

# InstructME: An Instruction Guided Music Edit And Remix Framework with Latent Diffusion Models

Bing Han<sup>\*1 †</sup>, Junyu Dai<sup>\*2</sup>, Xuchen Song<sup>\*2</sup>, Weituo Hao<sup>2</sup>, Xinyan He<sup>2</sup>, Dong Guo<sup>2</sup>,  
Jitong Chen<sup>2</sup>, Yuxuan Wang<sup>2</sup>, Yanmin Qian<sup>1</sup>

<sup>1</sup>X-LANCE Lab, Department of Computer Science and Engineering, Shanghai Jiao Tong University

<sup>2</sup>ByteDance

{daijunyu.6, xuchen.song}@bytedance.com, {hanbing97, yanminqian}@sjtu.edu.cn

## Abstract

Music editing primarily entails the modification of instrument tracks or remixing in the whole, which offers a novel reinterpretation of the original piece through a series of operations. These music processing methods hold immense potential across various applications but demand substantial expertise. Prior methodologies, although effective for image and audio modifications, falter when directly applied to music. This is attributed to music's distinctive data nature, where such methods can inadvertently compromise the intrinsic harmony and coherence of music. In this paper, we develop InstructME, an **Instruction guided Music Editing** and remixing framework based on latent diffusion models. Our framework fortifies the U-Net with multi-scale aggregation in order to maintain consistency before and after editing. In addition, we introduce chord progression matrix as condition information and incorporate it in the semantic space to improve melodic harmony while editing. For accommodating extended musical pieces, InstructME employs a chunk transformer, enabling it to discern long-term temporal dependencies within music sequences. We tested InstructME in instrument-editing, remixing, and multi-round editing. Both subjective and objective evaluations indicate that our proposed method significantly surpasses preceding systems in music quality, text relevance and harmony. Demo samples are available at <https://musicedit.github.io/>

## Introduction

Music editing involves performing basic manipulations on musical compositions, including such atomic operations as the inclusion or exclusion of instrumental tracks and the adjustment of pitches in specific segments. On top of these atomic operations, remixing can be understood as an advanced version of music editing that mixes various atomic operations with style and genre considered (Fagerjord 2010). Both atomic operations and remix can be handled by using a text-based generative model. In music editing, the text would be natural language-based editing instructions, such as *"adding a guitar track"*, *"replacing a piano track with a violin"*, etc. Models from the text-generated image domain seem to be adaptable to the music editing

scenario. However, unlike image generation, models need to pay attention to the music harmony in addition to understanding the text and generating it. For example, when introducing a guitar track, care is taken to harmonize its rhythm, chord progression, and melodic motifs with the original audio framework, thus ensuring that overall consistency and coherence are maintained. Therefore, for successful music editing, the model should be able to: (i) understand editing instructions and generate music stems; (ii) ensure the compatibility of the part being processed with the original music source.

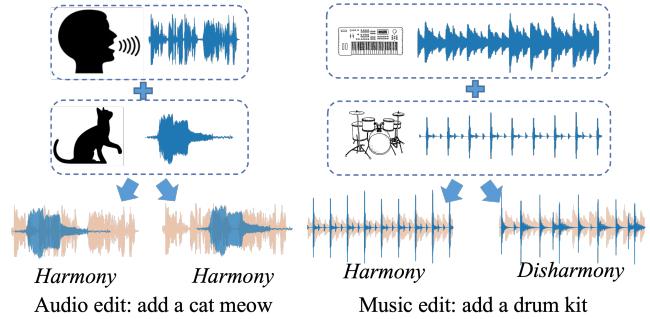


Figure 1: Left is audio edit: Each audio component is independent that does not necessitate the consideration of interdependence. Right is music edit: Harmony in pitch, intensity, rhythm, and timbre must be taken into account.

Lately, a multitude of endeavours pertaining to text-based image or audio manipulation (Hertz et al. 2022; Lugmayr et al. 2022; Meng et al. 2021; Wang et al. 2023) have attracted considerable attention due to their noteworthy performance within their respective domains. However, the distinct data properties and generative prerequisites inherent to the domain of music preclude the direct applicability of these methods to the sphere of music editing. In image editing, it is feasible to maintain consistency over the residual regions by employing masking techniques, thereby confining attention solely to the objects to be generated. However, this underlying principle proves inapplicable to the domain of musical data, as shown in Figure 1 the interwoven nature of individual tracks across both temporal and frequency do-

<sup>\*</sup>These authors contributed equally.

<sup>†</sup>Interns at ByteDance.

mains prevents the straightforward implementation of such an approach. Perhaps the most similar method to ours is (Wang et al. 2023) for audio editing. However, the method is mainly applied to the editing of sound effects. Unlike music tracks, the individual sound effects are independent of each other, so there is no need to consider whether the sound effects are in harmony with each other or not.

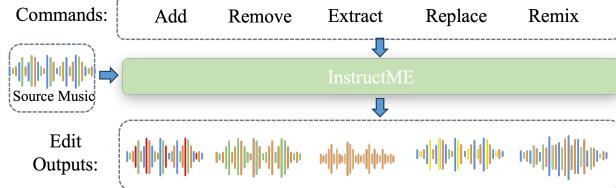


Figure 2: A brief illustration of InstructME. Given source music and command text, InstructME generates a piece of music that is harmonious and complies with the command requirement.

In order to bridge the gap between text-based generative models and music editing tasks, we propose InstructME, an instruction-guided music editing framework based on latent diffusion models. For simplicity, we limit music editing operations to adding, removing, extracting, replacing, and remixing. As shown in Figure 2, InstructME takes text instructions and source music as input, and outputs the target music accordingly. To maintain the consistency of the music before and after editing, we utilize the multi-scale aggregation strategy and incorporate the chord progression matrix into the semantic space (Kwon, Jeong, and Uh 2022; Jeong, Kwon, and Uh 2023) during the source music encoding process to ensure harmony. During training, we employ the chunk transformer to model long-term temporal dependencies of music data in a segmented chunk-wise manner and train the model on collected 417 hours of music data. For testing, we evaluate the model in terms of three aspects: music quality, text relevance and harmony. Experimental results of public and private datasets demonstrate that InstructME outperforms the previous system.

