MELFUSION: Synthesizing Music from Image and Language Cues using Diffusion Models

Sanjoy Chowdhury*[1,3†] Sayan Nag*[2,3†] K J Joseph[3]
Balaji Vasan Srinivasan[3] Dinesh Manocha[1]

[1]University of Maryland, College Park [2]University of Toronto [3]Adobe Research

{sanjoyc,dmanocha}@umd.edu sayan.nag@mail.utoronto.ca {josephkj,balsrini}@adobe.com

Project page - https://schowdhury671.github.io/melfusion_cvpr2024/

Abstract

Music is a universal language that can communicate emotions and feelings. It forms an essential part of the whole spectrum of creative media, ranging from movies to social media posts. Machine learning models that can synthesize music are predominantly conditioned on textual descriptions of it. Inspired by how musicians compose music not just from a movie script, but also through visualizations, we propose MELFUSION*, a model that can effectively use cues from a textual description and the corresponding image to synthesize music.* MELFUSION *is a text-to-music diffusion model with a novel "visual synapse", which effectively infuses the semantics from the visual modality into the generated music. To facilitate research in this area, we introduce a new dataset MeLBench, and propose a new evaluation metric IMSM. Our exhaustive experimental evaluation suggests that adding visual information to the music synthesis pipeline significantly improves the quality of generated music, measured both objectively and subjectively, with a relative gain of up to* ***67.98%*** *on the FAD score. We hope that our work will gather attention to this pragmatic, yet relatively under-explored research area.*

1. Introduction

Music is an essential tool for creative professionals and content creators. It can complement and set the mood for an accompanying still image, animation, video, or even text descriptions while creating a social media post. Finding music that matches a specific setting, can indeed be an arduous task. A conditional music generation approach, that can synthesize a music track by analyzing the visual content and the textual description can find a wide range of practical applications in various fields including social media.

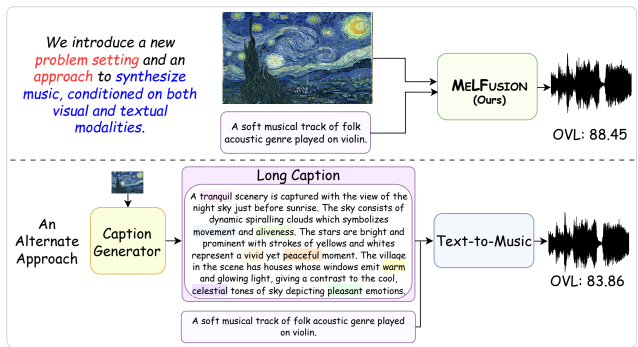

Figure 1. We present MELFUSION, a music diffusion model equipped with a novel "visual synapse", that can effectively infuse image semantics into a text-to-music diffusion model. This task indeed requires a detailed understanding of the concepts in the image. An alternate approach like using a caption generator to convert image to text space to be further used with existing text-to-music methods leads to a sub-optimal overall audio quality (OVL) score. Our approach can knit together complementary information from both modalities to synthesize high-quality music.

Inspired by the progress in generative modeling of images, music generation has also garnered significant attention from the community [1, 30, 42]. Recently, Agostinelli et al. [1], Copet et al. [4] proposed conditioning in the form of melody or humming. While Sheffer and Adi [43] pursue image-guided audio generation. Despite these efforts, music generation conditioned on multiple modalities like text and image, is largely uncharted.

Images are more expressive [13] than text-only information and capture more fine-grained semantic information about various visual aspects. For example, as depicted in Fig 1, to generate a musical track that goes well with a given image, without indeed using it, one has to make the tedious effort of producing long, descriptive captions (either generated by an image captioning model or human annota-

*Equal contribution.
†Work done during internship at Adobe Research.

tors) before employing a typical text-to-music generation model. Moreover, the model has to be supplied with critical attributes like *'tranquil'*, *'aliveness'* etc (highlighted in figure) to aptly capture the essence of the image. This poses a major bottleneck in the scalability of such systems especially for social media content creators and necessitates direct image conditioning with textual control in music generation.

Music is indeed different from generic audio. Music contains an arrangement of elements structured to form a coherent and complete entity. These musical elements include melody, harmony, rhythm, dynamics, and form [41, 45]. Unlike audio, music contains harmonies from different instruments forming intricate structures. Prior studies show [6, 11, 18, 36, 48, 52] that the human brain is extremely sensitive to disharmony. As a result, the margin of error especially in producing musical pieces is low compared to generic audio tracks. This makes music generation a harder task as the model should be equipped to control the fine-grained nuances of a composition involving melody, the interplay of the instruments, and genre.

An alternative to generating music would be to retrieve them. Retrieval-based systems [20, 33] struggle to 'match' the right track for a given input prompt thereby limiting their practical applicability in open-world scenarios primarily because (a) they tend to search from a pre-existing collection of tracks and (b) finding the correct association between the input prompt and the audio track can be challenging. The problem is inherently complex due to the multifaceted nature of music and the abstract associations between auditory experiences and other sensory modalities.

To overcome these shortcomings, we introduce the first music generation model that can be conditioned on image and text instruction. We observe that the features from a pre-trained text-to-image diffusion model that consumes the DDIM-inverted latent of the image can guide a text-to-audio diffusion model. Our key novelty is to facilitate this information exchange by incorporating a "visual synapse" to the text-to-music model, which includes a set of parameters that learn to combine the signals from both modalities.

We summarise our main contributions below:

(1) We formalize a novel task of generating music that is consistent with a reference image and an associated text prompt.

(2) We present MELFUSION, a novel diffusion model that can address this pragmatic task.

(3) We introduce MeLBench dataset comprising 11,250 $\langle \text{image}, \text{text}, \text{music} \rangle$ triplets. To the best of our knowledge, this is the largest collection of these three modalities. Further, we extend the MusicCaps [1] dataset by supplementing the text, and music pairs with suitable images extracted from corresponding YouTube videos or the web.

(4) In order to quantitatively establish the correspondence between the image-music pairs we propose a new metric IMSM. We demonstrate that the score follows human perception closely, through a user study.

(5) Finally, our exhaustive experimental results reveal that our approach outperforms existing text-to-music generation pipelines on both subjective as well as objective evaluation with a relative gain of up to **67.98%** on FAD score, thereby setting a new benchmark for multi-modal music synthesis.

