

SteerMusic: Enhanced Musical Consistency for Zero-shot Text-guided and Personalized Music Editing

Xinlei Niu^{1*}, Kin Wai Cheuk², Jing Zhang¹, Naoki Murata², Chieh-Hsin Lai², Michele Mancusi³, Woosung Choi², Giorgio Fabbro³, Wei-Hsiang Liao², Charles Patrick Martin¹, Yuki Mitsufuji²

¹Australian National University, Canberra, Australia;

²Sony AI, Tokyo, Japan;

³Sony Europe B.V., Stuttgart, Germany

xinlei.niu@anu.edu.au

Abstract

Music editing is an important step in music production, which has broad applications, including game development and film production. Most existing zero-shot text-guided editing methods rely on pretrained diffusion models by involving forward-backward diffusion processes. However, these methods often struggle to preserve the musical content. Additionally, text instructions alone usually fail to accurately describe the desired music. In this paper, we propose two music editing methods that improve the consistency between the original and edited music by leveraging score distillation. The first method, *SteerMusic*, is a coarse-grained zero-shot editing approach using delta denoising score. The second method, *SteerMusic+*, enables fine-grained personalized music editing by manipulating a concept token that represents a user-defined musical style. *SteerMusic+* allows for the editing of music into user-defined musical styles that cannot be achieved by the text instructions alone. Experimental results show that our methods outperform existing approaches in preserving both music content consistency and editing fidelity. User studies further validate that our methods achieve superior music editing quality.

Code — <https://github.com/sony/steermusic>

Demonstration page — <https://steermusic.pages.dev/>

Extended version — <https://arxiv.org/abs/2504.10826>

Introduction

Text-guided diffusion probabilistic models (DPMs) (Ho, Jain, and Abbeel 2020; Lai et al. 2025) have shown impressive performance in generating diverse high-quality audio samples, including music, speech, and sounds (Mariani et al. 2023; Zhang et al. 2024a; Niu et al. 2024). These text-to-audio (TTA) diffusion model trained with large scale datasets that can generate diverse samples conditioned on the natural language prompts specified. Consequently, the text-guided music editing task was proposed, which edits music by modifying the corresponding text prompts of the source music. Unlike controllable music generation, music editing task modifies an existing piece of music, which has

*Work done during the internship at Sony AI.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

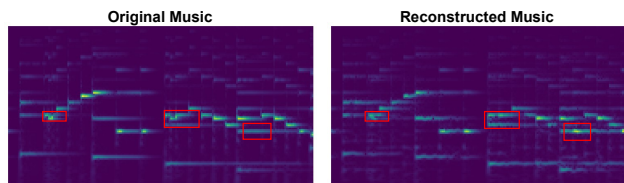


Figure 1: The distortion of the reconstructed melody (CQT1-PCC=0.721) after only 20 DDIM inversion steps.

two primary objectives: preserving original musical content and ensuring alignment between the edited music and the desired target.

Existing music editing methods focus on training music editing models from scratch (Copet et al. 2023) or fine-tuning pretrained TTA models (Zhang et al. 2024b), both of which require additional datasets or computational costs. Inspired by recent advances in image editing (Brooks, Holynski, and Efros 2023; Huberman-Spiegelglas, Kulikov, and Michaeli 2024; Hertz et al. 2022; Zhang et al. 2021), emerging music editing methods have instead pursued zero-shot techniques to reduce computational overhead. Existing zero-shot text-guided music editing pipelines (Zhang et al. 2024c; Manor and Michaeli 2024; Liu et al. 2024a) introduce noise into the source music during the forward diffusion process to suppress high-frequency components (e.g., timbral information) and subsequently perform editing during the denoising phase based on target guidance in the diffusion latent space. The process of retrieving a noisy latent representation from the data is commonly referred as inversion step. Due to the imperfect diffusion inversion process, the latent representation obtained may not fully preserve the original music content. This issue becomes more serious when the source prompt used as the condition cannot accurately capture the detailed characteristics of the music input (Kawar et al. 2023; Paissan et al. 2023). We refer to the distortion in the inverted latent representation as an “inversion error”. Notably, this distortion occurs even during the reconstruction of only a few of DDIM inversion steps (Song, Meng, and Ermon 2021), which alters the melodic information in the reconstructed results compared to the original music as shown in the CQT spectrogram (Brown 1991) in Fig. 1. In

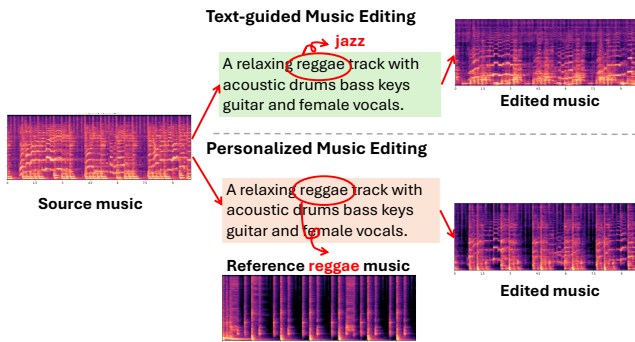


Figure 2: SteerMusic: Steering the music style with text-guided music editing or personalized music editing.

music editing, such distortion can even be compounded, resulting in a failure to preserve instruction-irrelevant content in original music. Although methods such as textual inversion (Gal et al. 2022) have been proposed to mitigate this issue by tuning the embeddings for near-lossless audio reconstruction (Niu, Zhang, and Martin 2024), there is no way to manipulate the target text prompts within the tuned source textual embeddings in editing tasks. A more promising solution to avoid inversion error is the delta denoising score (DDS) (Hertz, Aberman, and Cohen-Or 2023), a score distillation method that performs editing directly in the data space, which defines a differentiable function rendering the source input. DDS computes the difference in denoising scores between the source and target prompts through a single forward step. This approach eliminates the dependency of a full or partial forward diffusion process that could introduce inversion errors. By operating in the data space, DDS enables high-fidelity editing while preserving instruction-irrelevant content on the input.

Text-guided editing enables flexible and intuitive modifications, requiring users to provide only an arbitrary text instruction to perform the desired edit. However, one limitation of text-guided music editing is that it lacks fine-grained control over the direction and nuance of editing. For instance, editing a guitar performance into the one played by person A with a guitar brand B. Text-based editing alone struggles to specify the exact “guitar” required. Moreover, person A and guitar brand B might be unseen concepts for the models, rendering attempts to specify these words in the text prompt ineffective. To enhance user-personalized controllability in text-to-music generation, DreamSound (Plitsis et al. 2024) introduces a pioneering approach that adapts image personalization techniques (Gal et al. 2022; Ruiz et al. 2023) to the music domain, allowing the extraction of user-defined musical characteristics from reference audio. Besides, DreamSound also demonstrates the potential of leveraging the personalization techniques to perform personalized music editing by manipulating learned musical concepts on the noisy source latent through the denoising process of a personalized diffusion model. Despite their attempts, DreamSound still suffers from the inversion error, and struggles to preserve music content while editing the

specific concept given in the text prompt.

In this paper, we propose two music editing methods, *SteerMusic* and *SteerMusic+*, that can be easily adapted to existing text-to-music DPM based on the score distillation technique. We summarize our key contributions as follows.

1. We propose *SteerMusic*, a zero-shot text-guided music editing pipeline based on a DDS framework, which focuses on coarse-level editing, producing high-fidelity results while preserving source music contents.
2. We propose *SteerMusic+*, a personalized music editing method that leverages user-defined musical concepts to enable customized editing. *SteerMusic+*, an extension of *SteerMusic*, enables editing results that are not attainable through text prompts alone. For example, from reggae to the customized reggae given the reference as in Fig. 2.
3. We provide extensive experiments to demonstrate that the proposed methods produce superior editing results compared to the existing state-of-the-art methods in terms of musical consistency and edit fidelity.