Our key contributions can be summarized as:

- To the best of our knowledge, we propose the first instruction guided music editing framework applicable for both atomic and advanced operations.
- We point out the special problem of consistency and harmony in music editing domain and develop multi-scale aggregation and chord condition via chunk transformer to solve it.
- We propose quantitative evaluation metrics for music editing tasks in terms of music quality, text relevance and harmony.
- Our proposed method InstructME surpasses previous systems through thorough subjective and objective tests.

## Related Work

### Text guided Generation

Generating a new version of accompaniment for a track directly with targeted properties (e.g. genre, mood, instru-

ments) adhering is a viable approach to accomplish the objectives of editing or remixing. Recent studies (Huang et al. 2023b; Liu et al. 2023a; Agostinelli et al. 2023; Schneider, Jin, and Schölkopf 2023; Huang et al. 2023a; Lam et al. 2023; Copet et al. 2023) have already succeeded in generating plausible music that reflects key music properties (e.g. genre, mood, etc) that are depicted in a given text. However, there is no guarantee for them to generate tracks that are harmonious with a given track while keeping the given one or specified part of it unchanged. Another work (Donahue et al. 2023) proposed a generative model, which trained over instrumentals given vocals, generating coherent instrumental music to accompany input vocals. But it has no way for users to control the generation process, not to mention interactive editing, which is important for an intelligent editing tool as it applies feedback from users to make a more preferable output as in (Holz 2023).

## Audio Editing and Music Remixing

(Huang et al. 2023b; Liu et al. 2023a) propose zero-shot audio editing by utilizing pre-trained text-to-audio latent diffusion models, which seem flexible but not accurate enough for the editing process. Moreover, there is no guarantee for those audio generation models that are trained with general purposes to achieve a good editing effect in the editing specialized usage scenario. Due to this, AUDIT (Wang et al. 2023) proposed a general audio editing model based on a latent diffusion and denoising process guided by instructions. Certainly, as previously stated in the introduction section, this framework necessitates certain enhancements to effectively cater to music-related tasks.

For remixing, although the text-guided generative systems mentioned above can also perform generation conditions on a given recording (Liu et al. 2023a; Lam et al. 2023), or more specifically, melodies (Agostinelli et al. 2023; Copet et al. 2023), the generated music can only preserve the tune of the conditional melodies, the original tracks, such as vocal, will not directly feature in the output music. This kind of conditional-generated music is traditionally known as music covers, not remixes. Likewise, past studies (Yang et al. 2022; Yang and Lerch 2020b; Wierstorf et al. 2017) have attempted to apply neural networks to the task of music remixing. These methods, which are often incorporated with source separation models, primarily viewed music remixing as a task of adjusting the gain of individual instrument sources of an audio mixture. But music remixing is not just limited to manipulating the gain of different sources of the recording itself, it can also involve incorporating other materials to create something new (Waysdorff 2021).

## Methodology

In this section, we will provide an overview of the InstructME architecture and the process of instruction-based music editing, as illustrated in Figure 3. Additionally, we will explain strategies aimed at improving editing consistency and harmony, as well as approaches to achieving more sophisticated music editing operations.

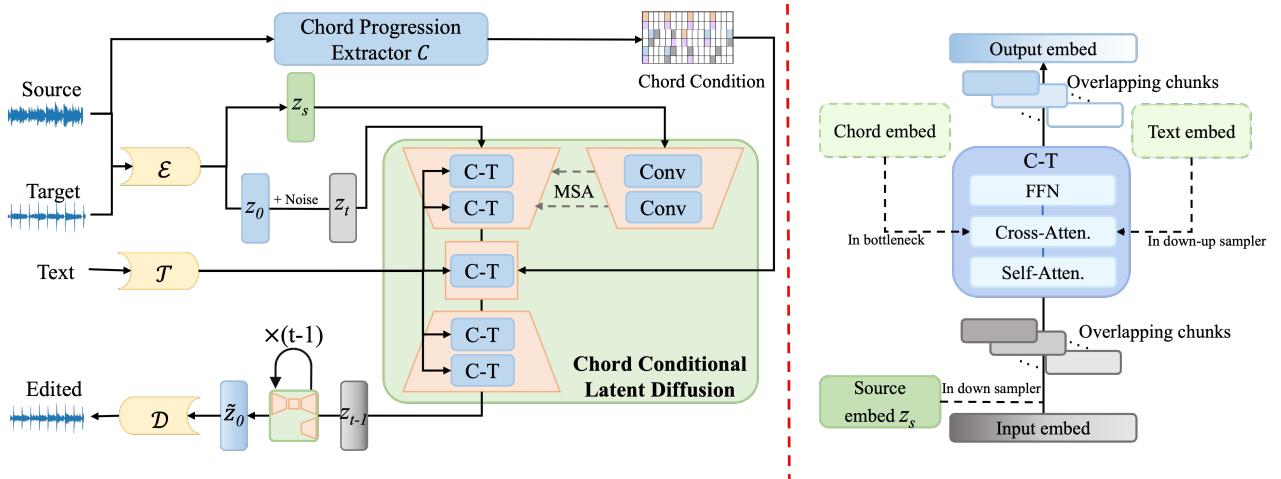


Figure 3: Left: Overview of InstructME diffusion process for music editing. Audio signal is processed by VAE (encoder  $\mathcal{E}$  and decoder  $\mathcal{D}$ ), meanwhile extractor ( $\mathcal{C}$ ) extracts the chord matrix of source music and together with text embedding extracted by  $\mathcal{T}$  as condition information, latent embedding  $z_s$  and  $z_t$  are fused by multi-scale aggregation and converted by chunk transformer to produce the final edited music. Right: Architecture of chunk transformer (C-T) blocks which in various positions of U-net will selectively incorporate chord or text embedding, and  $z_s$  will only input when chunk transformer is in down sampler.