2. Related Works

Music Generation Approaches: Music generation has garnered significant attention for a considerable amount of time. While some approaches [8, 35, 50] deploy GANs to tackle this task, Ycart et al. [51] introduced recurrent neural networks to model polyphonic music. Bassan et al. [2] proposed an unsupervised segmentation using ensemble temporal prediction errors. Jukebox [7] tackles the long context of raw audio using a multiscale VQ-VAE to compress it to discrete codes, modeling those using autoregressive Transformers. Another stream of work [14, 50] that predicts the MIDI notes to produce music has gained popularity in this space. However, the scope of these approaches is relatively limited as they need additional decoders to produce the musical pieces from the notations.

MusicLM [1] generates high-fidelity music from text descriptions by casting the process of conditional music generation as a hierarchical sequence-to-sequence modeling task. Mubert [34] is an API-based service that employs a Transformer backbone. The encoded prompt is used to match the music tags and the one with the highest similarity is used to query the audio generation API. MusicGen [4] comprises a single-stage transformer LM together with efficient token interleaving patterns. This eliminates the need for hierarchical upsampling. Despite significant progress, none of these approaches utilize the semantic information of images to condition the audio generation.

Diffusion Models for Music Generation: With the prolific success of diffusion models in conditional image generation, there have been recent efforts in music generation using them. Riffusion [12] base their algorithm on fine-tuning a stable diffusion model [39] on mel-spectrograms of music pieces from a paired music-text dataset. This is one of the first text-to-music generation methods. Moûsai [42] is a cascading two-stage latent diffusion model that is equipped to produce long-duration high-quality stereo music. Noise2Music [22] introduced a series of diffusion models, a generator, and a cascade model. The former generates an intermediate representation conditioned on text, while the later can produce audio conditioned on the intermediate representation of the text. MeLoDy [30] pursues an LM-guided diffusion model by reducing the forward pass bottleneck and applies a novel dual-path diffusion mode. We find that the visual guidance that is incorporated into our approach significantly enhances the music generation quality when compared to all these approaches. We elaborate this further in Sec. 4.4.

Diffusion Models for Audio Generation: Diffusion-based methods [23–25, 29, 31, 37] achieve remarkable results in speech synthesis too. FastDiff [23] deploys time-aware location-variable convolutions of diverse receptive field patterns to efficiently model long-term time dependencies with adaptive conditions. AudioLDM [32] is a text-to-audio system that is built on a latent space to learn continuous audio representations from contrastive language-audio pretraining (CLAP) embeddings. Ghosal et al. [17] simplifies the architecture of AudioLDM, and uses FLAN-T5 [3] as the text encoder. Another line of work [15, 49] involves text-conditional discrete diffusion models to generate discrete tokens as a representation for spectrograms. However, the quality of the sound produced by such methods leaves room for improvements in terms of both subjective and objective qualities, thereby limiting their practical usability. In contrast to these approaches, our method generates music samples conditioned on visual and textual signals.

3. Synthesizing Music from Image and Text

We propose to learn a conditional distribution $\mathcal{M}(\boldsymbol{w}|\boldsymbol{I}, \boldsymbol{Y})$, that can generate music waveforms $\boldsymbol{w}$ from an image $\boldsymbol{I}$ and a paired textual description $\boldsymbol{Y}$. We materialize $\mathcal{M}$ as MELFUSION, a diffusion model that can succinctly interleave the semantic cues from the image and textual modality while generating acoustically pleasing music.

Fig. 2 provides an overview of our approach. On a high level, our novel methodology consists of two subcomponents: 1) an approach to extract relevant visual information from the image conditioning $\boldsymbol{I}$ and 2) a method to induce this conditioning into the text-to-music generative model, in a parameter efficient way. We describe each of these in the subsequent subsections.

3.1. Extracting Visual Guidance

Latent diffusion models (LDMs) for text-to-image generation [39] have had phenomenal success in generating high-quality images that are well-grounded in their textual conditioning. We hypothesize that the latent representations and their transformations encode rich semantic knowledge, that can guide our audio diffusion model. In our exploration, we make use of a pre-trained Stable Diffusion model [39]. It contains a VQ-VAE [46] for encoding and decoding the image to the latent space, a text encoder, and a UNet [40] that carries out the diffusion process on the latent. The UNet contains an encoder, a bottleneck layer, and a decoder. Each encoder and the decoder further contain a set of blocks with cross-attention layers, self-attention layers, and convolutional layers. Given any intermediate latent image feature $\boldsymbol{f} \in \mathbb{R}^{(w\times h)\times d}$, a single self-attention [47] operation consist of $\boldsymbol{Q} = \boldsymbol{W}^q\boldsymbol{f}$, $\boldsymbol{K} = \boldsymbol{W}^k\boldsymbol{f}$, $\boldsymbol{V} = \boldsymbol{W}^v\boldsymbol{f}$:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}, \quad (1)$$

where d_k is the dimension of the query and key features. During cross-attention, the key and value matrices operate on the external text conditioning $\boldsymbol{c} \in \mathbb{R}^{s\times d_k}$: $\boldsymbol{K} = \boldsymbol{W}^k\boldsymbol{c}$, $\boldsymbol{V} = \boldsymbol{W}^v\boldsymbol{c}$. Here, $\boldsymbol{W}^q, \boldsymbol{W}^k$ and $\boldsymbol{W}^v$ are the attention weight matrices that transform either the image features or text conditions into the output of each block.

We want to transfer over the semantic information that is present within these attention layers corresponding to the image $\boldsymbol{I}$ into the music LDM. For this, we first invert $\boldsymbol{I}$ into the latent space using DDIM Inversion [44] to get $\boldsymbol{z}_T^I$. This will guarantee that we will be able to generate $\boldsymbol{I}$ from $\boldsymbol{z}_T^I$. Next, we do the reverse diffusion steps using a pre-trained text-to-image LDM starting from $\boldsymbol{z}_T^I$ and save the *self-attention features* $\boldsymbol{K} = \boldsymbol{W}^k\boldsymbol{f}$, $\boldsymbol{V} = \boldsymbol{W}^v\boldsymbol{f}$, to be injected into the music LDM. The intuition behind leveraging the self-attention features is that they control the feature transformations responsible for generating the visual semantics of the image. This is mathematically evident from Eq. (1). In the subsequent section, we elaborate on how we construct the "synapse" that can transfer the guidance information from $\boldsymbol{I}$ to the music-diffusion model.

