Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on AI, the Arts and Creativity

MusicMagus: Zero-Shot Text-to-Music Editing via Diffusion Models

Yixiao Zhang[1], **Yukara Ikemiya**[2], **Gus Xia**[3], **Naoki Murata**[2], **Marco A. Martínez-Ramírez**[2], **Wei-Hsiang Liao**[2], **Yuki Mitsufuji**[2], **Simon Dixon**[1]

[1]C4DM, Queen Mary University of London
[2]Sony AI
[3]Music X Lab, MBZUAI
{yixiao.zhang, s.e.dixon}@qmul.ac.uk, {yukara.ikemiya, naoki.murata, marco.martinez, weihsiang.liao, Yuhki.Mitsufuji}@sony.com, gus.xia@mbzuai.ac.ae

Abstract

Recent advances in text-to-music generation models have opened new avenues in musical creativity. However, music generation usually involves iterative refinements, and how to edit the generated music remains a significant challenge. This paper introduces a novel approach to the editing of music generated by such models, enabling the modification of specific attributes, such as genre, mood and instrument, while maintaining other aspects unchanged. Our method transforms text editing to *latent space manipulation* while adding an extra constraint to enforce consistency. It seamlessly integrates with existing pretrained text-to-music diffusion models without requiring additional training. Experimental results demonstrate superior performance over both zero-shot and certain supervised baselines in style and timbre transfer evaluations. We also showcase the practical applicability of our approach in real-world music editing scenarios.

1 Introduction

Recent advances in text-to-music generation have unlocked new possibilities in musical creativity [Zhang *et al.*, 2020; Zhao and Xia, 2021; Lu *et al.*, 2023; Min *et al.*, 2023; Agostinelli *et al.*, 2023; Schneider *et al.*, 2023; Copet *et al.*, 2023; Chen *et al.*, 2023]. However, a significant challenge persists in how to *edit* the generated results as music production usually involves iterative refinements. Building on this momentum, we regard 'text-to-music editing' as the process of using text queries to edit music, and we see two major types of operations: *inter-stem editing*, such as adding or removing instruments (e.g., "add a saxophone" or "remove the drums"), and *intra-stem editing*, which involves modifications within the same stem, such as adding effects or changing instruments (e.g., "add reverb to this stem" or "transfer the timbre of the specified notes"). In this context, a "stem" refers to an individual track or component within a music piece, such as a specific instrument or vocal part. The primary focus of this paper is on the latter, *intra-stem editing*.

One of the fundamental challenges of text-to-music editing is the difficulty of accommodating flexible text operations in both dataset construction and model training. This is not only a matter of data pair scarcity, but also the complexity inherent in the vast array of possible text-based edits that can be applied to music. Existing research [Wang *et al.*, 2023; Han *et al.*, 2023; Hussain *et al.*, 2023] has primarily focused on manually constructing datasets. However, these models are constrained to a few predefined operations, which undermines their effectiveness in text-to-music editing that requires flexibility and variety. This highlights the need for a new approach that moves away from traditional supervised learning reliant on specific data pairs and towards a more adaptable, unsupervised, or zero-shot approach.

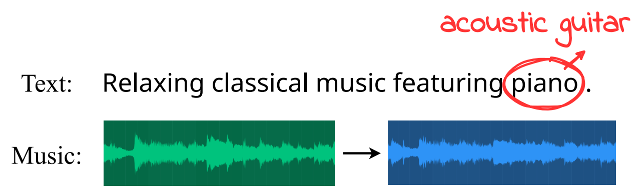

Figure 1: Text-to-music editing with MusicMagus. The edit from "piano" to "acoustic guitar" in the text prompt directly alters the corresponding musical attribute, while leaving others unchanged.

In this work, we introduce MusicMagus, which focuses on text-based *intra-stem* music editing. Leveraging the inherent capabilities of pre-trained diffusion models, MusicMagus is able to perform zero-shot editing without requiring additional training pairs. As illustrated in Figure 1, we utilise word swapping to direct the editing process. This operation is implemented as a manipulation within the diffusion model's semantic space. Recognizing the sensitivity of the diffusion process, where minor alterations can propagate significant changes, we employ an additional constraint to ensure that the resultant music maintains the structural integrity and stylistic coherence of the original music.

Although we mainly focus on the editing of music clips generated from diffusion models, we also discuss how to edit real-world music audio by the Denoising Diffusion Implicit Model (DDIM) inversion [Song *et al.*, 2021] technique.

In summary, our main contributions are as follows:

1. We propose a flexible and user-friendly text-to-music editing method using word swapping.
2. We contribute MusicMagus, a system capable of zero-shot music editing on diverse tasks without any depen-

dence on *paired* training data.

3. Comparative experiments validate that MusicMagus outperforms existing zero-shot methods and some supervised approaches in critical tasks such as style and timbre transformation. [1] [2]

2 Related Work

2.1 Text-to-Music Generation

Text-to-music generation models in the audio domain broadly fall into two categories: autoregressive (AR) models, primarily language model (LM) based, operating on discrete audio representations, and diffusion-based models working with continuous latent representations [Ho *et al.*, 2020]. AR models like MusicLM [Agostinelli *et al.*, 2023], MeLoDy [Lam *et al.*, 2023] and MusicGen [Copet *et al.*, 2023] excel in creating longer and higher-quality audio sequences but are limited by higher inference times, which can be challenging for interactive applications such as music generation and editing. Conversely, diffusion models, including Möusai [Schneider *et al.*, 2023], AudioLDM 2 [Liu *et al.*, 2023a], Jen-1 [Li *et al.*, 2023], and Tango [Ghosal *et al.*, 2023], offer advantages in parallel decoding but require numerous diffusion steps for high-quality output, and often struggle with generating longer audio sequences. Recently, MagNet [Jeong *et al.*, 2024] offers a novel, hybrid approach to music generation. Combining the best of AR and diffusion models, it starts with autoregressive sequence generation and finishes with parallel decoding. This method effectively balances quality and efficiency in music production.