## Instruction To Music Editing

InstructME accepts music audio  $\mathbf{x}_s$  and editing instructions  $y$  as input, and produces new audio  $\mathbf{x}$  that adheres to the given instructions. We utilize text and audio encoders to transform the data into a latent representation. For each text instruction  $y$ , a pretrained T5 (Raffel et al. 2020) converts it into sequence of embeddings  $\mathcal{T}(y) \in \mathbb{R}^{L \times D}$ , similar to (Wang et al. 2023). For each audio segment  $\mathbf{x}_s \in \mathbb{R}^{T \times 1}$ , a variational auto-encoder (VAE) transforms the waveform into a 2D latent embedding  $\mathbf{z}_s \in \mathbb{R}^{\frac{T}{r} \times C}$ . Using text and audio embeddings as conditions, a diffusion process (Song et al. 2020; Ho, Jain, and Abbeel 2020) produces embeddings of new audio samples, which the VAE decoder then converts back to audio waveforms.

The VAE used by InstructME consists of an encoder  $\mathcal{E}$ , a decoder  $\mathcal{D}$  and a discriminator with stacked convolutional blocks. The decoder reconstructs the waveform  $\hat{\mathbf{x}}$  from the latent space  $\mathbf{z}$  and there is no vocoder like (Wang et al. 2023; Liu et al. 2023a). The discriminator was used to enhance the sound quality of generated audio through adversarial training. We provide the model and training details in the Appendix.

**Diffusion Model** Diffusion model contains two processes. The forward process is a standard Gaussian noise injection process. At time step  $t$ ,

$$q(z_t | z_0) = \mathcal{N}(z_t; \sqrt{\alpha_t} z_0, (1 - \bar{\alpha}_t) \epsilon) \quad (1)$$

where  $\alpha_t = 1 - \beta_t$  and  $\bar{\alpha}_t = \prod_{t=1}^T \alpha_t$  are scheduling hyperparameters.

In the reverse process, we employ a time-conditional U-Net  $\epsilon_\theta$  (Ronneberger, Fischer, and Brox 2015; Rombach et al. 2022) as the denoise model backbone. At time step  $t$ , conditioning on embeddings of text  $\mathcal{T}(y)$  and source music  $z_s$ , this denoise model attempts to restore the original

latent  $z_0$  of target music from noisy  $z_t$ . For model optimization, we use reweighted bound (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) as objective function:

$$\mathcal{L}_{DM} = \mathbb{E}_{\epsilon, t, z_0} \|\epsilon - \epsilon_\theta(t, \mathcal{T}(y), z_s, z_t)\|_2^2 \quad (2)$$

with  $t$  uniformly sampled from  $[1, T]$  during the training. In the end, we pass  $z_0$  through the decoder  $\mathcal{D}$  to obtain the waveform music.

The U-Net layers utilize transformers with self and cross attention as building blocks. In the down sampler layers, source audio embeddings and generated embeddings merge into the input of the self-attention layer. Cross attention is employed for text conditions in each down-up sampler layer, and for chord conditions in the bottleneck layer.

**Efficient Diffusion** The self-attention of lengthy music sequences is computationally expensive. To alleviate this problem, we employ the chunk transformer to model long-term temporal dependencies in a chunk-wise manner. Outlined in Figure 3 (Right), the process involves three steps: segmentation of  $T$ -frame embeddings into  $K$ -frame chunks with 50% overlap, individual chunk processing through a transformer layer, and fusion to merge overlapping output chunks into  $T$  frames.

At each layer of the chunk transformer, a token from a  $K$ -frame chunk can observe  $\frac{3K}{2}$  neighboring frames. By stacking multiple layers of chunk transformer, the U-Net acquires an expansive receptive field, enabling effective modeling of long-term dependencies. Compared with oracle transformer's complexity  $\mathcal{O}(T^2)$ , chunk transformer has lower computational cost  $\mathcal{O}(2 * \lceil \frac{T}{K} \rceil * K^2) = \mathcal{O}(TK)$ . In addition to faster inference and lower memory consumption, the chunk-wise modeling approach decreases the model's reliance on sequence length, learning invariant representations. This minimizes performance degradation caused by duration differences in training and sampling.

## Improving Consistency and Harmony

To make the diffusion model more suitable for music editing tasks, we propose an enhanced U-Net with several modifications including multi-scale aggregation and chord condition.

**Multi-Scale Aggregation** Contrary to the music generation tasks (Huang et al. 2023a), music editing tasks require the preservation of certain content and properties from the original music. In order to maintain coherence between the original and edited music, AUDIT (Wang et al. 2023) directly concatenates the source music channel  $z_t$  with the target music channel  $z_s$  at the U-Net’s input. It leans heavily on the invariance of some low-level and local music features, which might pose challenges or limitations when applied to more complex music manipulation tasks. To more effectively capture the high-level characteristics of the source music, we introduce a multi-scale aggregation (MSA) strategy as depicted in Figure 3 (Left). The source music embeddings  $z_s$  are input to a multi-layer convolution encoder, yielding feature maps with varying resolutions for the corresponding U-Net layers. This strategy has been proven effective in high-resolution image generation (Karras et al. 2020).

**Chord-Conditional** The Chord progression is a key element in defining a piece’s musical harmony. We adopt a chord progression recognition model  $\mathcal{C}$  (Cheuk et al. 2022) to extract the chord probability embedding  $p$  of the source music and then emphasize it explicitly during the denoise process. (Kwon, Jeong, and Uh 2022; Jeong, Kwon, and Uh 2023) discover the semantic latent space in the bottleneck of diffusion has nice properties to accommodate semantic image manipulation. Inspired by them, we incorporate the chord progression representation  $p \in \mathbb{R}^{d_p \times T_p}$  in the bottleneck feature map  $h \in \mathbb{R}^{d_h \times T_h}$  of U-Net with cross-attention mechanism (Vaswani et al. 2017). With chord progression condition extracted by  $\mathcal{C}$ , in the bottleneck layer of U-Net, the objective function in Equation 2 can be rewritten as:

$$\mathcal{L}_{CDM} = \mathbb{E}_{\epsilon, t, z_0} \|\epsilon - \epsilon_\theta(t, p_s, z_s, z_t)\|_2^2 \quad (3)$$

where  $p_s$  denotes chord progression matrix of source music  $x_s$ , encoded by  $\mathcal{C}$ .