3.2. Text-to-Music LDM with Visual Synapse

Inspired by recent text-to-audio [17, 32] generation approaches, our text-to-music model is also formulated as a latent diffusion model. During training, the music waveform $\boldsymbol{w}$ is first converted to a spectrogram $\boldsymbol{s} \in \mathbb{R}^{E\times F}$, which is a visual representation obtained via Fourier Transformation on $\boldsymbol{w}$. E and F denote the number of time slots and frequency slots respectively. Then we encode $\boldsymbol{S}$ using AudioVAE [32] to get a latent representation $\boldsymbol{z}_1^M \in \mathbb{R}^{C\times E/r\times F/r}$, where C is the number of channels and r is the compression level.

The forward diffusion process involves corrupting $\boldsymbol{z}_1^M$ using a Markovian noise process q, which gradually adds noise to $\boldsymbol{z}_1^M$ through $\boldsymbol{z}_T^M$ over T steps with the following Gaussian function:

$$q(\boldsymbol{z}_t^M|\boldsymbol{z}_{t-1}^M) = \mathcal{N}(\boldsymbol{z}_t^M; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}^M, \beta_t\mathbf{I}), \quad (2)$$

where β_t is a predetermined variance schedule. This iterative sampling process can be approximated by a deterministic non-Markovian process as follows [44]:

$$\begin{aligned} q(\boldsymbol{z}_t^M|\boldsymbol{z}_1^M) &= \mathcal{N}(\boldsymbol{z}_t^M; \sqrt{\bar{\gamma}_t}\boldsymbol{z}_1^M, (1-\bar{\gamma}_t)\mathbf{I}) && (3)\\ &= \sqrt{\bar{\gamma}_t}\boldsymbol{z}_1^M + \epsilon\sqrt{(1-\bar{\gamma}_t)}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) && (4) \end{aligned}$$

where $\gamma_t = 1 - \beta_t$ and $\bar{\gamma}_t = \prod_{r=0}^{t}\gamma_r$.

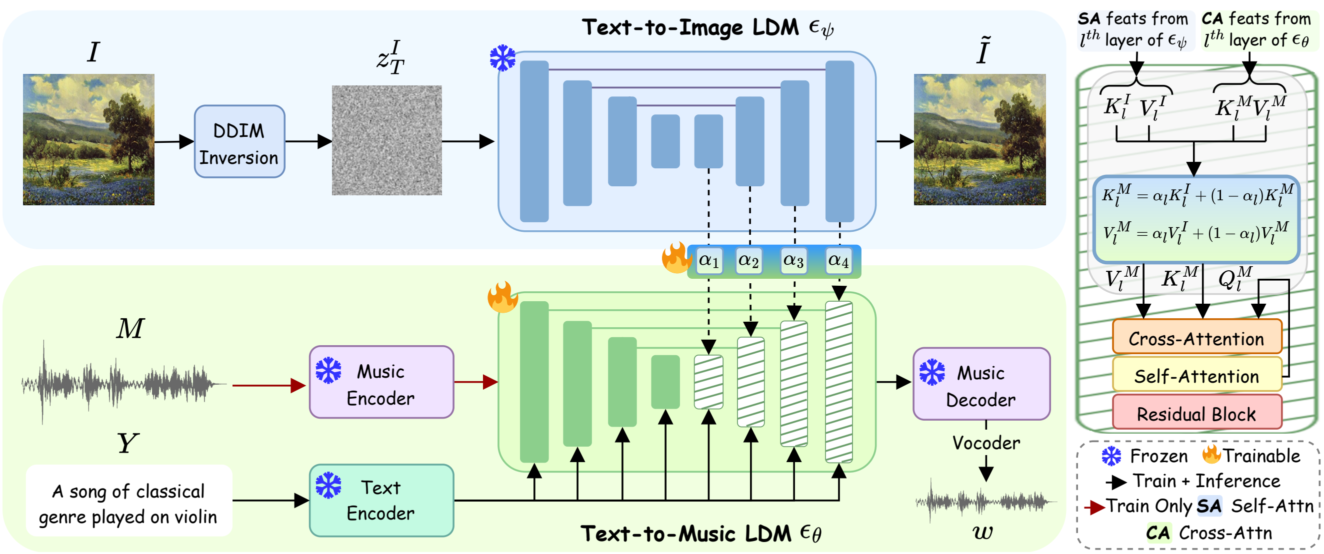

Figure 2. Our approach MELFUSION generates music waveform $\boldsymbol{w}$ conditioned on an image $\boldsymbol{I}$ and a given textual instruction $\boldsymbol{Y}$. Visual semantics from $\boldsymbol{I}$ is instilled into a text-to-music diffusion model (bottom green box) using a pre-trained and frozen text-to-image diffusion model (top blue box). The image $\boldsymbol{I}$ is first DDIM inverted into a noisy latent $\boldsymbol{z}_T^I$. The self-attention features from the decoder layers of the text-to-image LDM that consumes $\boldsymbol{z}_T^I$ is infused into the cross-attention features of text-to-music LDM decoder layers, modulated by learned α parameters. This fusion operation that happens in the decoder (green stripes) is detailed on the right side of the figure. The music encoder projects the spectrogram representation of the music to the latent space, and the music decoder retrieves back the spectrograms. Finally, a vocoder generates the waveform $\boldsymbol{w}$ from the spectrograms. Please refer to Sec. 3 for more details.

In the reverse diffusion process, an LDM $\epsilon_\theta(\cdot,\cdot,\cdot)$ (implemented as a UNet), learns to de-noise $\boldsymbol{z}_T^M \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to recover $\boldsymbol{z}_1^M$. The architecture of the UNet is kept exactly similar to the text-to-image UNet described in Sec. 3.1. To incorporate the additional guidance from image conditioning the *cross-attention* key and value features $\boldsymbol{K}_l^M$ and $\boldsymbol{V}_l^M$ in each of the decoder layer l of the UNet is modified as follows:

$$\boldsymbol{K}_l^M = \alpha_l \boldsymbol{K}_l^I + (1-\alpha_l)\boldsymbol{K}_l^M \tag{5}$$

$$\boldsymbol{V}_l^M = \alpha_l \boldsymbol{V}_l^I + (1-\alpha_l)\boldsymbol{V}_l^M, \tag{6}$$

where $\boldsymbol{K}_l^I$ and $\boldsymbol{V}_l^I$ are the *self-attention* features for the corresponding layer l of the image conditioning LDM from Sec. 3.1. Most importantly, the convex combination between these features is modulated by *learned layer specific α parameters*. We find that this simple formulation elegantly incorporates the image guidance into the text-to-music diffusion model without hampering its expressivity. As the α parameters facilitate the information exchange between the text-to-audio and text-to-image diffusion models, analogous to how a synapse in a nervous system facilitates the transfer of electrical and chemical signals between neurons, we refer to this handshake as the *visual synapse of a text-to-music LDM*.