Related Work

Text-guided Music Generation and Editing

Earlier music generation work focused on low-level control signals with strict temporal alignment, such as lyrics (Yu, Srivastava, and Canales 2021; Dhariwal et al. 2020) and MIDI (Yang, Chou, and Yang 2017) conditioning. Recently, high-level semantic prompts have gained popularity (Liu et al. 2023a, 2024b; Kundu, Singh, and Iwahori 2024; Huang et al. 2023b; Chowdhury et al. 2024; Agostinelli et al. 2023; Kreuk et al. 2022; Huang et al. 2023a; Le Lan et al. 2024; Li et al. 2024; Saito et al. 2025). More recent studies further explore melody prompts from reference music (Copet et al. 2024; Novack et al. 2024b,a; Hou et al. 2025; Chen et al. 2024; Kumari et al. 2023), enabling more precise, user-driven generation.

Music editing transforms existing audio according to target conditions while preserving the original content. Earlier approaches train models from scratch (Wang et al. 2023a; Agostinelli et al. 2023; Copet et al. 2023; Hou et al. 2025; Mariani et al. 2023) or fine-tune pretrained models (Paissan et al. 2023; Zhang et al. 2024b; Tsai et al. 2024; Han et al. 2023; Liu et al. 2023b). Recent work explores zero-shot methods (Zhang et al. 2024c; Manor and Michaeli 2024; Liu et al. 2024a) using pretrained TTA generator. However, these pipelines rely on a forward-backward diffusion process, which can introduce inversion errors as in Fig. 3 (a).

Personalized Music Generation and Editing

DreamSound (Plitsis et al. 2024) explores the possibility of capturing musical concepts from the given reference music using a personalization diffusion model (Ruiz et al. 2023; Gal et al. 2022; Kumari et al. 2023). This method allows users to generate new music samples by incorporating the captured personalized musical concept token into the text prompts. In addition to personalized music generation (Plitsis et al. 2024; Chen et al. 2024), DreamSound further extends their method to personalized music editing. However,

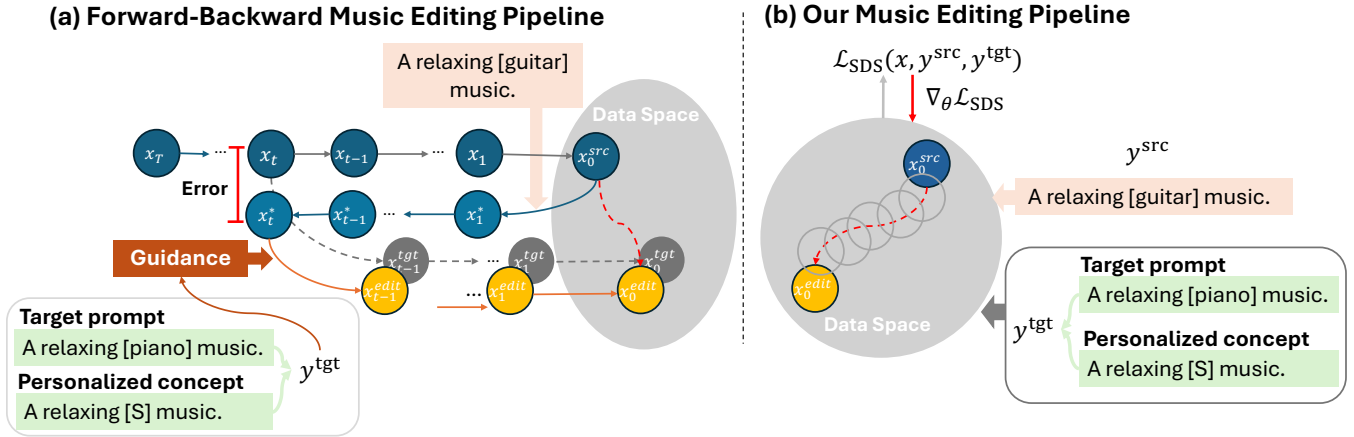


Figure 3: Overview of two music editing pipelines: (a) shows the conventional approach, which performs editing during denoising after an inversion process in the diffusion latent space; (b) shows our solution, which directly edits in data space by optimizing the differentiable function $x = g(\theta)$. The differentiable function is initialized with x^{src} . $[S]$ denotes a user-defined concept token, and gray circles represent the optimization trajectory from source to target.

we noticed that personalized music editing is still immature. Existing methods (Plitsis et al. 2024) struggle to maintain the musical consistency while editing into the desired music concept captured in the reference audio, which use the same music editing pipeline illustrated in Fig. 3 (a).

Preliminaries

Score distillation refines generated samples using the score (i.e., the gradient of the log-density) from a pretrained diffusion model ϵ_ϕ to enforce predefined constraints. It is typically implemented via probability density distillation, where gradients from the source diffusion model are used to iteratively refine a differentiable function until the desired outcome is achieved. Score distillation sampling (SDS) (Poole et al. 2022) pioneered score distillation by optimizing a differentiable function $x = g(\theta)$ to match a target prompt y^{tgt} , where the function $g(\theta)$ renders the source input x with parameters θ . It minimizes the loss $\mathcal{L}_{\text{Diff}} = \mathbb{E}_{t,\epsilon} [w(t) \|\epsilon_\phi(x_t, y^{\text{tgt}}, t) - \epsilon\|_2^2]$, where x_t is a noisy version of x at time t . By omitting the UNet Jacobian, the gradient is $\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, x = g(\theta), y^{\text{tgt}}) = \mathbb{E}_{t,\epsilon} [w(t) (\epsilon_\phi(x_t, y^{\text{tgt}}, t) - \epsilon) \frac{\partial x}{\partial \theta}]$, where y^{tgt} is the target prompt, x_t is a noisy latent of $x = g(\theta)$ at time step t , and $w(t)$ is a weighting function. However, SDS often produces blurry results in image editing (Wang et al. 2023b; Hertz, Aberman, and Cohen-Or 2023). The delta denoising score (DDS) (Hertz, Aberman, and Cohen-Or 2023) addresses this issue by computing the delta score between the source prompt y^{src} and the target prompt y^{tgt} .

In image editing, DDS refines only regions relevant to the target prompt y^{tgt} , preserving the rest of the image. Given source input x^{src} with prompt y^{src} and target prompt y^{tgt} , the

gradient over θ is

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{DDS}}(\phi, x = g(\theta), y^{\text{tgt}}, x^{\text{src}}, y^{\text{src}}) \\ = \mathbb{E}_{t,\epsilon} [w(t) (\epsilon_\phi(x_t, y^{\text{tgt}}, t) - \epsilon_\phi(x_t^{\text{src}}, y^{\text{src}}, t)) \frac{\partial x}{\partial \theta}] \quad (1) \end{aligned}$$

where x_t and x_t^{src} share the same sampled noise ϵ and the timestep t .

Although different variants of SDS and DDS have been proposed for the image domain (Yu, Yang, and Zhang 2025; Hertz, Aberman, and Cohen-Or 2023; Nam et al. 2024; Lin et al. 2025), the application of DDS in music editing remains underexplored. In the next section, we will explain how to incorporate DDS into music editing tasks.

Method

In this section, we introduce two music editing methods: *SteerMusic* for zero-shot text-guided editing, and *SteerMusic+* for personalized editing using user-defined concepts from reference music. Unlike the forward-backward pipeline in Fig.3 (a), our methods edit directly in a data space (Fig.3 (b)), yielding better musical content consistency.