## Towards Advanced Music Editing - Remix

For diffusion models, there exist two primary strategies for achieving controllable generation. One of these is classifier guidance (CG) (Dhariwal and Nichol 2021; Liu et al. 2023b), which utilizes a classifier during the sampling process and mixes its input gradient of the log probability with the score estimate of diffusion model. It is flexible and controllable, but tends to suffer a performance degradation (Ho and Salimans 2022). Another approach, named classifier-free guidance (CFG) (Ho and Salimans 2022; Nichol et al. 2021; Ramesh et al. 2022; Saharia et al. 2022), achieves the same effect through training a conditional diffusion model directly without a guidance classifier. This method performs better but requires a large amount of data with diverse text descriptions, which is difficult for our InstructME trained with source-target paired data. In this work, to attain a trade-off between quality and controllability, we adopt both clas-

sifier and classifier-free guidance to achieve the controllable editing of Remix operations.

We specify instrument and genre tags with CFG by incorporating these tags into text commands to train the conditional diffusion models. During the training, we discard our text condition  $y$  randomly with a certain probability  $p_{CFG}$  following (Liu et al. 2023a; Wang et al. 2023). Then, in the sampling, we can estimate the noise  $\hat{\epsilon}_\theta(t, \mathcal{T}(y), p_s, z_s, z_t)$  with a linear combination of the conditional and unconditional score estimates:

$$\begin{aligned} \hat{\epsilon}_\theta(t, \mathcal{T}(y), p_s, z_s, z_t) &= (1 - w)\epsilon_\theta(t, p_s, z_s, z_t) \\ &+ w\epsilon_\theta(t, \mathcal{T}(y), p_s, z_s, z_t) \end{aligned} \quad (4)$$

where  $w$  can determine the strength of guidance.

To achieve finer-grained semantic control with weakly-associated, free-form text annotations, we apply classifier guidance during sampling with a pre-trained MuLan (Huang et al. 2022), which can project the music audio and its corresponding text description into the same embedding space. The guidance function we use is:

$$F(x_t, y) = \|E_L(y) - E_M(x_t)\|_2^2 \quad (5)$$

where  $E_L(\cdot)$  and  $E_M(\cdot)$  denote the language and music encoders respectively. Then, by adding the gradient on estimated  $x_t$ , we can guide the generation

$$\hat{x}_t = x_t + s\nabla_{x_t}F(x_t, y) \quad (6)$$

with factor  $s$  to control the guidance scale.

## Experiments Setup

### Dataset

We collected 417 hours of music audio. Each audio file consists of multiple instrumental tracks. We resampled audios to 24khz sample rate and divided them into non-overlapping 10-second clips.

For each audio clip, we select pairs of versions with varying instrument compositions and generate a text instruction based on the instrument differences. We use the clips generated before to prepare the triplet data  $\langle$ text instruction, source music, target music $\rangle$  including remixing (1 Million), adding (0.3M) and replacement (0.3M), extracting (0.2M) and removing (0.2M) respectively. These music triplet data are referred to as the ‘in-house data’. We show our detailed data processing methods in Appendix.

**Evaluation Data** We evaluate the models on both in-domain data and out-domain data.

- In-domain data: We split the in-house data randomly into two parts and use one subset to generate triplet data for evaluating the models.
- Out-domain data: To demonstrate the robustness of the system, we also evaluate the models on the Synthesized Lakh (Slakh) Dataset (Manilow et al. 2019) which is a dataset of multi-track audio and has no overlap with the training data.

Dataset	Model	Task	FAD <sub>VGG</sub> ( $\downarrow$ )	Instruction Acc. ( $\uparrow$ )	Chord Rec. Acc. ( $\uparrow$ )	Pitch His. ( $\uparrow$ )	IO Interval ( $\uparrow$ )
In-house	AUDIT	Extract	1.67	0.39	0.82	0.62	0.54
		Remove	1.73	0.65	0.86	0.64	0.53
		Add	1.25	0.73	0.72	0.64	0.54
		Replace	1.50	0.62	0.83	0.63	0.51
		Avg.	1.54	0.60	0.81	0.63	0.53
	InstructME	Extract	1.54	0.56	0.86	0.69	0.68
		Remove	1.68	0.80	0.88	0.72	0.66
		Add	1.22	0.73	0.75	0.72	0.66
		Replace	1.39	0.62	0.86	0.71	0.67
		Avg.	<b>1.45</b>	<b>0.68</b>	<b>0.84</b>	<b>0.71</b>	<b>0.67</b>
Slakh	AUDIT	Extract	4.91	0.52	0.66	0.62	0.52
		Remove	1.92	0.57	0.57	0.61	0.51
		Add	3.11	0.87	0.58	0.63	0.47
		Replace	4.08	0.78	0.55	0.62	0.47
		Avg.	<b>3.50</b>	0.68	0.59	0.62	0.49
	InstructME	Extract	5.04	0.66	0.71	0.70	0.71
		Remove	1.87	0.79	0.65	0.70	0.69
		Add	3.15	0.87	0.66	0.74	0.67
		Replace	3.97	0.83	0.65	0.74	0.69
		Avg.	<b>3.50</b>	<b>0.79</b>	<b>0.67</b>	<b>0.72</b>	<b>0.69</b>

Table 1: **Objective Evaluation Results** of different edit tasks on In-house and Slakh datasets. Avg. is the average result of several edit tasks including extract, remove, add and replace. FAD reflects the music quality, Instruction Acc., and Chord Rec. Acc., pitch His. and IO Interval can measure the harmony of edited music.

## Evaluation Metric

Music is sounds that are artificially organized in relation to the sensational moments, with complex interplay and multi-layered perceptual impact between pitch, intensity, rhythm and timbre. Defining a single suitable metric to fully evaluate music is challenging, (Agostinelli et al. 2023; Huang et al. 2023a) focus evaluation on signal quality and semantics, whereas (Lv et al. 2023; Ren et al. 2020; Yang and Lerch 2020a) propose more direct evaluation approach based on musicality indicators, in order to achieve a more comprehensive evaluation of music, we proposed the following metrics to objectively evaluate the performance of edited music in three aspects:

**Music Quality.** We use the fréchet audio distance (FAD)<sup>1</sup>(Kilgour et al. 2019) to measure the quality between edited music and target music, the audio classification model is implemented with VGGish(Hershey et al. 2017).