There is also a growing emphasis on controllability in text-to-music generation models. Coco-mulla [Lin *et al.*, 2023] utilizes drum patterns and chord progressions, enhancing MusicGen's conditional music generation capabilities. Similarly, Music ControlNet [Wu *et al.*, 2023a] and DITTO [Novack *et al.*, 2024] apply multiple controls over a pretrained diffusion model for tailored music creation. Mustango [Melechovsky *et al.*, 2023] integrates metadata control within the diffusion-based TANGO [Ghosal *et al.*, 2023] framework; whereas Jen-1 Composer [Yao *et al.*, 2023] and StemGen [Parker *et al.*, 2023] generate new stems conditioned on existing stems, thus capitalizing on pre-existing musical elements for generation.

2.2 Text-to-Music Editing

Text-to-music editing encompasses two distinct types of operations: inter-stem and intra-stem editing. Inter-stem editing refers to operations conducted on one stem (such as adding or removing stems) that are conditioned on another stem, whereas intra-stem editing involves modifications within the stem itself, such as adjusting the instrument, genre, or mood.

Compared to text-based image editing [Hertz *et al.*, 2022; Parmar *et al.*, 2023; Hu *et al.*, 2024], research on text-to-music editing is relatively limited. Models like InstructME [Han *et al.*, 2023] and M^{2}UGen [Hussain *et al.*, 2023] demonstrate capabilities in both inter-stem and intra-stem editing, allowing for structural changes and detailed modifications within stems, but they often require extra training and specific data. Loop Copilot [Zhang *et al.*, 2023], an AI agent, employs a combination of existing models to facilitate compositional editing, yet it does so without altering the fundamental architecture or interface of the original models. Furthermore, the tag-conditioned timbre [Wu *et al.*, 2023c] and style [Dai *et al.*, 2018] transfer task can be seen as a precursor to this task, however labels are not as flexible as text.

In contrast, our model introduces a novel intra-stem editing approach. While it also operates without additional training, our approach distinctively utilizes the latent capacities of pre-trained diffusion-based models. This method enables efficient text-to-music editing, leveraging existing model structures without necessitating their combination or alteration.

Concurrently, a recent work [Manor and Michaeli, 2024] also attempts the similar task of text-based audio editing with diffusion models. We encourage the readers to refer to the work for more details.

[1]Work done during Yixiao's internship at Sony AI.

[2]Code and demo: https://bit.ly/musicmagus-demo.

3 Background

MusicMagus utilizes a pretrained diffusion model [Ho *et al.*, 2020] for text-to-music editing, eliminating the need for additional training. Specifically, we use a pretrained AudioLDM 2 model [Liu *et al.*, 2023a] as the backbone model, since it is the most accessible text-to-music diffusion model with publicly available weights. AudioLDM 2 employs a variational autoencoder (VAE) [Kingma and Welling, 2013] to compress a music audio spectrogram into a latent low-dimensional space. It then trains a latent diffusion model (LDM) on this latent space to generate new samples from Gaussian noise conditioned on text inputs. During generation, the LDM takes a condition y, generates a latent variable z_0, and uses the VAE decoder to produce the music spectrogram x, which can be then converted into a waveform using an external vocoder, such as HiFi-GAN [Su *et al.*, 2020].

During training, the LDM performs a forward diffusion process, which is defined as a Markov chain that gradually adds Gaussian noise to the latent representation of the data over T steps. This process can be represented as:

$$z_t = \sqrt{\alpha_t} z_{t-1} + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $t = 1, 2, \ldots, T$, z_t is the latent variable at step t, α_t is a variance schedule for the noise, and ϵ is a noise vector drawn from a standard Gaussian distribution. The process starts with z_0 being the initial latent representation of the data and ends with z_t being a sample from the Gaussian noise distribution.

The inference process in LDMs is the reverse of the forward process. It starts with a sample from the Gaussian noise distribution z_t and aims to recover the original data representation z_0. This is achieved by a series of denoising steps that can be described by the following formulation:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t)\right) + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\epsilon_\theta(z_t, t)$ is a neural network that predicts the noise added at step t, and $sigma_t$ represents the standard deviation of the noise added. The network ϵ_θ

is trained to minimize the difference between the predicted noise and the actual noise added during the forward process.

For simplicity, we denote the formula (2) as:

$$z_{t-1} = \text{Denoise}(z_t, \epsilon_\theta, t). \tag{3}$$

To decrease computational demands, denoising diffusion implicit models (DDIM) [Song *et al.*, 2021] introduced a modified approach which enables significantly fewer sampling steps (e.g., between 50 and 100, whereas DDPMs usually have 1000 steps) during inference, while having a negligible effect on the quality of the generated output.

4 Method

To illustrate our idea, we refer to the example in Figure 1. Initially, a music clip, denoted as x, is generated from the text prompt "Relaxing classical music featuring piano", which we refer to as y. The next step involves altering this text prompt by substituting "piano" with "acoustic guitar", thereby creating a new prompt y'. Our aim is to produce a revised music piece x', where only the specified attribute is changed, while maintaining all other aspects.