SteerMusic: Zero-shot Text-guided Music Editing

We introduce *SteerMusic*, a zero-shot text-guided music editing method that performs editing in the data space. In this setting, our goal is to edit a source music signal x^{src} by modifying its corresponding text prompt y^{src} , which y^{src} is a brief description of the source music that includes the specific musical attribute intended for modification. The modified text prompt y^{tgt} acts as the target prompt to guide the editing. To obtain desirable results, the musical content shared by y^{tgt} and y^{src} should be preserved, changing only the content that is distinct in y^{tgt} . For example, if the only change in y^{tgt} compared to y^{src} is replacing only the word “piano” with “guitar” while keeping the rest of the sentence

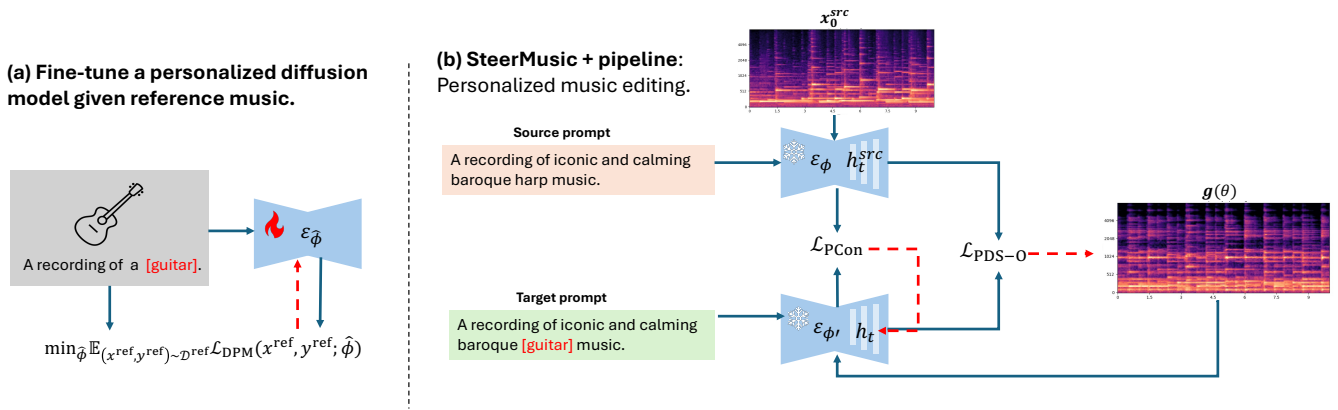


Figure 4: Overview of the SteerMusic+ pipeline: (a) Personalized diffusion model (PDM) fine-tuned using \mathcal{D}^{ref} and a user-defined [guitar] concept token. (b) Personalized editing using the PDM $\epsilon_{\phi'}$ from (a). Red dashed lines indicate gradient flows.

unchanged, the edited music should preserve the melody and tempo, while modifying only the musical instrument.

To achieve this goal, we adopt the DDS, which has been previously explored only in the image domain. To the best of our knowledge, we are the first to investigate the potential application in music editing. Following DDS, we define $x = g(\theta)$ rendering a source music signal x^{src} . We picked $g(\theta) = \theta$ as the differentiable function in Eq. 1, where we initialize $\theta = x_0^{\text{src}}$. i.e. $x = x_0^{\text{src}}$. In SteerMusic, we set ϵ_{ϕ} as a pretrained TTA or text-to-music DPMs.

Similar to Hertz, Aberman, and Cohen-Or (2023), the delta score in Eq. 1 steers the optimization process toward the target prompt while reducing the noisy editing direction commonly associated with vanilla SDS, leading to enhanced edit fidelity. SteerMusic method enables flexible editing using text instructions alone; however, it is limited to coarse-grained editing, as text prompts often lack the precision information to capture fine-grained musical details.

SteerMusic+: Personalized Music Editing

As mentioned before, text-guided editing lacks customization for precise music editing, such as transferring music to a specific style. To enable fine-grained personalized editing, we propose *SteerMusic+*, an extension of SteerMusic. In this setting, we have a set of source music and prompt pair $\{x^{\text{src}}, y^{\text{src}}\}$. The target prompt y^{tgt} is constructed by modifying y^{src} to manipulate a user-defined concept token $[S]$, representing the desired customization (see example in Fig. 4 (b)). SteerMusic+ uses two pretrained diffusion models:

- A pretrained diffusion probabilistic model (DPM), denoted as ϵ_{ϕ} , which serves as a reference for maintaining consistency with the original music content; and
- A personalized diffusion model (PDM), denoted as $\epsilon_{\phi'}$, is fine-tuned on a small set of data containing reference music to capture the user-defined concept $[S]$ and guide the editing process toward the desired direction.

We define successful personalized editing by the criteria:

- The instruction-irrelevant part (i.e., $\{y^{\text{tgt}} \cap y^{\text{src}}\}$) in the source music should be maintained.

- The edited musical attributes should perceptually align with the intended personalized musical concept $[S]$.

We now present the SteerMusic+ method that enables personalized music editing with enhanced music consistency.

Personalized Diffusion Model (PDM) serves a foundational part in SteerMusic+. Since training a PDM has been extensively studied in Plitsis et al. (2024); Ruiz et al. (2023); Gal et al. (2022); Kumari et al. (2023), we assume the availability of a pretrained text-to-music PDM in SteerMusic+, as SteerMusic+ is a plug-in pipeline compatible with existing PDMs. The text-to-music PDM $\epsilon_{\phi'}$ captures the user-defined musical concept S by fine-tuning a pretrained DPM ϵ_{ϕ} under a small set of reference music $\mathcal{D}^{\text{ref}} = \{(x^{\text{ref}}, y^{\text{ref}})_n\}_{n=1}^N$, which N can be as few as 1 (Plitsis et al. 2024). The fine-tuning is achieved via optimizing the objective

$$\phi' \in \arg \min_{\hat{\phi}} \mathbb{E}_{(x^{\text{ref}}, y^{\text{ref}}) \sim \mathcal{D}^{\text{ref}}} \mathcal{L}_{\text{DPM}}(x^{\text{ref}}, y^{\text{ref}}; \hat{\phi}) \quad (2)$$

where $\hat{\phi}$ is initialized with a pretrained DPM weights ϕ . As illustrated in Fig. 4 (a), the prompt y^{ref} takes form of “a recording of a $[S]$ ”, where the placeholder $[S]$ corresponds to a defined new concept word embedding. During inference, the PDM can generate music with the newly learned concept (e.g., “A disco song with a $[S]$ ”). The dataset \mathcal{D}^{ref} consists of reference audio clips that encapsulate the concept users aim to extract. According to Plitsis et al. (2024), the concept represents a musical style, which can be either an instance of instrument sounds or a specific genre that cannot be yielded even with the most detailed textual description. For instance, if the objective is to capture the user’s guitar playing style, the reference audio should feature performances on the user’s guitar playing. Conversely, if the goal is to capture the concept of jazz, the reference audio should consist of recordings that exemplify the same jazz style.

While the PDM can generate new music based on concepts from \mathcal{D}^{ref} , it cannot directly edit existing music using y^{tgt} that contains the concepts. Next, we introduce key components of SteerMusic+ that enable personalized editing with PDMs.