**Text Relevance.** We define the instruction accuracy(IA) metric to indicate the relevance of the text-music pair, the proposed editing tasks are all related to music tags such as instrument, mood and genre, so we calculate instruction accuracy according to the edited music tags and input command while tags are recognized with tagging models which implemented with (Lu et al. 2021).

**Harmony.** We introduce three metrics for quantitative evaluation:

- *Chord Recognition Accuracy (CRA).* Chord Recognition Accuracy measures the harmony coherence between edited music and target music. We acquire the chord progression sequences of both source and target music in

the initial step, while the chord progression recognition model is implemented by (Cheuk et al. 2022). Then the alignment of these sequences is computed to determine the chord recognition accuracy.

- *Pitch Class Histogram (PCH)*<sup>2</sup>. The pitch class histogram is a pitch content representation that is octave-independent. It uses 12 classes to represent the chromatic scale, which spans from 4 to 40. We calculate the distribution of pitches classes according to this histogram.
- *Inter-Onset Interval (IOI)*<sup>2</sup>. Inter-onset interval refers to the time between two note onsets within a bar. In our case, We quantize the intervals into 32 classes and calculate the distribution of interval classes. Regarding PCH and IOI, we further compute the averaging Overlapped Area of their distributions to quantize the musical harmony in terms of pitch and onset aspects.

For objective evaluation, we generate 800 triplet data for each music editing task and evaluate them with these objective metrics.

## Results and Analysis

### Objective Evaluation Results

We compare against AUDIT (Wang et al. 2023) trained on the same data and in the same VAE latent space, as our baseline system in all experiments. As shown in Table 1, InstructME outperforms AUDIT in terms of music quality, text relevance and harmony. Specifically, operations such as extract and remove are tasks with definite answers, and these tasks emphasize the precision of generation. Alternatively,

<sup>1</sup><https://github.com/gudgud96/frechet-audio-distance>

<sup>2</sup><https://github.com/RichardYang40148/mgeval>

operations such as add, replace, and remix are tasks with indeterminate answers, and these tasks are more creatively oriented. These two types of tasks require the model to achieve stable and diverse output based on an accurate understanding of textual instructions. From Table 1, it is observed that InstructME improves musical quality by 5.84%, text relevance by 13.33% and harmony up to 26.42% compared to AUDIT on In-house dataset. These results demonstrate that our approach provides rich generated content in addition to capturing the difference between textual instructions.

Model	FAD	IA	CRA	PCH	IOI
AUDIT	0.49	0.69	0.59	0.60	0.46
InstructME	<b>0.45</b>	<b>0.73</b>	<b>0.70</b>	<b>0.72</b>	<b>0.64</b>

Table 2: Objective evaluation results of remixing on In-house dataset.

To study the generalization ability of InstructME, we also test it on the public available dataset Slakh (Manilow et al. 2019) which is more challenging in maintaining harmony by including unseen chord progressions. Table 1 shows that InstructME achieves comparable performance with AUDIT on musical quality but better results on text relevance and harmony. Particularly for chord recognition accuracy, our method surpasses the previous method by 13.56%.

As a novel and unique task proposed in this paper, we also evaluated the performance of InstructME and AUDIT on remix operation. As Table 2 indicates, compared to AUDIT, InstructME achieves a significant improvement in harmony related metrics(CRA by 18.6%, PCH by 20% and IOI by 39%), and also achieves better results in music quality and text relevance.

## Subjective Evaluation Results

Model	Length	Quality	Relevance	Harmony
AUDIT	10s	2.79	2.94	3.01
	30s	2.33	2.23	2.19
	60s	2.24	2.09	2.05
InstructME	10s	3.35	3.59	3.54
	30s	2.62	3.05	2.63
	60s	2.62	2.93	2.48

Table 3: Subjective results of editing on music of different duration.

We also conduct a subjective evaluation by outsourcing 10 testing samples(10s clip) per editing task for each labor. Mean Opinion Score(MOS) of scale 5 is used to compare the music quality, text relevance and harmony of two methods. To be more representative of real-world application scenarios, we also respectively generate 30-second and 60-second audio results, leveraging the chunk transformer’s insensitivity to output length. A subjective evaluation of these generated long-duration music was also performed. We wrap all

results in Table 3 and the MOS scores show the superiority of our method over the baseline.

## Case Study

**Chord Condition and Multi-Scale Aggregation** We perform ablation experiments to study the impact of the chord condition and multi-scale aggregation. The objective evaluation results of training InstructME without chord condition and multi-scale aggregation are listed in Table 4 respectively. The absence of chord conditioning is demonstrated to lead to a deterioration in harmony-related metrics such as CRA, PCH, and IOI. This observation underscores the critical role of the chord conditioning mechanism in holding the harmonicity of music editing. Moreover, the notable decline in FAD, subsequent to the deactivation of Multi-Scale Aggregation within the U-Net architecture, indicates its significant contribution to preserving the audio quality of music during the editing process.

Metric	InstructME	w/o Chord Condition	w/o MSA
FAD( $\downarrow$ )	1.45	1.46( $\uparrow 0.01$ )	1.53( $\uparrow 0.08$ )
IA ( $\uparrow$ )	0.68	0.63( $\downarrow 0.05$ )	0.66( $\downarrow 0.02$ )
CRA( $\uparrow$ )	0.84	0.81( $\downarrow 0.03$ )	0.83( $\downarrow 0.01$ )
PCH( $\uparrow$ )	0.71	0.64( $\downarrow 0.07$ )	0.72( $\uparrow 0.01$ )
IOI ( $\uparrow$ )	0.67	0.53( $\downarrow 0.14$ )	0.67( $\downarrow 0.00$ )

Table 4: Impact of chord condition and multi-scale aggregation strategies. MSA denotes multi-scale aggregation here. Mean value over all edit operations for each metric is provided.