The explanation of our idea is twofold. In Section 4.1, we detail the method for altering the text prompt in the semantic domain. Subsequently, in Section 4.2, we discuss our approach to enforce suitable constraints over the cross-attention map during diffusion to preserve the integrity of the remaining elements of the music.

4.1 Finding Editing Direction

In this section, we introduce a strategy to calculate a difference (Δ) vector in the latent space to guide the editing direction. This method is chosen over direct word swapping as it better preserves semantic coherence and contextual relevance, especially in cases of varying phrase lengths and complex content alterations. We will further explain it in Section 4.2; besides, previous research finds that similar operations can facilitate a more robust edit, especially when the keywords subject to modification are sparsely represented in the training dataset [Parmar *et al.*, 2023].

We first introduce the text embedding method in AudioLDM 2. AudioLDM 2 uses a two-branch text encoder to embed the text prompt y to two embeddings: $E = \{E_{\text{T5}}, E_{\text{GPT}}\}$, where E_{T5} encodes the sentence-level representation, and E_{GPT} captures the more fine-grained semantic information inside y.

First, the FLAN-T5 [Chung *et al.*, 2022] encoder, utilizing a T5 model [Raffel *et al.*, 2020], encodes y into a feature vector $E_{\text{T5}} \in \mathbb{R}^{L\times 1024}$, where L represents the sentence length. In parallel, the CLAP [Wu *et al.*, 2023b] text encoder leverages a RoBERTa [Liu *et al.*, 2019] model to transform y into a flattened vector $E_{\text{CLAP}} \in \mathbb{R}^{1\times 512}$:

$$\begin{cases} E_{\text{T5}} = \text{T5}(y), \\ E_{\text{CLAP}} = \text{CLAP}(y). \end{cases} \tag{4}$$

Then, E_{T5} and E_{CLAP} are linearly projected to $P \in \mathbb{R}^{768}$. A GPT-2 model, pre-trained on an AudioMAE [Huang *et al.*, 2022], is then employed to auto-regressively generate 8 new tokens $E_{\text{GPT}} \in \mathbb{R}^{8\times 768}$:

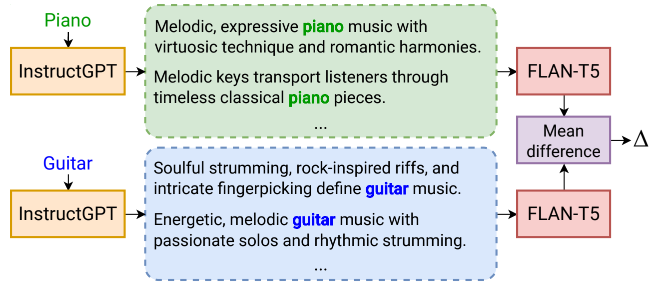

Figure 2: The pipeline of finding the editing direction Δ. We first use InstructGPT to generate a large number of captions and then calculate the mean difference between the two embedding sets.

$$E_{\text{GPT}} = \text{GPT-2}(\text{Proj}(E_{\text{T5}}, E_{\text{CLAP}})). \tag{5}$$

The LDM takes both E_{T5} and E_{GPT} as input in the diffusion process:

$$\epsilon_\theta = \epsilon_\theta(z_t, E, t), \tag{6}$$
$$z_{t-1} = \text{Denoise}(z_t, \epsilon_\theta, E, t). \tag{7}$$

Similarly, the new prompt y' can be encoded to $E' = \{E'_{\text{T5}}, E'_{\text{GPT}}\}$. Our goal is to find $E^{\text{edit}} = \{E^{\text{edit}}_{\text{T5}}, E^{\text{edit}}_{\text{GPT}}\}$.

We use the following method to find the editing vector Δ, as shown in Figure 2:

1. We first generate a multitude of music-related captions using a pretrained InstructGPT model [Ouyang *et al.*, 2022]. These captions are designed to contain the original and new keywords.
2. Subsequently, we input these two sets of captions into the FLAN-T5 encoder and compute the mean embeddings for each set of encoded vectors.
3. The final step is calculating the difference between these two mean embeddings, which is then employed as the vector for the editing direction Δ.

We employ different strategies to edit E_{T5} and E_{GPT}. For E_{T5}, the edited embedding is:

$$E^{\text{edit}}_{\text{T5}} = E_{\text{T5}} + \Delta. \tag{8}$$

The aforementioned editing method encounters challenges when applying Δ to E_{GPT}. The core issue is that E_{GPT} is obtained through the GPT-2 model, where the addition of a Δ to the embedding may not constitute a semantically valid operation. Consequently, in practical applications, we resort to using $E^{\text{edit}}_{\text{GPT}} = E'_{\text{GPT}}$, which is derived directly from encoding the new prompt.