Finally, the parameters of the LDM θ and the α parameters are trained end-to-end with the following loss function:

Algorithm 1 MELFUSION: Training

Input: Image: $\boldsymbol{I}$; Text: $\boldsymbol{Y}$; Music: $\boldsymbol{M}$; Pre-trained Text-to-Image LDM: $\epsilon_\psi(\cdot,\cdot,\cdot)$; Image Encoder: $\mathcal{E}^I(\cdot)$; Music Encoder: $\mathcal{E}^M(\cdot)$; Text Encoder: $\mathcal{T}^M(\cdot)$; Text-to-Music LDM: $\epsilon_\theta(\cdot,\cdot,\cdot)$; Number of Diffusion Steps: T.
Output: Trained Text-to-Music LDM: $\epsilon_\theta(\cdot,\cdot,\cdot)$, Learned mixing coefficient α, for each decoder layer l of LDM: $\{\alpha_l\}$.

```
z_T^I ← DDIM_Invert(E^I(I))                      ▷ Initialize Image Latent.
{ε_1^M, ··· ε_T^M} ← Forward_Diffusion(E^M(M))  ▷ Targets.
z_T^M ~ N(0, I)                                  ▷ Initialize Music Latent.
c ← T^M(Y)                                       ▷ Encoding Text.
for t ∈ {T, ··· , 1} do                          ▷ For each denoising step.
    for each layer l in decoder of LDM do
        K_l^I, V_l^I ← Self-attention features of ε_ψ(z_t^I, ∅, t).
        K_l^M, V_l^M ←Cross-attention features of ε_θ(z_t^M, c, t).
        K_l^M ← α_l K_l^I + (1 − α_l) K_l^M      ▷ Key update.
        V_l^M ← α_l V_l^I + (1 − α_l) V_l^M      ▷ Value update.
    L ← ||ε_t^M − ε_θ(z_t^M, c, t)||^2           ▷ Eq. (7)
    Optimize θ and all α parameters to reduce L.
return ε_θ(·,·,·), {α_l}.
```

$$\mathcal{L} = \mathbb{E}_{t\sim[1,T],\boldsymbol{z}_1^M,\boldsymbol{\epsilon}_t^M\sim\mathcal{N}(\mathbf{0},\mathbf{I})} ||\boldsymbol{\epsilon}_t^M - \epsilon_\theta(\boldsymbol{z}_t^M, \boldsymbol{c}, t)||^2 \tag{7}$$

3.3. Overall Framework

We summarize the overall flow of MELFUSION during training in Algorithm 1. Our key novelty is to introduce a

Algorithm 2 MELFUSION: Sampling

Input: Image: $\boldsymbol{I}$; Text: $\boldsymbol{Y}$; Pre-trained Text-to-Image LDM: $\epsilon_\psi(\cdot,\cdot,\cdot)$; Image Encoder: $\mathcal{E}^I(\cdot)$; Text Encoder: $\mathcal{T}^M(\cdot)$; Trained Text-to-Music LDM: $\epsilon_\theta(\cdot,\cdot,\cdot)$; Learned mixing coefficient α, for each decoder layer l of LDM: $\{\alpha_l\}$; Number of Diffusion Steps: T; Music Decoder: $\mathcal{D}^M(\cdot)$; Vocoder $\mathcal{V}(\cdot)$.

Output: Music Waveform: $\boldsymbol{w}$

1: $\boldsymbol{z}_T^I \leftarrow$ DDIM_Invert$(\mathcal{E}^I(\boldsymbol{I}))$ ▷ *Initialize Image Latent.*
2: $\boldsymbol{z}_T^M \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ *Initialize Music Latent.*
3: $\boldsymbol{c} \leftarrow \mathcal{T}^M(\boldsymbol{Y})$ ▷ *Encoding Text.*
4: **for** $t \in \{T, \cdots, 1\}$ **do** ▷ *For each denoising step.*
5: **for** each layer l in decoder of LDM **do**
6: $\boldsymbol{K}_l^I, \boldsymbol{V}_l^I \leftarrow$ Self-attention features of $\epsilon_\psi(\boldsymbol{z}_t^I, \emptyset, t)$.
7: $\boldsymbol{K}_l^M, \boldsymbol{V}_l^M \leftarrow$Cross-attention features of $\epsilon_\theta(\boldsymbol{z}_t^M, \boldsymbol{c}, t)$.
8: $\boldsymbol{K}_l^M \leftarrow \alpha_l \boldsymbol{K}_l^I + (1-\alpha_l)\boldsymbol{K}_l^M$ ▷ *Key update.*
9: $\boldsymbol{V}_l^M \leftarrow \alpha_l \boldsymbol{V}_l^I + (1-\alpha_l)\boldsymbol{V}_l^M$ ▷ *Value update.*
10: $\boldsymbol{z}_t^M \leftarrow \boldsymbol{z}_t^M - \epsilon_\theta(\boldsymbol{z}_t^M, \boldsymbol{c}, t)$ ▷ *Reverse Diffusion Step.*
11: $\boldsymbol{s} \leftarrow \mathcal{D}^M(\boldsymbol{z}_0^M)$ ▷ *Generate Spectrograms.*
12: $\boldsymbol{w} \leftarrow \mathcal{V}(\boldsymbol{s})$ ▷ *Generate Waveform from Spectrograms.*
13: **return** $\boldsymbol{w}$.

channel through which we can guide the text-to-music diffusion model toward the semantic concepts contained in the corresponding image conditioning. This "synapse" is detailed in Line 7 to Line 10. The rest of the algorithm follows the standard LDM training flow.