**Stability and Diversity** In order to investigate the stability and diversity of generation within the InstructME framework, as depicted in Figure 4(a), we employ a spectral diagram to visually represent each editing task. Each row contains the source music, the ground truth music and three samples generated by our method. As mentioned in Section , different editing operations require different modeling capabilities. For example, the remix demands creativity because of different interpretations of the same sound source. However, the remove manipulation requires accuracy since the model should accurately recognize instruments mentioned in the textual instructions. For tasks requiring precision, our model consistently generates results congruent with the ground truth. For creativity-oriented tasks, the visual representation illustrates the diversity present in the spectrogram of the sampled music segments, underscoring the capacity of InstructME to conceive and construe novel compositions derived from existing audio sources.

**Consistency and Harmony** In Figure 4(b), we present an illustrative study to elucidate aspects of consistency and harmony. In the example, we take “Add acoustic guitar” as the textual instruction. The waveforms in Figure 4(b)(1) indicate that the beat timings before and after editing are meticulously synchronized, which demonstrates InstructME’s capacity to maintain temporal consistency through the editing process. The temporal consistency is also observed in the corresponding spectrograms in Figure 4(b)(2), which exhibit

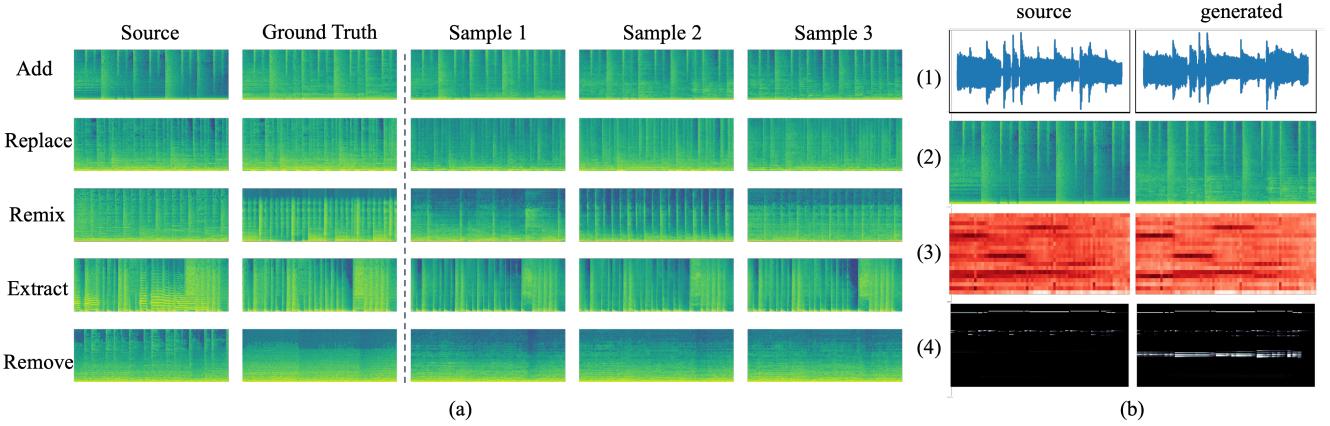


Figure 4: (a): Visualization of different editing tasks with three samples. All music segments are shown by spectrograms. (b): Comparison between source and edited music from four perspectives: (1) waveform (2) spectrogram (3) chord matrix (4) pitch matrix. The instruction command in the example is “add acoustic guitar”.

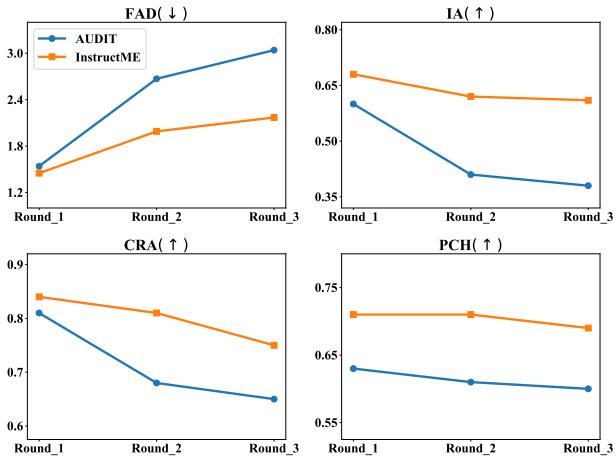


Figure 5: Line chart of objective evaluation results of three-round editing. Compared to AUDIT, InstructME shows a smaller decrease in metrics related to harmony and text relevance, while also exhibiting a smaller degradation in music quality(FAD).

energy spikes at identical times. Subsequently, we proceed to extract and compare the chord progression matrices from the source and generated musical segments as portrayed in Figure 4(b)(3). The intensity of the color is indicative of the predicted chord probability. Upon scrutinizing the probability patterns between the source and generated music, it is concluded that our method is able to preserve the musical harmony of the source. The last row in Figure 4(b) is the pitch matrix. In this representation, the uppermost pair of white stripes corresponds to the pitch pertaining to the piano and drum, respectively. Remarkably, the mere variations of these two lines between source and generated music are evidence of the consistency maintenance. The third stripe corresponds to the guitar’s pitch information. Furthermore, the

absence of other instruments from the generated music validates our model’s precise and controlled behavior.

In pursuit of a more comprehensive understanding of our model’s persistence in consistency and harmony, we undertake a series of multi-round editing experiments, comparing the musical outcomes of InstructME and AUDIT following repeated editing iterations. Results are listed in Figure 5. Notably, both our method and the baseline approaches display a gradual diminishment in performance as the iterative editing processes unfold. However, InstructME exhibits a comparatively marginal deterioration, particularly in terms of music harmony. Importantly, the remaining metrics remain well within an acceptable range, attesting to the robustness of the model’s performance.

## Conclusion

In this work, we introduce InstructME, a music editing and remixing framework based on latent diffusion models. For InstructME, we enhance the U-Net with multi-scale aggregation and chord condition to improve the harmony and consistency of edited music, and introduce chunk transformer to extend the long-term music generation capabilities. To evaluate the efficacy of music editing results, we establish several quantitative metrics and conduct experimental trials to validate them. Our findings indicate that the proposed InstructME outperforms the baselines in both subjective and objective experiments, which shows that our InstructME can effectively edit source music based on simple editing instructions, while preserving certain musical components and generating harmonious results that align with the semantic information conveyed in the instructions. In future endeavors, we aim to expand the scope and practical value of InstructME by exploring more intricate music editing tasks, such as structured editing.