Finally, we have the edited embeddings:

$$E^{\text{edit}} = \{E_{\text{T5}} + \Delta, E'_{\text{GPT}}\}. \tag{9}$$

4.2 Adding Constraints over Cross-Attention

Diffusion models exhibit inherent randomness in their generation output. By setting a fixed random seed and using the same text prompts, we can reproduce the same musical output. However, even minor variations in the text prompt can

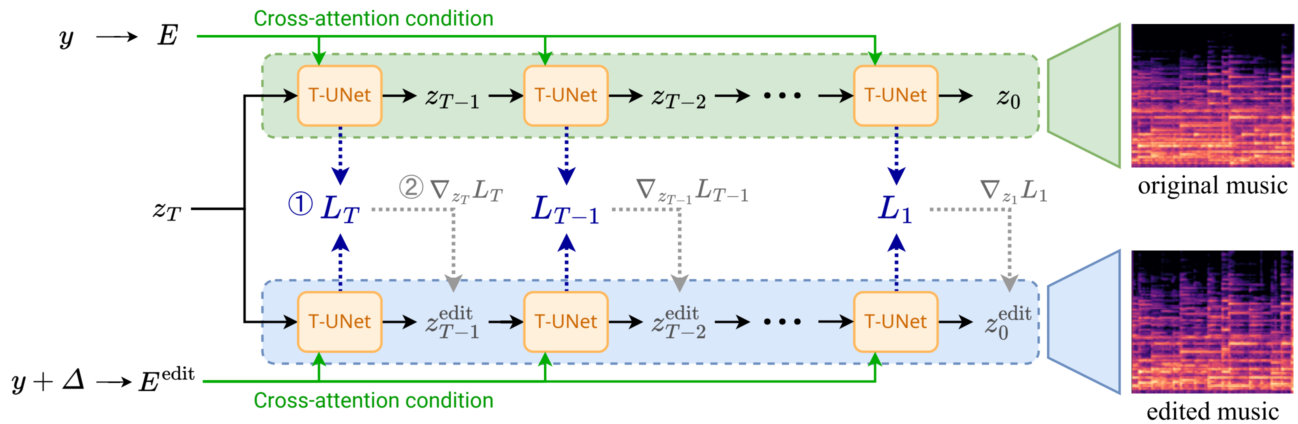

Figure 3: The workflow of the MusicMagus model. To constrain the diffusion model at timestep t, we need to: (1) calculate the L2 loss L_t between the cross-attention map M^{edit}_t and M^{origin}_t; (2) compute the gradient of L_t with respect to z_t, and then perform a single-step optimization to update $\epsilon^{\text{edit}}_\theta$ of the diffusion model.

result in significantly different music clips. Previous studies have demonstrated that imposing external constraints on the cross-attention map between the text condition and the diffusion latent space enhances the consistency of the music generation, particularly for the remaining attributes that need to remain unchanged [Hertz *et al.*, 2022; Parmar *et al.*, 2023; Tumanyan *et al.*, 2023]. Building on this concept, we introduce a method designed to constrain the text-to-music diffusion model specifically for editing purposes.

To begin, we examine the acquisition of the cross-attention map. During the denoising process at timestep t, the model computes the cross-attention score between the encoded text $\{E_{\text{T5}}, E_{\text{GPT}}\}$ and the intermediate features of LDM ϵ_θ:

$$\begin{aligned} &\text{Attention}(Q, K, V) = M \cdot V, \\ &\quad \text{where } M = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right). \end{aligned} \tag{10}$$

In this context, $Q = W_Q\phi(z_t),\ K = W_k E,\ V = W_v E$ are defined, where $W = \{W_Q, W_K, W_V\}$ represents projection layers, and $E = \{E_{\text{T5}}, E_{\text{GPT}}\}$ are the text embeddings. AudioLDM 2 proposes the T-UNet architecture, which is distinct from the UNet architecture, to extract intermediate spatial features $\phi(x_t)$. T-UNet incorporates a transformer block after each encoder and decoder block's convolution operation, and the cross-attention occurs in the transformer block's final layer. The term d denotes the dimension of the projected keys and queries.

As illustrated in Figure 3, to apply the editing, we first reconstruct the music x with the original text embeddings E. We record the cross-attention maps for each timestep $t \in [1, T]$:

$$M^{\text{origin}} = \{M^{\text{origin}}_1, ..., M^{\text{origin}}_T\}. \tag{11}$$

Then we use the edited text embeddings E^{edit} to generate an edited music clip. Similarly, at timestep t, we have a cross-attention map M^{edit}_t.

At each timestep t, we apply the constraint by calculating the L_2 loss between M^{origin}_t and M^{edit}_t:

$$L_t = \|M^{\text{edit}}_t - M^{\text{origin}}_t\|_2. \tag{12}$$

We then compute the gradient $\nabla_{z_t} L_t$ and perform a single-step optimization with the step length α:

$$\epsilon^{\text{edit}}_\theta = \epsilon_\theta(z_t - \alpha\nabla_{z_t} L_t, E^{\text{edit}}, t). \tag{13}$$

Subsequently, we execute the t-step denoising process using the updated ϵ^{edit}:

$$z_{t-1} = \text{Denoise}(z_t, \epsilon^{\text{edit}}_\theta, E^{\text{edit}}, t). \tag{14}$$

This optimization is applied at every step until the denoising process is completed. Experimental results of the ablation studies validate that this constraint significantly enhances structural consistency during denoising.

To effectively utilize the cross-attention constraint, employing Δ for editing is essential. This method is crucial, especially when dealing with cases that involve substituting text of varying lengths, exemplified by replacing a shorter expression with a longer one (such as "piano" $\rightarrow$ "acoustic guitar"). Utilizing Δ maintains the uniformity of embedding lengths during the editing process. In contrast, techniques like word swapping can alter these lengths, leading to discrepancies between M^{edit} and M^{origin}, and consequently, errors in calculating L_t. Furthermore, Δ facilitates the insertion of words at different sentence positions without disrupting the position-related cross-attention maps, ensuring the attention mechanism remains focused on the correct semantic context.

5 Experiments

In the domain of text-to-music editing, comprehensive model evaluation is inherently challenging due to the countless number of possible editing schemes. To address this, we focus on two key aspects: timbre transfer and style transfer, and compare our model's performance against established baselines in these areas. This comparison is conducted through both objective and subjective testing methodologies.