Personalized Delta Score (PDS) is an extension of Eq. 1 to enable personalized music editing. We define the PDS loss as $\mathcal{L}_{\text{PDS}}(\phi', \phi, x = g(\theta), y^{\text{tgt}}, x^{\text{src}}, y^{\text{src}}) = \mathbb{E}_{t, \epsilon} [w(t) \|\epsilon_{\phi'}(x_t, y^{\text{tgt}}, t) - \epsilon_{\phi}(x_t^{\text{src}}, y^{\text{src}}, t)\|_2^2]$. By omitting the UNet Jacobian, the gradient over θ is given by $\nabla_{\theta} \mathcal{L}_{\text{PDS}} = \mathbb{E}_{t, \epsilon} [w(t) (\epsilon_{\phi'}(x_t, y^{\text{tgt}}, t) - \epsilon_{\phi}(x_t^{\text{src}}, y^{\text{src}}, t)) \frac{\partial x}{\partial \theta}]$, where x_t and x_t^{src} share the same sampled noise ϵ at the time step t . This modified delta score can be decomposed into two components: the score of $\epsilon_{\phi'}$ provides the desired direction to guide the editing to match the target prompt with the concept [S]. The score of ϵ_{ϕ} reduces the noisy direction of unintended modification areas. The delta score between the PDM $\epsilon_{\phi'}$ and the DPM ϵ_{ϕ} may not produce an effective direction toward y^{tgt} , as $\epsilon_{\phi'}$ shifted to the reference distribution \mathcal{D}^{ref} . We introduce an additional component to compensate for the distribution shift induced by the score of $\epsilon_{\phi'}$.

Distribution Shift Regularization. To bridge the distribution gap between $\epsilon_{\phi'}$ and ϵ_{ϕ} , we introduce a regularization term to regularize the edited score to the personalized diffusion model $\epsilon_{\phi'}$. We wish to minimize the distribution shift between two diffusion models by adding a constraint as

$$\min_{\theta} \mathcal{L}_{\text{PDS}}(\phi', \phi, x = g(\theta), y^{\text{tgt}}, x^{\text{src}}, y^{\text{src}}), \quad (3)$$

$$\text{subject to } \mathcal{L}_{\text{shift}}(\phi', \phi, x = g(\theta), y^{\text{tgt}}) - \zeta \leq 0.$$

where ζ is a small amount of constant; the regularization term is $\mathcal{L}_{\text{shift}}(\phi', \phi, x = g(\theta), y^{\text{tgt}}) = \mathbb{E}_{t, \epsilon} [w(t) \|\epsilon_{\phi'}(x_t, y^{\text{tgt}}, t) - \epsilon_{\phi}(x_t, y^{\text{tgt}}, t)\|_2^2]$. The gradient in respect to θ is given by $\nabla_{\theta} \mathcal{L}_{\text{shift}} = \mathbb{E}_{t, \epsilon} [w(t) (\epsilon_{\phi'}(x_t, y^{\text{tgt}}, t) - \epsilon_{\phi}(x_t, y^{\text{tgt}}, t)) \frac{\partial x}{\partial \theta}]$. Therefore, the overall gradient through θ is

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{\text{PDS-O}}(\phi', \phi, x = g(\theta), y, x^{\text{src}}, y^{\text{src}}) \\ = \underbrace{\nabla_{\theta} \mathcal{L}_{\text{PDS}}(\phi', \phi, x = g(\theta), y^{\text{tgt}}, x^{\text{src}}, y^{\text{src}})}_{\text{'Delta score points to edit direction'}} \\ + \lambda \underbrace{\nabla_{\theta} \mathcal{L}_{\text{shift}}(\phi', \phi, x = g(\theta), y^{\text{tgt}})}_{\text{'Delta score regularizes distribution shift'}} \end{aligned} \quad (4)$$

where λ is a constant that adjusts regularization strength.

Personalized Contrastive (PCon) Loss. Eq. 4 formulates a regularized reference guided editing direction, where the regularized delta score encourages alignment with the target prompt while mitigating the distribution shift. However, it does not explicitly enforce the fidelity to the concept, which may lead to suboptimal editing quality. To further enhance the fidelity of the edit, we incorporate a PCon loss between temporal features, which is modified from a patch-wise contrastive loss (Nam et al. 2024). PCon loss extracts intermediate features h_l^{src} and h_l that pass through the residual block and the self-attention block from ϵ_{ϕ} conditioned on y^{src} and $\epsilon_{\phi'}$ conditioned on y^{tgt} at the l -th self-attention layer, respectively. The features are then reshaped to size $\mathbb{R}^{T_l \times F_l \times C_l}$, where T_l , F_l , and C_l represent the size of the temporal, spatial, and channel dimensions in the l -th layer, respectively. The patch corresponding to the temporal location on the feature map h_l^{src} is designated as ‘positive’, and vice versa. The PCon loss is defined as

$$\mathcal{L}_{\text{PCon}}(x, x^{\text{src}}) = \mathbb{E}_h \left[\sum_l \sum_{t'} \ell(h_l^{t'}, h_l^{\text{src}, t'}, h_l^{\text{src}, T_l \setminus t'}) \right] \quad (5)$$

$$\ell(h, h^+, h^-) = -\log \left(\frac{\exp(h \cdot h^+ / \tau)}{\exp(h \cdot h^+ / \tau) + \exp(h \cdot h^- / \tau)} \right)$$

where $t' \in \{1, \dots, T_l\}$ represents the temporal location query patch, the positive patch as $h_l^{\text{src}, t'}$ while the other patches as $h_l^{\text{src}, T_l \setminus t'}$. $\exp(h \cdot h^+ / \tau)$ is a positive sample with the same temporal location, $\exp(h \cdot h^- / \tau)$ is a negative sample with a mismatched temporal location in the self-attention features, τ is a temperature parameter as $\tau > 0$.

The gradient of $\mathcal{L}_{\text{PCon}}(x, x^{\text{src}})$ will propagate to the hidden state of self-attention layers h in personalized diffusion $\epsilon_{\phi'}$. Given that the personalized diffusion $\epsilon_{\phi'}$ has a distribution shift over the reference dataset \mathcal{D}^{ref} , the $\mathcal{L}_{\text{PCon}}$ explicitly encourages feature similarity at the frequency domain in self-attentions, particularly for attributes that distinguish the target concept. This reinforcement leads the model to prioritize concept consistency over strict temporal alignment with the source music. Consequently, $\mathcal{L}_{\text{PCon}}$ amplifies the distinctive characteristics of the target concept of $\epsilon_{\phi'}$ in SteerMusic+, ensuring that the edited music maintains stronger fidelity to the desired style while allowing structural variations.

Experiments

Evaluation Metrics

We evaluate music editing objectively based on two aspects: musical consistency (content preservation before and after editing) and editing fidelity. We follow Manor and Michaeli (2024) and calculate the following objective metrics for measurement. To evaluate *musical consistency*, we use:

- **Fréchet Audio Distance (FAD)** (Kilgour et al. 2019) which measures the distributional difference between source and edited music (lower the better). We calculate FAD based on both VGGish (Hershey et al. 2017) and clap-laion-music (Wu et al. 2023) embeddings, denoted as FAD_{VGGish} and FAD_{CLAP} respectively.
- **LPAPS** (Iashin and Rahtu 2021), an audio version of LPIPS (Zhang et al. 2018), which quantifies the consistency of the edited audio relative to the source audio (lower the better).
- **Top-1 Constant-Q Transform Pearson Correlation Coefficient (CQT1-PCC)**, which measures melody consistency between the source and edited music (higher the better). The CQT1-PCC (Brown 1991) extracts the main melody of the music audio and has been shown to outperform traditional chroma-based features in representing melodic characteristics (Hou et al. 2025). While existing metrics such as FAD and LPAPS provide insight into audio quality and perceptual similarity, they fall short in comprehensively capturing melodic structure. Moreover, existing transcription models (Cwitkowitz et al. 2024; Bittner et al. 2022; Gardner et al. 2022; Chang et al. 2024; Mancusi et al. 2025) effectively extract melodies from real music, they are unreliable for synthesized audio. To address these limitations, we introduce CQT1-PCC as a supplementary objective metric, specifically designed to quantify melodic consistency in generative

Method	FAD _{CLAP} ↓	FAD _{Vggish} ↓	CQT1-PCC↑	LPAPS↓	CLAP↑	MOS-P↑	MOS-T↑
DDIM	0.477	4.713	0.330	5.377	0.264	1.37	1.91
SDEdit	0.638	6.749	0.169	6.208	0.218	0.92	1.68
MusicMagus	0.593	7.631	0.338	5.243	0.238	2.11	1.57
ZETA	0.509	3.380	0.293	5.458	0.252	1.22	1.60
SteerMusic	0.278	2.426	0.480	3.772	<u>0.259</u>	2.92	2.50

Table 1: Model comparison on zero-shot text-guided music editing task using the ZoME-Bench dataset.

music editing. This metric enables a more targeted evaluation of whether the core melodic structure of the source audio is retained in the edited result.