## 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*.
- Cheuk, K. W.; Choi, K.; Kong, Q.; Li, B.; Won, M.; Hung, A.; Wang, J.-C.; and Herremans, D. 2022. Jointist: Joint learning for multi-instrument transcription and its applications. *arXiv preprint arXiv:2206.10805*.
- Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and Controllable Music Generation. *arXiv preprint arXiv:2306.05284*.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.
- Donahue, C.; Caillon, A.; Roberts, A.; Manilow, E.; Esling, P.; Agostinelli, A.; Verzetti, M.; Simon, I.; Pietquin, O.; Zeghidour, N.; et al. 2023. SingSong: Generating musical accompaniments from singing. *arXiv preprint arXiv:2301.12662*.
- Fagerjord, A. 2010. After convergence: YouTube and remix culture. *International handbook of internet research*, 187–200.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- 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 *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 131–135. IEEE.
- 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.
- Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Holz, D. 2023. Midjourney. Artificial Intelligence platform. Accessible at <https://www.midjourney.com>. <https://www.midjourney.com/>. Accessed: 2023-07-31.
- Huang, Q.; Jansen, A.; Lee, J.; Ganti, R.; Li, J. Y.; and Ellis, D. P. 2022. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*.
- 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. *arXiv preprint arXiv:2301.12661*.
- Jeong, J.; Kwon, M.; and Uh, Y. 2023. Training-free Style Transfer Emerges from h-space in Diffusion models. *arXiv preprint arXiv:2303.15403*.
- Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.
- Kilgour, K.; Zuluaga, M.; Roblek, D.; and Sharifi, M. 2019. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*, 2350–2354.
- Kwon, M.; Jeong, J.; and Uh, Y. 2022. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*.
- Lam, M. W.; Tian, Q.; Li, T.; Yin, Z.; Feng, S.; Tu, M.; Ji, Y.; Xia, R.; Ma, M.; Song, X.; et al. 2023. Efficient Neural Music Generation. *arXiv preprint arXiv:2305.15719*.
- Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023a. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.
- Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2023b. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 289–299.
- Lu, W.-T.; Wang, J.-C.; Won, M.; Choi, K.; and Song, X. 2021. SpecTNT: A time-frequency transformer for music audio. *arXiv preprint arXiv:2110.09127*.
- Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.
- Lv, A.; Tan, X.; Lu, P.; Ye, W.; Zhang, S.; Bian, J.; and Yan, R. 2023. GETMusic: Generating Any Music Tracks with a Unified Representation and Diffusion Framework. *arXiv preprint arXiv:2305.10841*.
- Manilow, E.; Wichern, G.; Seetharaman, P.; and Le Roux, J. 2019. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 45–49. IEEE.
- Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.
- Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
- Ren, Y.; He, J.; Tan, X.; Qin, T.; Zhao, Z.; and Liu, T.-Y. 2020. Popmag: Pop music accompaniment generation. In *Proceedings of the 28th ACM international conference on multimedia*, 1198–1206.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, 234–241. Springer.
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.
- Schneider, F.; Jin, Z.; and Schölkopf, B. 2023. Mo\^usai: Text-to-Music Generation with Long-Context Latent Diffusion. *arXiv preprint arXiv:2301.11757*.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, Y.; Ju, Z.; Tan, X.; He, L.; Wu, Z.; Bian, J.; and Zhao, S. 2023. AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models. *arXiv preprint arXiv:2304.00830*.
- Waysdorff, A. S. 2021. Remix in the age of ubiquitous remix. *Convergence*, 27(4): 1129–1144.
- Wierstorf, H.; Ward, D.; Mason, R.; Grais, E. M.; Hummersone, C.; and Plumbley, M. D. 2017. Perceptual evaluation of source separation for remixing music. In *Audio Engineering Society Convention 143*. Audio Engineering Society.
- Yang, H.; Firodiya, S.; Bryan, N. J.; and Kim, M. 2022. Don't Separate, Learn To Remix: End-To-End Neural Remixing With Joint Optimization. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 116–120. IEEE.
- Yang, L.-C.; and Lerch, A. 2020a. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9): 4773–4784.
- Yang, L.-C.; and Lerch, A. 2020b. Remixing music with visual conditioning. In *2020 IEEE International Symposium on Multimedia (ISM)*, 181–188. IEEE.

## Appendix

### Data

**Data Preprocess** For training the UniME, we collected 417 hours music audio files which contain different instrumental tracks, which is named in-house data. Then, we preprocess them by resampling their sample rate to 24khz and splitting them into non-overlapping 10-second clips with sliding window. For data cleaning,

- Normalize and unify the name of instruments.
- For each clip, combine the tracks with the same instrument. Eg: combine *acoustic guitar (left)* and *acoustic guitar (right)* into *acoustic guitar*.
- An energy-based Voice Activity Detection (VAD) is used to remove the clips with the most silence.

**Triplet Data Generation** Then we use the clips generated before to prepare the music-to-music triplet data including  $\{\text{instruction, source music, target music}\}$  for training the InstructME. As shown in Table 5, we only provide five different music edit tasks in this work. And the workflow of triplet generation is described in the following:

- **ReMix:** Firstly, we select one clip randomly from the database, mix all the instrument tracks and regard the mixed clip as source music. Then we obtain the rhythm and time step information of the source music. In the following, we retrieve another music clip which has the same rhythm as target music and aligns the source-target music pair with time step information. Finally, we generate the text command as the template listed in Table 5.
- **Add:** Similarly, we randomly choose one clip from the database in the first. And  $i$  ( $i \in [1, 2, 3, 4]$ ) instrument tracks of the clips will be selected and mixed to get the source music. Then, we can get the target music by selecting another instrument track and combining it with the source music.
- **Remove:** Removing task can be regarded as the reverse edit operation of adding, so we can generate the removing triplet data by reversing the pairs of adding data.
- **Extract:** For extracting task, we select one clip and choose one of the instrument tracks as target music. Then, we mix this track with some other instruments as the source music.
- **Replace:** To generate the replace data pairs, we choose two different instrument tracks and mix them with some other instruments.