5.1 Baselines

We benchmark our model against three distinct models in the field: AudioLDM 2 [Liu *et al.*, 2023a], Transplayer [Wu *et al.*, 2023c], and MusicGen [Copet *et al.*, 2023]. While our approach utilizes AudioLDM 2 as its backbone, AudioLDM 2 independently offers methods for both timbre and style transfer tasks, making it a relevant baseline.

AudioLDM 2. AudioLDM 2 is a diffusion-based model supporting unified speech, audio, and music generation at 16kHz. It follows the idea of AudioLDM and individually proposes a method for general audio style transfer. This is achieved through the interpolation of audio latents and subsequent denoising with a new prompt.

Transplayer. This state-of-the-art, diffusion-based model trained on POP909 [Wang *et al.*, 2020] and MAESTRO [Hawthorne *et al.*, 2019] dataset, specialising in timbre transfer at 16kHz. Unlike typical timbre transfer models that require training for each instrument pair, Transplayer is trained on multiple pairs, enabling versatile many-to-many timbre transfers.

MusicGen. A leading text-to-music generation model, MusicGen is a supervised model trained on a dataset of over 20,000 high-quality music pieces, generating 32kHz music. It uniquely allows for the inclusion of an extra melody condition, facilitating the style transfer task within the text-to-music generation process.

5.2 Metrics

We employ different metrics for subjective and objective experiments. For the subjective evaluation, we incorporate the following metrics, where OVL and REL are following [Kreuk *et al.*, 2023]:

Overall Quality (OVL). This metric is used to assess the overall music quality, encompassing aspects like sound clarity and musicality. It primarily evaluates whether the editing process enhances or diminishes the quality of the original music audio. The scoring for this metric ranges from 0 to 100.

Relevance (REL). REL measures the perceived semantic closeness between the edited music and the new text prompt. It is a subjective score, also ranging from 0 to 100.

Structural Consistency (CON). We define a new metric CON to evaluate the consistency of the pitch contour and structural aspects in the subjective test. Similar to the others, its scoring range is from 0 to 100.

The objective experiments utilize the following metrics:

CLAP Similarity (CLAP). [Wu *et al.*, 2023b] This metric assesses the semantic relevance between the edited music and the new text prompt. It utilizes a pretrained CLAP model, where a higher score indicates greater semantic similarity between the music and text, with scores ranging from 0 to 1. We implement it with the MuLaB library [Manco *et al.*, 2023].

Chromagram Similarity (Chroma). We use this new metric to gauge the preservation of pitch contours and rhythm patterns in the music. It involves computing the cosine similarity between the chromagrams of the original and edited music. A higher score suggests better retention of the structure and pitch contour, with values also ranging from 0 to 1. We implement this metric with the librosa library [McFee *et al.*, 2015].

5.3 Data Preparation

Objective Experiments

For the timbre transfer task, we conducted a random selection of 60 music audio samples generated by AudioLDM 2, covering three specific word swapping pairs: (piano $\rightarrow$ organ), (viola $\rightarrow$ piano), and (piano $\rightarrow$ acoustic guitar). The primary rationale behind choosing these pairs is the limited range of instrument pairs supported by the Transplayer model. Given that the quality of music generated by AudioLDM 2 can vary, we implemented a quality-based filtering process. This entailed excluding any music samples that fell below a predefined quality threshold, continuing this selection process until the requisite number of suitable samples was attained.

Building upon the methodology established for timbre transfer, we applied a similar approach to the music style transfer task. Our selection encompassed a diverse range of style conversions, including (jazz $\rightarrow$ classical), (country $\rightarrow$ metal), (jazz $\rightarrow$ metal), and (jazz $\rightarrow$ rock). For each of these style pairs, we employed a random selection process, ultimately curating a dataset comprising 50 samples in total.

We use a template to synthesize the text prompt: "A $\{mood\}$ $\{genre\}$ music with $\{timbre\}$ performance.", where mood is randomly chosen from a fixed set of {"upbeat", "relaxing", "peaceful"}.

Subjective Experiments

For the subjective test, we randomly selected a subset of data points from the objective test dataset. Specifically, 8 data points were chosen for the timbre transfer task and 5 data points for the style transfer task. Each data point included results from both the baseline models and our ablation studies. The results are shown in Tables 1 and 2.

5.4 Experimental Setup

We choose the *AudioLDM2-base* model [3] as our backbone model. During inference, we configure the DDIM steps to 100, and generate 5-second audio clips at a sampling rate of 16kHz. A uniform gradient step length ($\alpha = 0.04$) is applied for both timbre transfer and style transfer tasks. All inference is performed on a single NVIDIA A100 GPU.

For the Transplayer model, we utilize the official pretrained checkpoint [4] without any modifications to its weights or code. As for MusicGen, we opt for the *MusicGen-melody* checkpoint [5], which has 1.5B parameters. To maintain consistency, all generated samples from these models are subsequently downsampled to 16kHz resolution.

5.5 Results

Subjective Experiments

We conducted a subjective listening test for both the timbre transfer and style transfer tasks. This test involved disseminating an online survey within the Music Information Retrieval (MIR) community and our broader research network, which resulted in the collection of 26 complete responses. The gender distribution of the participants was 19

[3]https://huggingface.co/cvssp/audioldm2

[4]https://github.com/Irislucent/TransPlayer

[5]https://huggingface.co/facebook/musicgen-melody

Model name	Type	REL	OVL	CON	Avg.
AudioLDM 2	Zero-shot	15.7	49.9	**80.6**	48.7
Transplayer	Supervised	28.3	28.9	34.6	30.6
Ours *w/o L2 & Δ*	Zero-shot	78.0	61.6	50.4	63.3
Ours *w/o L2*	Zero-shot	**78.8**	**62.4**	51.3	64.2
Ours (final)	Zero-shot	76.2	62.1	66.6	**68.3**

Table 1: The subjective evaluation results on the timbre transfer task.