To evaluate *editing fidelity*, we use:

- **CLAP Score** (Wu et al. 2023) which measures the alignment between edited music and the target prompt in text-guided music editing (higher is better).
- **CDPAM** (Manocha et al. 2021), a perceptual audio metric that leverages deep learning representations to measure perceptual distance between audios such as music and speech (Jacobellis, Cummings, and Yadwadkar 2024; Hai et al. 2024; Gui et al. 2024). We use CDPAM to evaluate audio perceptual similarity between reference music and the edited result in personalized music editing (lower the better).

Subjective evaluation. We designed a mean opinion score (MOS) study to evaluate target editing fidelity (MOS-T) and source content preservation (MOS-P) by asking participants to rate the results from 1-Bad to 5-Excellent (Sector 1996) for randomly selected edited samples. We provide more experimental details in our extended version.

Zero-shot Text-guided Music Editing

In this part, we evaluate our SteerMusic method on the zero-shot text-guided music editing task.

Dataset. We use the ZoME-Bench dataset (Liu et al. 2024a) which includes 1,000 10-second audio samples from MusicCaps (Agostinelli et al. 2023), each paired with source and target text prompts. We evaluate our models on four well-defined editing tasks that require modifying a specific aspect of the audio while preserving the original melody: change instrument (131 clips), change genre (134), change mood (100), and change background (95). To assess long-form editing, we use the MusicDelta subset of MedleyDB (Bittner et al. 2016) with ranging from 20 seconds to 5 minutes, comprising 34 excerpts of varying styles and lengths, with prompts from Manor and Michaeli (2024).

Baseline. We compare SteerMusic with zero-shot text-guided music editing methods plug-in the same pretrained AudioLDM2 (Liu et al. 2024b), including SDEdit (Meng et al. 2021), DDIM (Song, Meng, and Ermon 2021), ZETA (Manor and Michaeli 2024), and MusicMagus (Zhang et al. 2024c). To ensure statistical reliability, experiments are conducted using multiple random seeds. We are unable to include MelodyFlow (Le Lan et al. 2024) and MEDIC (Liu et al. 2024a) due to the lack of source code. We exclude AudioEditor (Jia et al. 2025) and AudioMorphX (Liang et al.

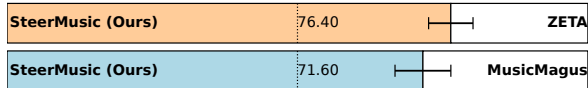


Figure 5: User preference for SteerMusic: percentage of users preferring our method over ZETA and MusicMagus.

2024), which are designed for general sound editing rather than music.

Experimental results. Tab.1 compares SteerMusic with zero-shot baselines across various style transfer tasks. SteerMusic achieves higher source consistency, shown by improved CQT1-PCC, lower LPAPS, and FAD. While DDIM attains a slightly higher CLAP score (+5e-3), its low CQT1-PCC and high LPAPS indicate poor preservation of source content due to lack of further source consistency constraints during denoising. In contrast, SteerMusic effectively balances source consistency and edit fidelity, fulfilling the core objective of music editing. Furthermore, our method attains the highest MOS scores and yields statistically significant improvements over all baseline models. An ANOVA test shows MOS-P = 2.92 with p-value = 7.37×10^{-27} and MOS-T = 2.5 with p-value = 3.24×10^{-7} . These results demonstrate that SteerMusic provides substantially better editing performance in terms of both source-music consistency and editing fidelity as perceived by human listeners.

To assess real-world applicability, we further evaluate our method on MusicDelta dataset in Tab. 3, which consists of varying lengths of music clips. MusicMagus fails in this experiment as it was designed for 5-second editing and doesn't support long-form music, and hence is excluded. SteerMusic consistently outperforms other baselines in terms of edit fidelity and source consistency, demonstrating its robustness and effectiveness in handling longer and more complex music editing.

User preference study. We evaluate SteerMusic with the user preference study following the design in Manor and Michaeli (2024). To reduce cognitive load and improve reliability, we compare SteerMusic with the top-performing baselines, MusicMagus and ZETA. In this study, users were asked to answer a sequence of 20 questions, each question contains original music, an editing instruction, and two edited results. Users were instructed to select the edited result that better matches the instruction while preserving the main content of the original music. We collected 25 full responses, which the participants having a minimum of 1 and average of 5 years of music training. As shown in Fig. 5, our

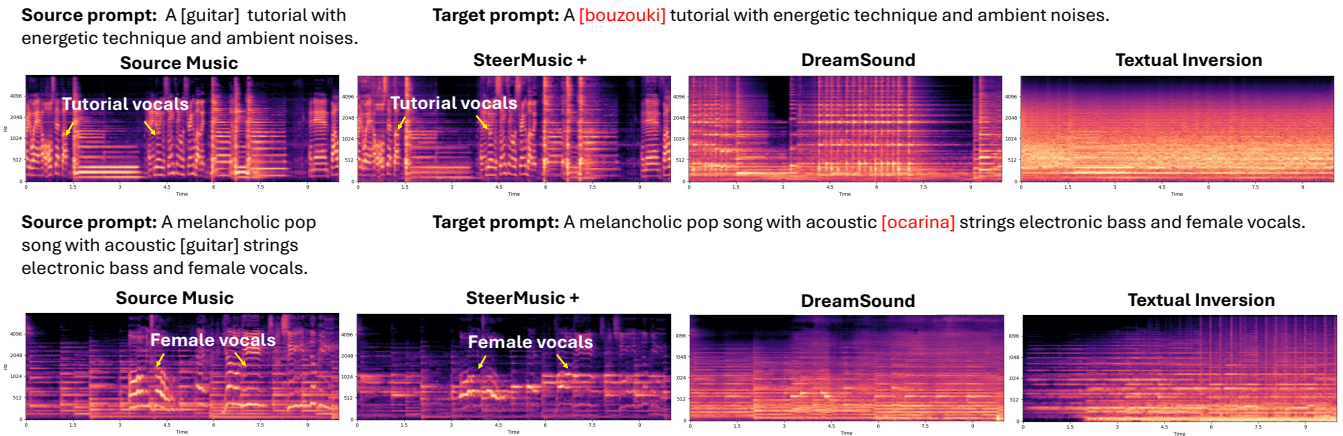


Figure 6: A visualization of edited results between SteerMusic+ and baselines in personalized music editing. SteerMusic+ preserves instruction-irrelevant musical content on the source music.