During inference, we make use of the trained text-to-image and text-to-music diffusion models, along with the learned α parameters. As seen in Lines 8 and 9 in Algorithm 2, the cross-attention features of the text-to-music LDM decoder are updated to incorporate the visual conditioning in each denoising step. Once the denoising (Line 10) is complete, the latent representation is projected back into a spectrogram using the decoder of Audio VAE [32], and then the waveform is generated using HiFi-GAN vocoder [26] in Lines 11 and 12 respectively.

4. Experiments and Results

To complement our newly introduced problem setting which generates music conditioned on visual and textual modality, we introduce a new dataset, a new evaluation metric, and come up with a strong baseline by extending a state-of-the-art text-to-music method to consume image modality. We explain each of these in the subsequent sections.

4.1. Datasets

To the best of our knowledge, there is no publicly available dataset that contains the $\langle$Image, Text, Music$\rangle$ triplets that are required to train and evaluate MELFUSION. We collect a new dataset MeLBench, which contains 11,250 manually annotated triplets of $\langle$Image, Text, Music$\rangle$. Further, we extend the MusicCaps [1] dataset which contains $\langle$Text, Music$\rangle$ pairs by adding the corresponding image.

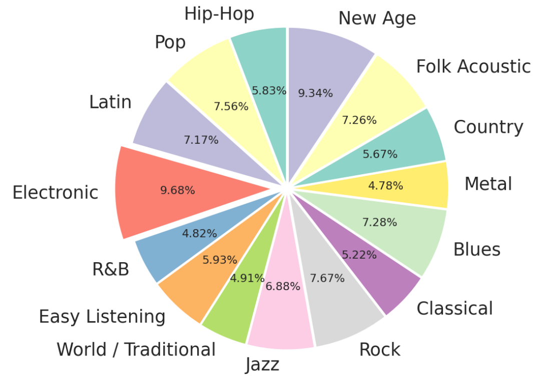

Figure 3. The distribution of different genres in MeLBench.

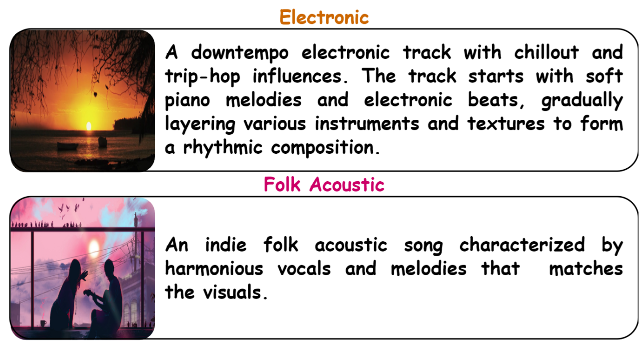

Figure 4. Some image and text pairs from MeLBench. We include more examples in the Appendix.

MeLBench: We hired 18 professional annotators to find 10-second snippets of YouTube videos corresponding to 15 predefined genres. The annotators are trained musicians with at least 5 years of practice. For each of these videos, they were asked to provide (a) a free-form text description for up to three sentences, expressing the composition and (b) any other music-related details such as describing the genre, mood, tempo, singer voices, instrumentation, dissonances, rhythm, etc. A carefully selected frame and music from the snippet along with text description from annotators forms $\langle$Image, Text, Music$\rangle$ triplets. We perform strict sanity checks to ensure the quality of these triplets in MeLBench. Fig. 4 shows some image and text samples from the dataset and Fig. 3 shows the distribution of different genres in MeLBench. Before annotating YouTube snippets (containing music-albums, art-performance, ensembles etc.), they were asked to check for complementary relevance between visuals and music. Further, we perform manual validation to filter lower-quality samples. We include more examples and more statistics of the dataset in the Appendix.

Extended MusicCaps: MusicCaps [1] is a subset of the AudioSet [16] dataset, which contains music and a corresponding textual description of the same. We carefully choose two images from the web or YouTube that can go well with each datapoint in MusicCaps, thereby extending $\langle \text{Text}, \text{Music} \rangle$ pairs to $\langle \text{Image}, \text{Text}, \text{Music} \rangle$ triplets. We defer more details to the Appendix.

4.2. Evaluation Metrics

We use objective evaluation and human subjective evaluation metrics to measure the efficacy of MELFUSION.

4.2.1 Objective Evaluation: Following previous works [17, 28, 32], Fréchet Audio Distance (FAD), Fréchet Distance (FD) and KL Divergence scores are used for objective evaluation. FAD [28] is a perceptual metric that is adapted from Fréchet Inception Distance (FID) for the audio domain. It uses a VGG-like backbone [19] for feature extraction. FD is similar to FAD but uses PANNs [27] as the feature extractor. Unlike reference-based metrics, FAD and FD measure the distance between the generated audio distribution and the real audio distribution without using any reference audio samples. On the other hand, KL Divergence [28] is a reference-dependent metric that computes the divergence between the distributions of the original and generated audio samples based on the labels generated by a pre-trained classifier. While FAD is more related to human perception, KL Divergence captures the similarities between the original and generated audio signals based on broad concepts present in them.

FAD, FD, and KL Divergence score captures the 'goodness' of generated music, while it doesn't measure whether the generated music is consistent with the image conditioning. We identify this as a gap and propose IMSM metric.

Image Music Similarity Metric (IMSM): CLIP score is one of the widely used metrics for measuring the similarity between an image and a corresponding textual description. N pairs of images and texts are passed through respective encoders (pre-trained using CLIP loss [38]) to obtain corresponding feature embeddings which are used to compute CLIP score matrix $\mathcal{A}_{\text{CLIP}} \in \mathbb{R}^{N \times N}$. In a very similar fashion, CLAP scores are computed amongst N audio-text pairs yielding CLAP score matrix $\mathcal{A}_{\text{CLAP}} \in \mathbb{R}^{N \times N}$ [10]. It is worth noting that in both the matrices the columns represent text modality. This motivates us to develop a metric IMSM, which is a measure of the perceptual similarity between given image-music pairs bridged by the text modality. In particular, we use CLIP image and text encoders which are contrastively aligned [38] to compute the image and text feature embeddings. As a second step, we leverage language as the bridging modality by freezing the CLIP text encoder and aligning the music (audio) encoder via contrastive training [10]. Finally, for $\langle \text{Image}, \text{Text}, \text{Music} \rangle$ pairs we obtain IMSM by suitably combining $\mathcal{A}_{\text{CLIP}}$ and $\mathcal{A}_{\text{CLAP}}$ using the given mathematical expression:

$$\mathcal{A}_{\text{IMSM}} = \mathcal{A}_{\text{CLIP}} \, \mathcal{A}_{\text{CLAP}}^{T} \tag{8}$$