Model name	Type	REL	OVL	CON	Avg.
AudioLDM 2	Zero-shot	19.8	53.2	**84.2**	52.4
MusicGen	Supervised	63.3	**66.0**	48.2	59.1
Ours *w/o L2 & Δ*	Zero-shot	69.2	56.9	58.9	61.7
Ours *w/o L2*	Zero-shot	**71.3**	53.8	55.0	60.0
Ours (final)	Zero-shot	65.7	57.8	65.6	**63.1**

Table 2: The subjective evaluation results on the style transfer task.

Model name	Type	CLAP	Chroma	Avg.
AudioLDM 2	Zero-shot	0.16	0.72	0.44
Transplayer	Supervised	0.18	0.56	0.37
Ours *w/o L2 & Δ*	Zero-shot	0.33	0.68	0.51
Ours *w/o L2*	Zero-shot	**0.34**	0.69	0.52
Ours (final)	Zero-shot	0.33	**0.76**	**0.55**

Table 3: The objective evaluation results on the timbre transfer task.

Model name	Type	CLAP	Chroma	Avg.
AudioLDM 2	Zero-shot	0.18	**0.80**	**0.49**
MusicGen	Supervised	**0.24**	0.66	0.45
Ours *w/o L2 & Δ*	Zero-shot	0.22	0.65	0.44
Ours *w/o L2*	Zero-shot	0.22	0.67	0.45
Ours (final)	Zero-shot	0.21	0.77	**0.49**

Table 4: The objective evaluation results on the style transfer task.

males (76%) and 6 females (24%). Regarding musical experience, 5 participants (19.23%) had less than 1 year of experience, 5 (19.23%) had between 1 and 5 years, and the majority, 16 participants (61.54%), had more than 5 years of experience. This subjective test was approved by the ethics committee of both Sony AI and Queen Mary University of London (QMERC20.565.DSEECS23.129).

The data presented in Table 1 reveals that our proposed model exhibits superior performance in the timbre transfer task when compared to two baseline models. Specifically, AudioLDM 2 demonstrates a notable limitation in transferring to novel semantics, resulting in edited samples that closely resemble the original ones. This is evident from its low Relevance (REL) score and high Consistency (CON) score. Contrary to expectations, the performance of Transplayer is consistently inferior, suggesting that its generalization capability may be inadequate for complex tasks such as many-to-many instrument timbre transfer in practical applications. Our model is the best on the average of altering semantic content and maintaining structural integrity.

Insights gleaned from our ablation study further elucidate these findings. The inclusion of the additional constraint significantly enhances performance in terms of Structure Consistency (CON), highlighting its role in bolstering structural coherence. However, the subjective experiments indicate no marked difference in Relevance (REL) scores between the methods. This observation aligns with expectations, since the primary objective of Δ usage is to ensure the consistency of the cross-attention maps, particularly during complex editing operations or in scenarios involving underrepresented words demonstrated in Section 4.1, which may not be fully reflected by the current subjective test settings.

We also evaluated our model's performance in the style transfer task, as detailed in Table 2. Similar to the previous findings, our model demonstrates superior performance over the baseline models in this task as well.

AudioLDM 2 exhibits notable limitations in style transfer, with its performance being generally unstable; MusicGen, despite its downsampled audio quality from 32KHz to 16kHz, retains a high level of audio quality, as indicated by its high Overall Quality (OVL) score. However, MusicGen struggles with precisely preserving the original melody in the style transfer process, particularly in maintaining polyphonic melodies, which introduces some instability in its outputs.

In contrast, our method not only changes the semantics but also keeps that the overall quality is not diminished, resulting in the best average score; it also maintains the structural integrity and pitch consistency, which are critical in music style transfer.

Objective Experiments

We compare the performance of our model and the zero-shot and supervised baselines. The results for the timbre transfer and style transfer tasks are shown in Tables 3 and 4.

In the timbre transfer task (Table 3), our model demonstrated enhanced performance in semantic transfer. The incorporation of a constraint on the cross-attention mechanism largely improved pitch and rhythm accuracy, reinforcing the insights obtained from the subjective experiments. These results substantiate the efficacy of our model in maintaining semantic integrity while facilitating timbre transfer results.

Table 4 presents the findings for the style transfer task. Here, our model outperformed the baselines in terms of structural and pitch consistency. However, in terms of semantic transfer, the differences between our model and the baselines were less pronounced. This suggests that while our model excels in maintaining the structural and pitch elements during style transfer, the semantic changes are comparable to those achieved by the baseline models.