Method	FAD _{CLAP} ↓	FAD _{Vggish} ↓	CQT1-PCC↑	LPAPS↓	CDPAM↓	MOS-P↑	MOS-T↑
Textual Inv.	0.789	9.688	0.216	5.083	0.713	1.64	1.63
DreamSound	0.902	10.686	0.292	5.082	0.609	1.42	1.81
SteerMusic+	0.362	4.434	0.399	4.125	0.593	3.07	2.47

Table 2: Model comparison on personalized music editing task using the ZoME-Bench dataset.

method was clearly preferred over all competing methods.

Personalized Music Editing

In this part, we evaluate our SteerMusic+ method, which is designed for personalized music editing task.

Dataset. To cover both common and exotic concepts, we selected eight representative musical concepts from the 32 defined in Plitsis et al. (2024): four instruments style (Guitar, Bouzouki, Ocarina, Sitar) and four genres (Morricone, Reggae, Hiphop, Sarabande). Each concept includes a placeholder instruction and five 10-second reference clips from YouTube and FreeSound. We use the “change instrument” and “change genre” tasks from ZoME-Bench (Liu et al. 2024a), replacing the original target prompt with the selected concept token. Due to the lack of detailed instructions in MusicDelta, we use the standardized prompt as “A recording of a [style] song.”, where [style] is either the source style or the target style concept [S].

Baseline. We set two existing personalized music editing methods proposed by Plitsis et al. (2024) as the baselines, Textual inversion and DreamSound. Textual inversion optimizes a concept embedding, whereas DreamSound fine-tunes an AudioLDM2 with rare-token identifiers (Ruiz et al. 2023). Both methods perform personalized music editing by manipulating the concept token during denoising process. We follow the official codes provided by Plitsis et al. (2024) to obtain text-to-music PDM. We reproduce the personalized music editing methods by calculating noisy latent representations x_t from x_0^{src} of a DPM conditioned on the source prompt with a predefined shallow time step t using DDIM inversion (Song, Meng, and Ermon 2021), where $t = 30$. We

denoise x_t on the PDM¹ conditioned on the target prompt linked to the learned concept to obtain x_0^{tgt} . To ensure statistical reliability, experiments are conducted using multiple random seeds. We do not compare with Jen-1 DreamStyler (Chen et al. 2024), which is a personalized music generation method.

Experimental results. We plug SteerMusic+ into the PDM used in DreamSound¹. Tab. 2 shows that SteerMusic+ achieves superior musical consistency compared to the baselines. It indicates that the edited outputs of SteerMusic+ successfully preserve instruction-irrelevant music content in the source music. Furthermore, SteerMusic+ performs accurate edits that align well with the concepts captured from references, as indicated by the low CDPAM compared to the baseline methods. These objective evaluation results aligned with the subjective metrics, with SteerMusic+ obtaining significantly higher MOS-P and MOS-T than the baselines as measured by an ANOVA test (MOS-P with p-value 1.38×10^{-13} and MOS-T with p-value 1×10^{-3}). In long-form music editing (Tab. 3), SteerMusic+ still outperforms the baseline. We exclude Textual Inv. from comparison as it fails to produce meaningful results on MusicDelta, likely due to the limitation of its base model with complex inputs. Fig. 6 shows a personalized instrument style transfer example, where SteerMusic+ effectively preserves instruction-irrelevant content from the source music.

User preference study. We evaluate SteerMusic+ by a

¹ Textual inversion optimizes only the concept token embedding rather than fine-tuning a PDM, we denoise x_t using the DPM employed during DDIM inversion. Consequently, textual inversion is incompatible with SteerMusic+ pipeline, which relies on a PDM.

Method	FAD _{CLAP} ↓	FAD _{Vggish} ↓	CQT1-PCC↑	LPAPS↓	CDPAM↓	CLAP↑
DDIM	0.646	3.336	0.245	4.972	-	0.316
ZETA	0.665	3.789	0.296	5.071	-	0.319
SDEdit	0.818	8.757	0.137	5.996	-	0.310
SteerMusic	0.622	2.559	0.351	4.122	-	0.321
DreamSound	0.847	8.972	0.220	5.318	0.583	-
SteerMusic+	0.666	5.506	0.273	4.574	0.581	-

Table 3: Model comparison of SteerMusic and SteerMusic+ on MusicDelta dataset.

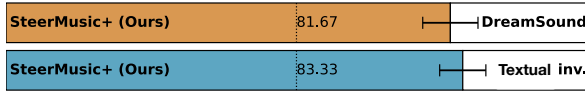


Figure 7: User preference for SteerMusic+: percentage of users preferring our method over the baselines, DreamSound and Textual Inv.

user preference study compared to DreamSound and Textual Inv. Users were asked to select the edited result that best matches the reference style while preserving the main content of the source music. We collected 24 full responses, excluding responses that are partially finished. Participants have a minimum of 1 and an average of 5 years of music training experience. As shown in Fig. 7, SteerMusic+ is clearly preferred by participants compared to baselines.

Ablation study. We conduct an ablation study to understand the effects of different components in SteerMusic+. We used the concept [bouzouki] as it is an uncommon musical instrument that typically requires the use of personalized models. The regularization weight λ , which acts as the Lagrange multiplier for the constraint in Eq. 4, was tested within $[-1,1]$ to balance the constraint enforcement and the optimization stability. Yellow dots indicate results obtained by \mathcal{L}_{PDS-O} in Eq. 4 with varying λ values. From the zoomed-in view in Fig. 8, as the CDPAM value decreases with λ , the loss of \mathcal{L}_{shift} within \mathcal{L}_{PDS-O} helps steer the editing toward the concept. However, the effect of \mathcal{L}_{shift} is minor, suggesting that the gradient of \mathcal{L}_{PDS} alone does not sufficiently capture the target concept during editing. As a result, the edited music struggles to align with the reference musical characteristics. In contrast, SteerMusic+ (with \mathcal{L}_{PCon} in Eq. 5) leads to a significant decrease in the CDPAM value, which indicates better alignment with the target concept. Since \mathcal{L}_{PCon} was proposed to explicitly enhance the characteristics of the target concept on the editing, we observe that \mathcal{L}_{shift} becomes more effective in this setting. Specifically, when λ is negative, the edit preserves the original content, while positive λ pushes the edit toward the target concept.

This highlights a fundamental trade-off between the source music consistency and the adherence to the target style, consistent with the findings of Manor and Michaeli (2024). In personalized music editing, where target characteristics are derived from reference music, shifting the output towards the concept often disrupts the original content. This differs from text-guided editing, which follows abstract

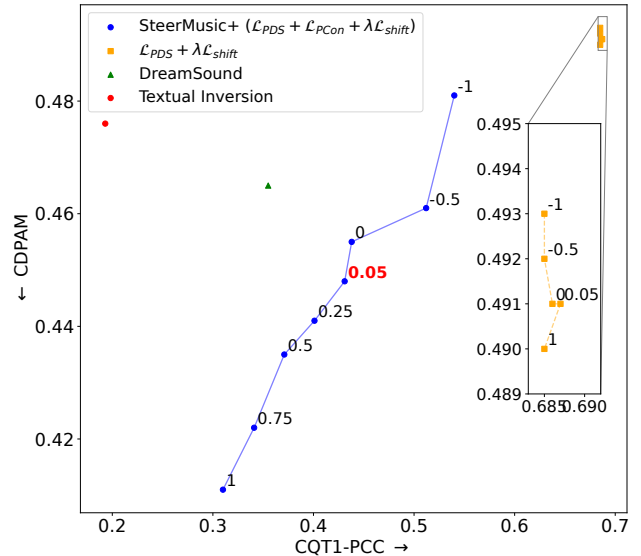


Figure 8: Ablation study on SteerMusic+ with adherence to music consistency vs. edit fidelity on the edited music with λ values in Eq. 4. The horizontal axis (CQT1-PCC) indicates source melody preservation; the vertical axis (CDPAM) indicates alignment with the target concept.

textual cues rather than concrete musical references. The key challenge in personalized editing is balancing the retention of the original music content, while integrating the distinctive attributes of the reference music. For practical use, we recommend keeping a smaller λ value (e.g., $\lambda = 0.05$) to avoid over-editing.