4.2.2 Subjective Evaluation: Following earlier works in text-to-audio generation [17, 28, 32], we use overall audio quality (OVL) and relevance to image-text inputs (REL) to analyze the results of our subjective user study involving 75 participants. They were presented with 100 randomly generated samples from MELFUSION. Each of the metrics (OVL and REL) is a score between 1-100 with 1 being the lowest. For the OVL score, the users are asked to assign a score based on how perceptually realistic the generated audio is, while the REL score requires them to carefully examine the image and the text prompts before providing a rating based on their relevance with the generated music. We add more details of the user study in the Appendix.

4.3. Baseline Methods

We compare MELFUSION against strong baselines to test its mettle. To the best of our knowledge, there doesn't exist a music diffusion model that is conditioned on visual and textual modality. Hence, we introduce two baselines: 1) caption the image with Instruct BLIP [5] and then pass it along with the caption to MusicLM [1]. We call this baseline **MusicLM + InstructBLIP**. 2) we adapt a recent text-to-audio diffusion model TANGO [17] into our setting, as explained next. We call its modified version TANGO++. Further, we compare ourselves with 7 other text-to-music approaches. We elaborate on them below:

TANGO++: TANGO [17] is a powerful text-to-audio model based on LDMs. They condition the diffusion model on text embeddings from FLAN-T5 [3] text encoder $\boldsymbol{z}_{text}$. To facilitate joint conditioning from text and image $\boldsymbol{I}$, we embed $\boldsymbol{I}$ to the latent space as $\boldsymbol{z}_{image}$ using a ViT [9] based CLIP encoder, and align them together through an Image-Text Contrastive loss. Once they are aligned, both the embeddings are fused and the LDM is jointly conditioned.

Text-to-Music Baselines: To bring out the utility of conditioning on both visual and textual modality, we compare MELFUSION with seven other text-to-music methods too: Riffusion [12], Mubert [34], MusicLM [1], Moûsai [42], Noise2Music [22], MeLoDy [30] and MusicGen [4]. We provide more details of each of these in the Appendix.

4.4. Results

We present exhaustive objective and subjective comparison of MELFUSION with the baseline approaches in Tab. 1. When compared with text-to-music approaches in the first section of the table, our results show the significant utility of adding extra visual conditioning on the generations. The fine-grained contextual information from visual modality is able to supplement the information from the corresponding text, thereby enhancing the quality of music generation. Further, MELFUSION is able to consistently outperform MusicLM + InstructBLIP and TANGO++ (which has similar conditioning as ours). This highlights the efficacy

Model	Txt	Img	MusicCaps						MeLBench					
			Objective metrics				Subjective metrics		Objective metrics				Subjective metrics	
			FAD↓	KL↓	FD↓	IMSM↑	OVL↑	REL↑	FAD↓	KL↓	FD↓	IMSM↑	OVL↑	REL↑
Riffusion [12]	✓	✗	13.40	1.19	-	-	79.48	75.60	14.06	1.42	32.64	-	80.11	76.26
MuBERT [34]	✓	✗	9.60	1.58	-	-	77.59	77.93	-	-	-	-	-	-
MusicLM [1]	✓	✗	4.00	1.01	-	-	81.51	82.65	3.62	0.93	23.44	-	83.86	84.27
Moûsai [42]	✓	✗	7.50	1.59	-	-	75.94	77.33	9.13	1.63	31.51	-	75.11	74.32
Noise2Music [21]	✓	✗	2.13	-	-	-	81.13	79.88	-	-	-	-	-	-
MeLoDy [30]	✓	✗	5.41	-	-	-	80.61	79.25	-	-	-	-	-	-
MusicGen [4]	✓	✗	3.40	1.23	-	-	83.57	83.18	3.28	1.21	23.60	-	84.61	83.25
MusicLM + InstructBLIP [5]	✓	✓	4.12	1.18	25.68	0.55	80.21	79.85	3.88	1.07	24.96	0.63	81.18	82.42
TANGO++	✓	✓	3.05	1.17	23.91	0.68	84.62	83.96	2.93	1.14	22.16	0.71	85.52	84.81
MELFUSION (Ours)	✓	✓	**1.12**	**0.89**	**22.65**	**0.76**	**86.78**	**85.92**	**1.05**	**0.72**	**20.49**	**0.83**	**88.45**	**87.39**
$\Delta_{\text{MELFUSION- MusicGen}}$	-	-	+67.05%	+27.64%	-	-	+3.84%	+3.29%	+67.98%	+40.49%	+13.17%	-	+4.53%	+4.97%

Table 1. Our proposed approach MELFUSION offers significant gains over state-of-the-art text-to-music methods (first section), and our adapted text-and-image conditioned baselines (second section) across multiple objective and subjective metrics on two datasets. IMSM is applicable only when the model is conditioned on visual modality. We skip comparison with MuBERT, Noise2Music, and MeLoDy on MeLBench dataset as their codebases are not public. Please refer to Sec. 4.4 for more details.

of our visual synapse, which infuses the right amount of visual conditioning to enable the model to synthesize perceptually congruent music tracks. Captions from InstructBLIP are superfluous and vague when compared to expert-annotated, high-quality MusicCaps captions on which MusicLM is trained. This distributional shift leads to a performance drop as shown in the table. TANGO++ uses contrastive loss to align CLIP image features and FLAN-T5 text features, further, we use simple addition for joint conditioning – these design choices can be sub-optimal.

Our subjective human evaluation in Tab. 1 also suggests that conditioning the music generation on both visual and textual modality improves its perceptual quality.

5. Discussions and Analysis

5.1 Analyzing the Design Choice of α parameters: α parameters introduced in Eq. (6) controls how the self-attention features from the blocks within the text-to-image diffusion model interact with the cross-attention features from the text-to-music diffusion model. MELFUSION has one alpha parameter per block within the decoders of both diffusion models. We vary this design choice in Tab. 2. Attaching the synapse in the decoder offers better performance. This is because the decoder controls the major transformations that contribute to generating the image. Further, learning different α per block helps to learn block-specific mixing co-efficient, which slightly improves the performance.