6 Discussion

6.1 Real Music Audio Editing

MusicMagus offers capabilities for editing real-world music audio, although it is noted that the performance may not match the editing of synthesized music audio generated from

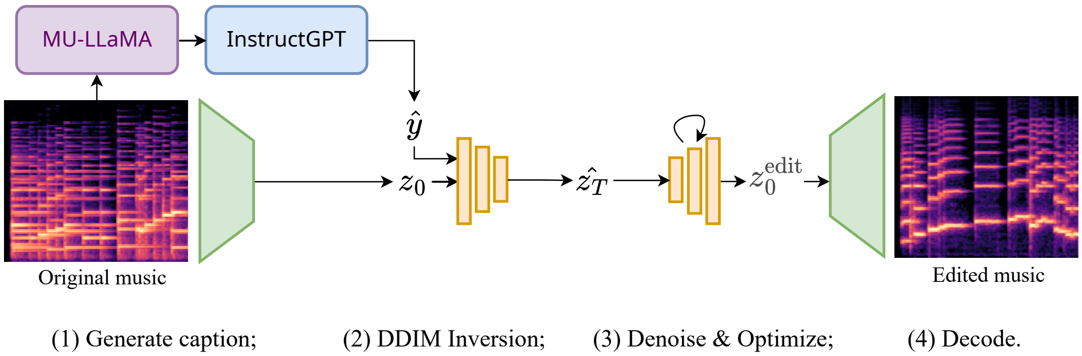

Figure 4: The diagram of the real music audio editing pipeline using MusicMagus with DDIM inversion and diffusion model editing.

diffusion models. We begin with the DDIM inversion to estimate the latent representation $\hat{z_T}$ of a given real music audio x. This step is crucial to facilitate editing with the diffusion model, as depicted in Figure 4.

The inversion requires a corresponding text prompt $\hat{y}$, which is initially generated by a pretrained music captioning model, MU-LLaMA [Liu *et al.*, 2023b]. Due to the discrepancy between the text prompt distributions of AudioLDM 2 and MU-LLaMA, the InstructGPT model is employed to refine the generated captions, aligning them more closely with AudioLDM 2's distribution. This refinement includes condensing the caption into a single, concise sentence and emphasizing essential characteristics such as the key instruments, mood, and genre.

DDIM inversion, while effective, is not a perfect reconstruction method. It faces a trade-off between the editability of the estimated latent $\hat{z_T}$ and its reconstruction fidelity [Hertz *et al.*, 2022]. A balance is sought by selecting an intermediate value for classifier-free guidance, set to 1. Additionally, the diffusion latent is typically modeled as Gaussian noise. To mitigate auto-correlation that may arise during inversion, we adopt a strategy from Parmar et al. [Parmar *et al.*, 2023], introducing autocorrelation regularization to diminish its impact, thereby enhancing the estimation of $\hat{z_T}$.

Subsequent to obtaining the estimated latent $\hat{z_T}$, the caption $\hat{y}$ is edited, and the MusicMagus editing algorithm is applied within the diffusion model framework to produce the edited music audio [6].

6.2 Limitations

The current implementation of MusicMagus, while effective, is built upon the AudioLDM 2 model, which is not without its constraints. One significant limitation is the model's challenge in generating multi-instrument music when such complexity is specified. This inherently restricts the scope of creative expression and diversity that the model can offer. The performance of AudioLDM 2 was not enhanced in our approach, which is an aspect we aim to address moving forward.

Moreover, our zero-shot method exhibits instability, as evidenced by a notable number of failure cases. These failures are often due to unsuccessful application of the delta and word-swapping techniques, highlighting an area ripe for improvement. Currently, the scope of alterations we can apply to the music is somewhat modest; our system struggles to introduce substantial changes, such as adding or removing an instrument, adding sound effects, etc., without compromising the overall structure and quality of the audio.

Another factor that confines our system is the inherent limitations of the base model itself. For instance, the diffusion process struggles with generating very long sequences, which in turn limits the practical applications of our model. Addressing this limitation could potentially open up new domains where longer sequence generation is essential.

Lastly, the audio quality, currently capped by the 16kHz sampling rate, is another significant limitation, often resulting in artifacts that can detract from the listener's experience. Enhancing the audio fidelity is an important step that will bring us closer to a model that can produce professional-grade audio, which is crucial for both consumer applications and artistic endeavors. The pursuit of higher audio quality and the reduction of artifacts are critical goals for our future work.

7 Conclusion

In conclusion, our research contributes a novel text-to-music editing framework that effectively manipulates selected musical aspects, such as timbre and style, without altering the remaining parts. Our method distinguishes itself by its compatibility with current diffusion models and its operational simplicity, not necessitating further training protocols. The empirical evidence from our studies confirms that our method advances the state-of-the-art, delivering enhanced performance in style and timbre transfer.

Although we have identified areas for improvement, such as the model's ability to handle complex multi-instrument compositions and the stability of zero-shot methods, these challenges provide a clear trajectory for our ongoing research. By incrementally refining the underlying model and expanding the editing capabilities, we aim to push the boundaries of automated music generation and editing further. The ultimate goal is to refine the underlying model, enabling the generation and editing of high-fidelity, nuanced, and diverse musical compositions with simple and intuitive human input while maximizing creative expressiveness.

[6]We provide listening samples at the demo page.

Ethical Statement

The subjective tests are approved by the ethics committee of both Sony AI and Queen Mary University of London (QMERC20.565.DSEECS23.129).

Acknowledgments

We want to thank Dr. Woosung Choi, Zineb Lahrichi, Liwei Lin and Liming Kuang for their contribution to this work. Yixiao Zhang is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by the China Scholarship Council, Queen Mary University of London and Apple Inc.

References

[Agostinelli *et al.*, 2023] 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.

[Chen *et al.*, 2023] 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. *arXiv preprint arXiv:2308.01546*, 2023.

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

[Copet *et al.*, 2023] 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.

[Dai *et al.*, 2018] Shuqi Dai, Zheng Zhang, and Gus G Xia. Music style transfer: A position paper. *arXiv preprint arXiv:1803.06841*, 2018.

[Ghosal *et al.*, 2023] 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.