Limitation. As shown in Fig. 9, the number of fine-tuning steps in the PDM significantly affects the performance of DreamSound and SteerMusic+. A small number of steps (e.g., 50) leads to poor concept learning and low edit fidelity, while a large number of steps (e.g., 200) causes overfitting and source music structural loss. These findings highlight the importance of balancing PDM fine-tuning steps to achieve both high edit fidelity and content preservation in personalized music editing settings. The editing fidelity of SteerMusic+ is fundamentally limited by the ability of the PDMs to capture the concept attributes. Consequently, failure cases can often be attributed to inherent shortcomings in the PDM’s representational capacity. We refer readers to Plitsis et al. (2024), which investigated the capacity of PDMs

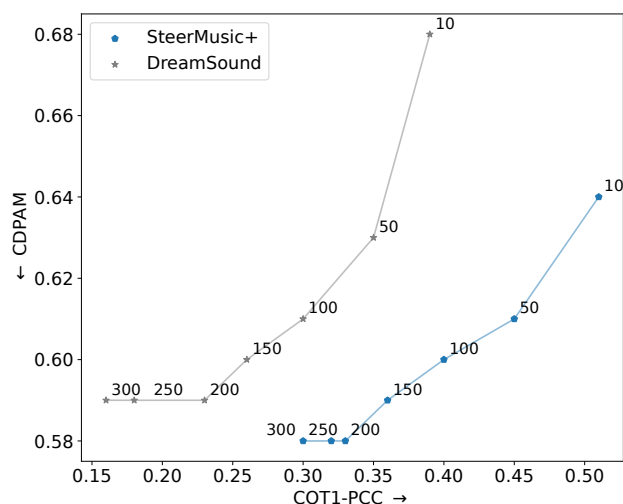


Figure 9: Adherence to music consistency vs. edit fidelity to edited results with different fine-tune steps in PDM. The horizontal axis (CQT1-PCC) indicates source melody preservation; the vertical axis (CDPAM) indicates alignment with the target concept.

to capture musical attributes and offers insights into improving this capacity.

Conclusion

We present two music editing methods, *SteerMusic* and *SteerMusic+*, from coarse-grained to fine-grained music editing. To the best of our knowledge, this is the first work to fully leverage DDS and score distillation in the music editing framework. Our methods address the limitations of prior approaches by enhancing musical consistency while producing high-fidelity edits aligned with target prompts. *SteerMusic+* further introduces a personalized editing pipeline that extracts user-defined style concepts from reference music for fine-grained control. We validate our methods through comprehensive experiments that show clear improvements over existing baselines in both consistency and fidelity. For practical application, the proposed methods could be extended by using a better-trained TTA diffusion backbone with a higher sampling rate to achieve high-fidelity editing results. In future studies, personalized music editing methods could focus on improving user-controllable editing outcomes for more nuanced and expressive edits.

References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Bittner, R.; Wilkins, J.; Yip, H.; and Bello, J. P. 2016. MedleyDB 2.0 Audio.

Bittner, R. M.; Bosch, J. J.; Rubinstein, D.; Meseguer-Brocal, G.; and Ewert, S. 2022. A lightweight instrument-

agnostic model for polyphonic note transcription and multi-pitch estimation. In *International Conference on Acoustics, Speech and Signal Processing*, 781–785. IEEE.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instruct-pix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Brown, J. C. 1991. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, 89(1): 425–434.

Chang, S.; Benetos, E.; Kirchhoff, H.; and Dixon, S. 2024. YourMT3+: Multi-Instrument Music Transcription with Enhanced Transformer Architectures and Cross-Dataset STEM Augmentation. In *International Workshop on MLSP*. IEEE.

Chen, B.; Li, P.; Yao, Y.; and Wang, A. 2024. JEN-1 DreamStyler: Customized Musical Concept Learning via Pivotal Parameters Tuning. *arXiv preprint arXiv:2406.12292*.

Chowdhury, S.; Nag, S.; Joseph, K.; Srinivasan, B. V.; and Manocha, D. 2024. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26826–26835.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36: 47704–47720.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Cwitkowitz, F.; Cheuk, K. W.; Choi, W.; Martínez-Ramírez, M. A.; Toyama, K.; Liao, W.-H.; and Mitsufuji, Y. 2024. Timbre-Trap: A Low-Resource Framework for Instrument-Agnostic Music Transcription. In *International Conference on Acoustics, Speech and Signal Processing*, 1291–1295. IEEE.

Dhariwal, P.; Jun, H.; Payne, C.; Kim, J. W.; Radford, A.; and Sutskever, I. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gardner, J. P.; Simon, I.; Manilow, E.; Hawthorne, C.; and Engel, J. 2022. MT3: Multi-Task Multitrack Music Transcription. In *International Conference on Learning Representations*.

Gui, A.; Gamper, H.; Braun, S.; and Emmanouilidou, D. 2024. Adapting frechet audio distance for generative music evaluation. In *International Conference on Acoustics, Speech and Signal Processing*, 1331–1335. IEEE.

Hai, J.; Wang, H.; Yang, D.; Thakkar, K.; Dehak, N.; and Elhilali, M. 2024. Dpm-tse: A diffusion probabilistic model for target sound extraction. In *International Conference on Acoustics, Speech and Signal Processing*, 1196–1200. IEEE.