Encoder	Decoder	Extended MusicCaps			MeLBench		
		FAD ↓	KL ↓	FD ↓	FAD ↓	KL ↓	FD ↓
Same α for all blocks.							
✓	✗	3.22	1.23	24.01	2.01	1.01	21.96
✓	✓	2.71	1.13	23.31	1.27	0.87	21.04
✗	✓	2.79	1.14	23.44	1.29	0.87	21.13
Different α for each block.							
✓	✗	2.03	1.10	23.36	1.92	0.93	21.28
✓	✓	1.13	0.94	22.81	1.07	0.76	20.53
✗	✓	**1.12**	**0.89**	**22.65**	**1.05**	**0.72**	**20.49**

Table 2. We systematically analyze our design choice of learnable α parameters. We vary the position of the synapse: encoder or decoder and also study whether we need the same or different α parameters for each block within them.

Learning Rate	Extended MusicCaps			MeLBench		
	FAD ↓	KL ↓	FD ↓	FAD ↓	KL ↓	FD ↓
0.5e-6	3.12	1.21	23.26	2.01	1.14	21.95
0.5e-4	1.38	0.95	22.81	1.39	0.88	20.86
1e-6	2.56	1.17	23.11	1.86	1.10	21.71
1e-5	**1.12**	**0.89**	**22.65**	**1.05**	**0.72**	**20.49**

Table 3. Sensitivity analysis on the learning rate for α parameters.

We also perform a sensitivity analysis on the learning rate (LR) used while learning α parameters in Tab. 3. Based on this analysis, we use a LR of $1e-5$ in our experiments.

5.2 Efficacy of Conditioning on both Modalities: In order to study the contribution of each modality on MELFUSION, we train three different variations of the model by selectively turning off visual conditioning and textual conditioning. We report the results in Tab. 4. We see significant improvement when conditioning on both modalities. This highlights how complementary semantic information from each modality can compose better music.

Text	Image	Extended MusicCaps			MeLBench		
		FAD ↓	KL ↓	IMSM ↑	FAD ↓	KL ↓	IMSM ↑
✓	✗	3.07	1.21	-	3.11	1.19	-
✗	✓	5.62	1.54	-	4.16	1.37	-
✓	✓	**1.12**	**0.89**	**0.76**	**1.05**	**0.72**	**0.83**

Table 4. Conditioning independently on each of the modalities leads to inferior music generation performance in this experiment.

5.3 Sensitivity Analysis: Tab. 5 reports the sensitivity analysis of changing the number of denoising steps T and the strength of classifier-free guidance during inference. Simi-

Varying Steps					Varying Guidance				
Guidance	Steps	FAD↓	KL↓	FD↓	Steps	Guidance	FAD↓	KL↓	FD↓
7	50	2.59	1.97	27.45	400	2	1.47	1.13	24.64
	200	1.35	1.12	25.22		7	**1.12**	**0.89**	**22.65**
	400	<u>1.12</u>	<u>0.89</u>	<u>22.65</u>		20	1.51	1.18	24.92
	600	1.09	0.88	22.57		30	1.63	1.29	25.38
	800	1.07	0.77	22.48		50	1.58	1.34	24.87

Table 5. Sensitivity analysis on the number of denoising steps T, and the strength of classifier-free guidance.

Text prompt length (in words)	Image	Objective metrics			Subjective metrics	
		FAD ↓	KL ↓	FD ↓	OVL ↑	REL ↑
$\geq 8 \leq 13$	✗	5.28	1.35	25.81	82.86	82.54
$\geq 14 \leq 19$	✗	3.13	1.20	23.11	85.25	85.16
≥ 20	✗	3.02	1.19	22.65	86.04	85.96
≤ 7	✓	**1.86**	**0.87**	**21.36**	**87.13**	**86.21**

Table 6. Performance of MeLFusion with varying verbosity of text prompts collected from MeLBench.

lar to the findings from Ghosal et al. [17], increasing T helps to generate more pleasing music. This can be attributed to enhanced refinement from more denoising. CFG strength of 7 gives the best result, and we use it in our experiments.

5.4 Verbose Text versus Image Conditioning: The visual synapse infuses fine-grained semantics from the image into text-to-music diffusion models. Another alternative to infuse such semantics would be to use verbose text descriptions. To study this, we remove the visual synapse from MeLFusion and train a music generation model conditioned only on text. Then, we vary the length of text prompts and report results in Tab. 6. Interestingly, we find that using image conditioning with a small text prompt outperforms using lengthier prompts. This underscores the utility of visual conditioning and the ability of visual synapses to modulate the LDM effectively.

5.5 Effect of visual-cues: We analyse the effect of using a different image and the same text prompt (please refer to the project page). When we change from the walkway image to the blue forest, the music becomes more calm and distant. We change the image to an abandoned amusement park, carnival orchestra, foggy seaside concert and forest at night. We observe a prevalence of eerie ambient sound echoing through the deserted park, occasional circus-inspired motifs, distant sounds of waves and foghorn-like effects and atmospheric strings imitating rustling leaves respectively.

5.6 Comparison with Text-to-Audio Methods: We include a comparison of MeLFusion with text-to-audio generation approaches in Tab. 7. We finetune their pre-trained checkpoints on our benchmark datasets for this experiment. The complementary information from both modalities allows our approach to outperform these methods too.

5.7 Effectiveness of IMSM: We conduct a user study with 64 participants to check whether the proposed IMSM metric is indeed capturing the relatedness between generated music and the conditioning image. We randomly choose 300

Method	MusicCaps			MeLBench		
	FAD ↓	KL ↓	FD ↓	FAD ↓	KL ↓	FD ↓
AudioLDM [32]	2.29	1.29	24.07	1.86	1.42	22.49
TANGO [17]	1.96	1.17	23.31	1.93	1.18	21.92
MeLFusion	**1.12**	**0.89**	**22.65**	**1.05**	**0.72**	**20.49**

Table 7. While comparing MeLFusion with state-of-the-art text-to-audio approaches, we see significant improvement in quality.

samples from Extended MusicCaps and MeLBench each. We compute the IMSM score, and also ask users for their image-music similarity on a scale of [0,1], for these samples. The average score from IMSM metric and the users for Extended MusicCaps and MeLBench are (0.76, 0.71) and (0.83, 0.85) respectively. The high correlation underscores the validity and usefulness of IMSM metric.

