertson, Adam Stark, and Eran Egozy. The Harmonix Set: Beats, Downbeats, and Functional Segment Annotations of Western Popular Music. 2019.

Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in PyTorch, 2020.

Puyuan Peng, Po-Yao Huang, Daniel Li, Abdelrahman Mohamed, and David Harwath. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv*, 2024.

Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An Efficient Architecture for Large-Scale Text-to-Image Diffusion Models. In *The Twelfth International Conference on Learning Representations*, October 2023.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification, October 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, December 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. AudioPaLM: A Large Language Model That Can Speak and Listen, June 2023.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive Learning of General-Purpose Audio Representations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879, June 2021. doi: 10.1109/ICASSP39728.2021.9413528.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A Dataset and Taxonomy for Urban Sound Research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 1041–1044, Orlando Florida USA, November 2014. ACM. ISBN 978-1-4503-3063-3. doi: 10.1145/2647868.2655045.

Justin Salamon, Oriol Nieto, and Nicholas J Bryan. Deep Embeddings and Section Fusion Improve Music Segmentation. November 2021.

Florian Schmid, Khaled Koutini, and Gerhard Widmer. Low-Complexity Audio Embedding Extractors, June 2023.

Roy Sheffer and Yossi Adi. I Hear Your True Colors: Image Guided Audio Generation, November 2022.

Shanti Stewart, Kleanthis Avramidis, Tiantian Feng, and Shrikanth Narayanan. Emotion-Aligned Contrastive Learning Between Images and Music, September 2023.

Didac Suris, Carl Vondrick, Bryan Russell, and Justin Salamon. It's Time for Artistic Correspondence in Music and Video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10554–10564, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.01031.

John Thickstun, Zaid Harchaoui, Dean Foster, and Sham M. Kakade. Invariances and Data Augmentation for Supervised Music Transcription, November 2017.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, September 2016.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning, May 2018.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. Audiobox: Unified Audio Generation with Natural Language Prompts.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers, January 2023a.

Yu Wang, Nicholas J. Bryan, Justin Salamon, Mark Cartwright, and Juan Pablo Bello. Who Calls The Shots? Rethinking Few-Shot Learning for Audio. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 36–40, New Paltz, NY, USA, October 2021. IEEE. ISBN 978-1-66544-870-3. doi: 10.1109/WASPAA52581.2021.9632677.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models, July 2023b.

Minz Won, Justin Salamon, Nicholas J Bryan, Gautham J Mysore, and Xavier Serra. Emotion Embedding Spaces for Matching Music to Stories. November 2021.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2CLIP: Learning Robust Audio Representations from Clip. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal*

Processing (ICASSP), pages 4563–4567, May 2022. doi: 10.1109/ICASSP43922.2022.9747669.

Shangda Wu, Dingyao Yu, Xu Tan, and Maosong Sun. CLaMP: Contrastive Language-Music Pre-training for Cross-Modal Symbolic Music Information Retrieval, June 2023a.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation, April 2023b.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete Diffusion Model for Text-to-sound Generation, July 2022.

Yi Yuan, Haohe Liu, Xubo Liu, Xiyuan Kang, Mark D. Plumbley, and Wenwu Wang. Latent Diffusion Model Based Foley Sound Generation System For DCASE Challenge 2023 Task 7, May 2023.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An End-to-End Neural Audio Codec, July 2021.

Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training, June 2021.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT, February 2020.

Hang Zhao, Chen Zhang, Bilei Zhu, Zejun Ma, and Kejun Zhang. S3T: Self-Supervised Pre-Training with Swin Transformer For Music Classification. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 606–610, May 2022. doi: 10.1109/ICASSP43922.2022.9746056.

Le Zhuo, Ruibin Yuan, Jiahao Pan, Yinghao Ma, Yizhi LI, Ge Zhang, Si Liu, Roger Dannenberg, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wenhu Chen, Wei Xue, and Yike Guo. LyricWhiz: Robust Multilingual Zero-shot Lyrics Transcription by Whispering to ChatGPT, July 2023.

Vita

Davin Lawrence is graduating from the University of Texas with a Bachelor and Master's degree in Computer Science. Prior to his scholastic career at UT, Davin received a degree in Music Business and Audio engineering from Austin Community College and worked for over a decade in the music industry in Austin, TX. Through his music career, he worked in music studios and music technology, as well as performed in several bands and most notably as a professional DJ. Post graduation, he hopes to incorporate his history in the music industry with the his new education to create musical instruments playable by all.

Address: davin.lawrence@utexas.edu

This thesis was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.