- Han, B.; Dai, J.; Hao, W.; He, X.; Guo, D.; Chen, J.; Wang, Y.; Qian, Y.; and Song, X. 2023. Instructme: An instruction guided music edit and remix framework with latent diffusion models. *arXiv preprint arXiv:2308.14360*.
- Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing*, 131–135. IEEE.
- Hertz, A.; Aberman, K.; and Cohen-Or, D. 2023. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2328–2337.
- Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.
- Hou, S.; Liu, S.; Yuan, R.; Xue, W.; Shan, Y.; Zhao, M.; and Zhang, C. 2025. Editing Music with Melody and Text: Using ControlNet for Diffusion Transformer. In *International Conference on Acoustics, Speech and Signal Processing*, 1–5.
- Huang, Q.; Park, D. S.; Wang, T.; Denk, T. I.; Ly, A.; Chen, N.; Zhang, Z.; Zhang, Z.; Yu, J.; Frank, C.; et al. 2023a. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*.
- Huang, R.; Huang, J.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; and Zhao, Z. 2023b. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, 13916–13932. PMLR.
- Huberman-Spiegelglas, I.; Kulikov, V.; and Michaeli, T. 2024. An edit friendly ddpn noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12469–12478.
- Iashin, V.; and Rahtu, E. 2021. Taming Visually Guided Sound Generation. In *British Machine Vision Conference*. BMVA Press.
- Jacobellis, D.; Cummings, D.; and Yadwadkar, N. J. 2024. Machine Perceptual Quality: Evaluating the Impact of Severe Lossy Compression on Audio and Image Models. *arXiv preprint arXiv:2401.07957*.
- Jia, Y.; Chen, Y.; Zhao, J.; Zhao, S.; Zeng, W.; Chen, Y.; and Qin, Y. 2025. AudioEditor: A Training-Free Diffusion-Based Audio Editing Framework. In *International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.
- Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.
- Kilgour, K.; Zuluaga, M.; Roblek, D.; and Sharifi, M. 2019. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Proc. Inter-speech 2019*, 2350–2354.
- Kreuk, F.; Synnaeve, G.; Polyak, A.; Singer, U.; Défossez, A.; Copet, J.; Parikh, D.; Taigman, Y.; and Adi, Y. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.
- Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.
- Kundu, S.; Singh, S.; and Iwahori, Y. 2024. Emotion-Guided Image to Music Generation. *arXiv preprint arXiv:2410.22299*.
- Lai, C.-H.; Song, Y.; Kim, D.; Mitsufuji, Y.; and Ermon, S. 2025. The Principles of Diffusion Models. *arXiv preprint arXiv:2510.21890*.
- Le Lan, G.; Shi, B.; Ni, Z.; Srinivasan, S.; Kumar, A.; Ellis, B.; Kant, D.; Nagaraja, V. K.; Chang, E.; Hsu, W.-N.; et al. 2024. High Fidelity Text-Guided Music Editing via Single-Stage Flow Matching. In *Audio Imagination: Advances in Neural Information Processing Systems 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.
- Li, P. P.; Chen, B.; Yao, Y.; Wang, Y.; Wang, A.; and Wang, A. 2024. Jen-1: Text-guided universal music generation with omnidirectional diffusion models. In *2024 IEEE CAI*, 762–769. IEEE.
- Liang, J.; Yuan, Y.; Jia, D.; Zhuang, X.; Liu, Z.; Chen, Y.; Chen, Z.; Wang, Y.; and Wang, Y. 2024. AudioMorphix: Training-free audio editing with diffusion probabilistic models.
- Lin, Y.-B.; Lin, K.; Yang, Z.; Li, L.; Wang, J.; Lin, C.-C.; Wang, X.; Bertasius, G.; and Wang, L. 2025. Zero-Shot Audio-Visual Editing via Cross-Modal Delta Denoising. *arXiv preprint arXiv:2503.20782*.
- Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D. P.; Wang, W.; and Plumbley, M. D. 2023a. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *International Conference on Machine Learning*.
- Liu, H.; Wang, J.; Li, X.; Huang, R.; Liu, Y.; Xu, J.; and Zhao, Z. 2024a. Medic: Zero-shot music editing with disentangled inversion control.
- Liu, H.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Tian, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2024b. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM TASLP*.
- Liu, S.; Hussain, A. S.; Sun, C.; and Shan, Y. 2023b. M^2 UGEN: Multi-modal Music Understanding and Generation with the Power of Large Language Models. *arXiv preprint arXiv:2311.11255*.
- Mancusi, M.; Halychanskyi, Y.; Cheuk, K. W.; Moliner, E.; Lai, C.-H.; Uhlich, S.; Koo, J.; Martínez-Ramírez, M. A.; Liao, W.-H.; Fabbro, G.; et al. 2025. Latent Diffusion Bridges for Unsupervised Musical Audio Timbre Transfer. In *International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

- Manocha, P.; Jin, Z.; Zhang, R.; and Finkelstein, A. 2021. CDPAM: Contrastive learning for perceptual audio similarity. In *International Conference on Acoustics, Speech and Signal Processing*, 196–200. IEEE.
- Manor, H.; and Michaeli, T. 2024. Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion. *International Conference on Machine Learning*.
- Mariani, G.; Tallini, I.; Postolache, E.; Mancusi, M.; Cosmo, L.; and Rodolà, E. 2023. Multi-source diffusion models for simultaneous music generation and separation. *arXiv preprint arXiv:2302.02257*.
- Meng, C.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.
- Nam, H.; Kwon, G.; Park, G. Y.; and Ye, J. C. 2024. Contrastive denoising score for text-guided latent diffusion image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9192–9201.
- Niu, X.; Zhang, J.; and Martin, C. P. 2024. SoundMorpher: Perceptually-Uniform Sound Morphing with Diffusion Model. *arXiv preprint arXiv:2410.02144*.
- Niu, X.; Zhang, J.; Walder, C.; and Martin, C. P. 2024. Soundlocd: An efficient conditional discrete contrastive latent diffusion model for text-to-sound generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, 261–265. IEEE.
- Novack, Z.; McAuley, J.; Berg-Kirkpatrick, T.; and Bryan, N. 2024a. DITTO-2: Distilled diffusion inference-time t-optimization for music generation. *arXiv preprint arXiv:2405.20289*.
- Novack, Z.; McAuley, J.; Berg-Kirkpatrick, T.; and Bryan, N. J. 2024b. Ditto: Diffusion inference-time t-optimization for music generation. *International Conference on Machine Learning*.
- Paissan, F.; Della Libera, L.; Wang, Z.; Ravanelli, M.; Smaragdis, P.; and Subakan, C. 2023. Audio editing with non-rigid text prompts. *arXiv preprint arXiv:2310.12858*.
- Plitsis, M.; Kouzelis, T.; Paraskevopoulos, G.; Katsouros, V.; and Panagakis, Y. 2024. Investigating personalization methods in text to music generation. In *International Conference on Acoustics, Speech and Signal Processing*, 1081–1085. IEEE.
- Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.
- Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.
- Saito, K.; Kim, D.; Shibuya, T.; Lai, C.-H.; Zhong, Z.; Takida, Y.; and Mitsufuji, Y. 2025. SoundCTM: Unifying Score-based and Consistency Models for Full-band Text-to-Sound Generation. In *The Thirteenth International Conference on Learning Representations*.
- Sector, I. T. U. T. S. 1996. *Methods for subjective determination of transmission quality*. International Telecommunication Union.
- Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.
- Tsai, F.-D.; Wu, S.-L.; Kim, H.; Chen, B.-Y.; Cheng, H.-C.; and Yang, Y.-H. 2024. Audio Prompt Adapter: Unleashing Music Editing Abilities for Text-to-Music with Lightweight Finetuning. *ISMIR*.
- Wang, Y.; Ju, Z.; Tan, X.; He, L.; Wu, Z.; Bian, J.; et al. 2023a. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36: 71340–71357.
- Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023b. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36: 8406–8441.
- Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.
- Yang, L.-C.; Chou, S.-Y.; and Yang, Y.-H. 2017. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*.
- Yu, Y.; Srivastava, A.; and Canales, S. 2021. Conditional LSTM-GAN for melody generation from lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1): 1–20.
- Yu, Z.; Yang, Z.; and Zhang, J. 2025. DreamSteerer: Enhancing Source Image Conditioned Editability using Personalized Diffusion Models. *Advances in Neural Information Processing Systems*, 37: 120699–120734.
- Zhang, K.; Li, Y.; Zuo, W.; Zhang, L.; Van Gool, L.; and Timofte, R. 2021. Plug-and-play image restoration with deep denoiser prior. *IEEE TPAMI*, 44(10): 6360–6376.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE CVPR*, 586–595.
- Zhang, X.; Liu, D.; Liu, H.; Zhang, Q.; Meng, H.; Perera, L. P. G.; Chng, E.; and Yao, L. 2024a. Speaking in wavelet domain: A simple and efficient approach to speed up speech diffusion model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 159–171.
- Zhang, Y.; Ikemiya, Y.; Choi, W.; Murata, N.; Martínez-Ramírez, M. A.; Lin, L.; Xia, G.; Liao, W.-H.; Mitsufuji, Y.; and Dixon, S. 2024b. Instruct-MusicGen: Unlocking Text-to-Music Editing for Music Language Models via Instruction Tuning. *arXiv preprint arXiv:2405.18386*.
- Zhang, Y.; Ikemiya, Y.; Xia, G.; Murata, N.; Martínez-Ramírez, M. A.; Liao, W.-H.; Mitsufuji, Y.; and Dixon, S. 2024c. MusicMagus: zero-shot text-to-music editing via diffusion models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 7805–7813.