[Han *et al.*, 2023] Bing Han, Junyu Dai, Xuchen Song, Weituo Hao, Xinyan He, Dong Guo, Jitong Chen, Yuxuan Wang, and Yanmin Qian. InstructME: An instruction guided music edit and remix framework with latent diffusion models. *arXiv preprint arXiv:2308.14360*, 2023.

[Hawthorne *et al.*, 2019] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019.

[Hertz *et al.*, 2022] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2024] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-Imagen: Image generation with multi-modal instruction. *arXiv preprint arXiv:2401.01952*, 2024.

[Huang *et al.*, 2022] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022.

[Hussain *et al.*, 2023] Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. M^{2}UGen: Multi-modal music understanding and generation with the power of large language models. *arXiv preprint arXiv:2311.11255*, 2023.

[Jeong *et al.*, 2024] Myeonghun Jeong, Minchan Kim, Joun Yeop Lee, and Nam Soo Kim. Efficient parallel audio generation using group masked language modeling. *arXiv preprint arXiv:2401.01099*, 2024.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kreuk *et al.*, 2023] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. In *The Eleventh International Conference on Learning Representations*, 2023.

[Lam *et al.*, 2023] 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.

[Li *et al.*, 2023] Peike Li, Boyu Chen, Yao Yao, Yikai Wang, Allen Wang, and Alex Wang. Jen-1: Text-guided universal music generation with omnidirectional diffusion models. *arXiv preprint arXiv:2308.04729*, 2023.

[Lin *et al.*, 2023] Liwei Lin, Gus Xia, Junyan Jiang, and Yixiao Zhang. Content-based controls for music large language modeling. *arXiv preprint arXiv:2310.17162*, 2023.

[Liu *et al.*, 2019] 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. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2023a] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023.

[Liu *et al.*, 2023b] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music understanding llama: Advancing text-to-music generation with question answering and captioning. *arXiv preprint arXiv:2308.11276*, 2023.

[Lu *et al.*, 2023] Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. MuseCoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*, 2023.

[Manco *et al.*, 2023] Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bodganov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, et al. The Song Describer Dataset: A corpus of audio captions for music-and-language evaluation. *arXiv preprint arXiv:2311.10057*, 2023.

[Manor and Michaeli, 2024] Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using DDPM inversion. *CoRR*, abs/2402.10009, 2024.

[McFee *et al.*, 2015] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, volume 8, pages 18–25, 2015.

[Melechovsky *et al.*, 2023] Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*, 2023.

[Min *et al.*, 2023] Lejun Min, Junyan Jiang, Gus Xia, and Jingwei Zhao. Polyffusion: A diffusion model for polyphonic score generation with internal and external controls. In *ISMIR 2023 Hybrid Conference*, 2023.

[Novack *et al.*, 2024] Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. Ditto: Diffusion inference-time t-optimization for music generation, 2024.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[Parker *et al.*, 2023] Julian D Parker, Janne Spijkervet, Katerina Kosta, Furkan Yesiler, Boris Kuznetsov, Ju-Chiang Wang, Matt Avent, Jitong Chen, and Duc Le. Stemgen: A music generation model that listens. *arXiv preprint arXiv:2312.08723*, 2023.

[Parmar *et al.*, 2023] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[Schneider *et al.*, 2023] Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. Möusai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023.

[Song *et al.*, 2021] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[Su *et al.*, 2020] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. *arXiv preprint arXiv:2006.05694*, 2020.

[Tumanyan *et al.*, 2023] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[Wang *et al.*, 2020] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, and Gus Xia. POP909: A pop-song dataset for music arrangement generation. In Julie Cumming, Jin Ha Lee, Brian McFee, Markus Schedl, Johanna Devaney, Cory McKay, Eva Zangerle, and Timothy de Reuse, editors, *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, pages 38–45, 2020.

[Wang *et al.*, 2023] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and Sheng Zhao. AUDIT: Audio editing by following instructions with latent diffusion models. *arXiv preprint arXiv:2304.00830*, 2023.

[Wu *et al.*, 2023a] Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J Bryan. Music ControlNet: Multiple time-varying controls for music generation. *arXiv preprint arXiv:2311.07069*, 2023.

[Wu *et al.*, 2023b] 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. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[Wu *et al.*, 2023c] Yuxuan Wu, Yifan He, Xinlu Liu, Yi Wang, and Roger B Dannenberg. Transplayer: Timbre style transfer with flexible timbre control. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[Yao *et al.*, 2023] Yao Yao, Peike Li, Boyu Chen, and Alex Wang. Jen-1 composer: A unified framework for high-fidelity multi-track music generation. *arXiv preprint arXiv:2310.19180*, 2023.

[Zhang *et al.*, 2020] Yixiao Zhang, Ziyu Wang, Dingsu Wang, and Gus Xia. BUTTER: A representation learning framework for bi-directional music-sentence retrieval and generation. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 54–58, 2020.

[Zhang *et al.*, 2023] Yixiao Zhang, Akira Maezawa, Gus Xia, Kazuhiko Yamamoto, and Simon Dixon. Loop Copilot: Conducting AI ensembles for music generation and iterative editing. *arXiv preprint arXiv:2310.12404*, 2023.

[Zhao and Xia, 2021] Jingwei Zhao and Gus Xia. Accomontage: Accompaniment arrangement via phrase selection and style transfer. In Jin Ha Lee, Alexander Lerch, Zhiyao Duan, Juhan Nam, Preeti Rao, Peter van Kranenburg, and Ajay Srinivasamurthy, editors, *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, pages 833–840, 2021.