5.8 Using IMSM to Measure Purity of the Datasets: We compute the IMSM scores for all image-music pairs present in both Extended MusicCaps and MeLBench datasets and obtain a score of 0.91 and 0.93 respectively. The purpose of this is to quantitatively establish that the curated samples are perceptually in sync and are meaningful. The high values of the IMSM scores demonstrate that the curated image samples are highly perceptually similar and have ample association with the musical compositions.

6. Conclusion and Future Works

We explore the utility of infusing image semantics into a text-to-music diffusion model, enabling us to generate music, consistent with both visual and textual semantics in this work. To the best of our knowledge, ours is the first effort towards such a multi-conditioned music generation. We develop MeLFusion with a novel "visual synapse" to effectively infuse the image semantics into music generation, introduce a new dataset MeLBench, and propose a new evaluation metric. We conduct exhaustive experimental analysis on MeLBench and a modified version of MusicCaps [1] and compare MeLFusion against 7 text-to-music methods, and a modified baseline. The results suggest: 1) the extra information from the image conditioning significantly boosts music generation quality 2) our "visual synapse" is effective in modulating and infusing the required semantic information into the generative process.

MeLFusion can be an essential tool for a creative professional or a social-media content creator who needs to generate music that can go well with their multi-modal post (consider a user posting about their recent picnic – their photos can be the image conditioning while a short description of the trip can be the textual input to MeLFusion). Creating music with semantic lyrics that can go well with a video can be some interesting open-ended follow-up works.

References

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. 1, 2, 5, 6, 7, 8

[2] Shahaf Bassan, Yossi Adi, and Jeffrey S Rosenschein. Unsupervised symbolic music segmentation using ensemble temporal prediction errors. *arXiv preprint arXiv:2207.00760*, 2022. 2

[3] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 3, 6

[4] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023. 1, 2, 6, 7

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. 2023. 6, 7

[6] Paromita Das, Somsubhra Gupta, and Biswarup Neogi. Measurement of effect of music on human brain and consequent impact on attentiveness and concentration during reading. *Procedia computer science*, 172:1033–1038, 2020. 2

[7] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 2

[8] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[10] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6

[11] Evelina Fedorenko, Josh H McDermott, Sam Norman-Haignere, and Nancy Kanwisher. Sensitivity to musical structure in the human brain. *Journal of neurophysiology*, 108(12):3289–3300, 2012. 2

[12] Seth Forsgren and Hayk Martiros. Riffusion-stable diffusion for real-time music generation, 2022. *URL https://riffusion.com/about*, 6. 2, 6, 7

[13] Aviv Gabbay, Niv Cohen, and Yedid Hoshen. An image is worth more than a thousand words: Towards disentanglement in the wild. *Advances in Neural Information Processing Systems*, 34:9216–9228, 2021. 1

[14] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 758–775. Springer, 2020. 2

[15] Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo. Vampnet: Music generation via masked acoustic token modeling. *arXiv preprint arXiv:2307.04686*, 2023. 3

[16] Jort F Gemmeke, Daniel PW 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. IEEE, 2017. 5

[17] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023. 3, 6, 8

[18] Assal Habibi and Antonio Damasio. Music, feelings, and the human brain. *Psychomusicology: Music, Mind, and Brain*, 24(1):92, 2014. 2

[19] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 6

[20] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*, 2022. 2

[21] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023. 7

[22] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023. 2, 6

[23] Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022. 3

[24] Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech synthesis. *arXiv preprint arXiv:2205.07211*, 2022.

[25] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605, 2022. 3

[26] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020. 5

[27] 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. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 6

[28] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022. 6

[29] Max WY Lam, Jun Wang, Dan Su, and Dong Yu. Bddm: Bilateral denoising diffusion models for fast and high-quality speech synthesis. *arXiv preprint arXiv:2203.13508*, 2022. 3

[30] Max WY Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, et al. Efficient neural music generation. *arXiv preprint arXiv:2305.15719*, 2023. 1, 2, 6, 7

[31] Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. *arXiv preprint arXiv:2106.06406*, 2021. 3

[32] 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. *arXiv preprint arXiv:2301.12503*, 2023. 3, 5, 6, 8

[33] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. Contrastive audio-language learning for music. *arXiv preprint arXiv:2208.12208*, 2022. 2

[34] 2023 Mubert Inc. Mubert. URL https://mubert.com/. Mubert inc. mubert. url https://mubert.com/, 2023,. *Mubert Inc. Mubert. URL https://mubert.com/, 2023,*, 2023. 2, 6, 7

[35] Aashiq Muhamed, Liang Li, Xingjian Shi, Suri Yaddanapudi, Wayne Chi, Dylan Jackson, Rahul Suresh, Zachary C Lipton, and Alex J Smola. Symbolic music generation with transformer-gans. In *Proceedings of the AAAI conference on artificial intelligence*, pages 408–417, 2021. 2

[36] Sam V Norman-Haignere, Nancy Kanwisher, Josh H McDermott, and Bevil R Conway. Divergence in the functional organization of human and macaque auditory cortex revealed by fmri responses to harmonic tones. *Nature neuroscience*, 22(7):1057–1060, 2019. 2

[37] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[41] Catherine Schmidt-Jones. *The basic elements of music*. Connexions, 2012. 2

[42] Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023. 1, 2, 6, 7

[43] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[45] David Temperley and Trevor de Clercq. Statistical analysis of harmony and melody in rock music. *Journal of New Music Research*, 42(3):187–204, 2013. 2

[46] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[48] Junjie Wu, Junsong Zhang, Xiaojun Ding, Rui Li, and Changle Zhou. The effects of music on brain functional networks: a network analysis. *Neuroscience*, 250:49–59, 2013. 2

[49] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. 3

[50] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*, 2017. 2

[51] Adrien Ycart, Emmanouil Benetos, et al. A study on lstm networks for polyphonic music sequence modelling. ISMIR, 2017. 2

[52] Robert J Zatorre. Music and the brain. *Annals of the New York Academy of Sciences*, 999(1):4–14, 2003. 2