**Text Instruction Templates** we provide five different music edit tasks in this work. And the workflow of triplet generation is described in Table 5, there are some specific examples of command:

- add distorted electric guitar
- add synthesizer
- add piccolo
- extract viola
- extract synth strings
- extract string section

- remove accordion
- remove string section
- remove clarinet
- replace distorted electric guitar with electric guitar
- replace guitar synth with synth pad
- replace flute with accordion
- remix with bass, piano, strings
- remix with guitar, bass, drums
- remix with drums, bass, guitar, piano
- remix to Rock genre
- remix to R&B genre
- remix to Jazz genre

Task	Text Command
Remix	Remix with {instrument/genre}
Add	Add {instrument}
Remove	Remove {instrument}
Extract	Extract {instrument}
Replace	Replace {instrument A} with {instrument B}

Table 5: The text command templates of each edit task.

### Metric Calculation

This section endeavors to expound upon the computation of evaluation metrics. As stated in the primary paper, a multitude of metrics have been chosen to evaluate edited music, primarily categorized into three dimensions: music quality, text relevance, and harmony. The corresponding metrics and computational methodologies are delineated as follows.

- **Fréchet Audio Distance (FAD).** FAD is defined as a measure of quality difference between a given audio and target audio (clean, high-quality audio). In contrast to extant audio evaluation metrics, FAD does not scrutinize individual audio clips. Rather, it compares embedding statistics derived from the complete evaluation set with those generated from a massive compilation of unaltered music. In general, FAD can be defined as:

$$F(\mathcal{N}_t, \mathcal{N}_e) = \|\mu_t - \mu_e\|^2 + \text{tr}(\Sigma_t + \Sigma_e - 2\sqrt{\Sigma_t \Sigma_e}) \quad (7)$$

where  $\mathcal{N}_e(\mu_e, \Sigma_e)$  is the evaluation set embeddings,  $\mathcal{N}_t(\mu_t, \Sigma_t)$  is the target set embeddings,  $\text{tr}$  is the trace of a matrix.

- **Instruction accuracy (IA).** As mentioned in the main paper, we calculate IA according to the tag difference between edited music and target music. When edited music has tag set as  $T_e$  and target music has tag set as  $T_t$ :

$$IA = \frac{1}{\max(|T_e|, |T_t|)} |T_e \cap T_t| \quad (8)$$

where  $|\cdot|$  denote the number of set.

Metrics	Task	1 point	2 points	3 points	4 points	5 points
Music Quality		The sound quality is very poor, basically it's all noise.	Mostly noise, some instruments can be heard, some instruments sound normal.	There is noise, but it does not affect the performance of the instruments, and most of the instruments sound normal.	There is almost no noise, and the timbre is normal.	There is no noise at all, and the sound of the instrument is very pure, like a professional recording studio.
Text Relevance	Editing for instruments	Nothing to do with the command, no instrument matches.	There are one or two instruments that match with command.	The more obvious instruments all match with command, but there are still at least two instruments that do not match.	Basically right, but missing one.	All correct.
	Editing for mood or genre	It doesn't match with command, and it's obvious that it's not right.	Not right, but at least not the opposite.	There is no right or wrong, the genre or mood is rather vague.	The direction is obvious, and it can basically be judged to be correct.	Direction is clear and correct.
Harmony	Atomic operation	It doesn't harmony with the source music, you can obviously hear it.	It's not very harmonious with the source music, but the onset is correct.	Although it doesn't harmonize with the source music, it doesn't affect hearing.	Basically the same as target.	Same or different from target, but the overall music is very nice and has a strong sense of harmony
	Remix	It doesn't match with command, and it's obvious that it's not right. It's incongruous with the vocal, you can obviously hear it.	It doesn't coordinate well with the vocals, only the alignment of the beat can be heard.	The beats are aligned, but also basically in harmony with the vocals.	The beats and chords are basically aligned, and it sounds like a song.	It is spliced together in complete harmony with the vocals and has a reinterpretation part.

Table 6: Quantitative benchmark and description of difference tasks.

- *Chord Recognition Accuracy (CRA)*. The chord accuracy is defined as:

$$CA = \frac{1}{N_t * N_c} \sum_{i=1}^{N_t} \sum_{j=1}^{N_c} I \left\{ C_{i,j} = C'_{i,j} \right\} \quad (9)$$

where  $N_t$  is the number of tracks,  $N_c$  is the number of chords,  $C_{i,j}$  and  $C'_{i,j}$  are the  $j$ -th chord in the  $i$ -th target track and edited track, respectively.

- *Overlapped Area of Pitch Class Histogram and Inter-Onset Interval*. We first obtain histograms of pitch (divided into 12 categories) and IOIs (divided into 32 categories) according to their respective classification numbers, and convert the histograms into probability density functions (PDFs) using Gaussian kernel density estimation. We then compute the feature overlapped area(OA) between the PDFs of edited music and target music, the

OA definition like:  $(\mathcal{D}_{OA})$

$$\mathcal{D}_{OA}^f = \frac{1}{N_t * N_b} \sum_{i=1}^{N_t} \sum_{j=1}^{N_b} OA(\mathcal{P}_{i,j}^f, \hat{\mathcal{P}}_{i,j}^f) \quad (10)$$

where  $f \in \{PCH, IOI\}$ ,  $N_t$  is the number of tracks,  $N_b$  is the number of bars, OA refers to the overlapping area of two distributions,  $\mathcal{P}_{i,j}^f$  and  $\hat{\mathcal{P}}_{i,j}^f$  are the distribution features in  $j$ -th bar and  $i$ -th track of edited music and target music.

### Subjective Evaluation

Every candidate participant in the subjective evaluation has undergone standard alignment training to ensure that all participants score under the same standard. The training benchmark is as Table 6 shows. Figure 6 shows the user interface presented to candidates.

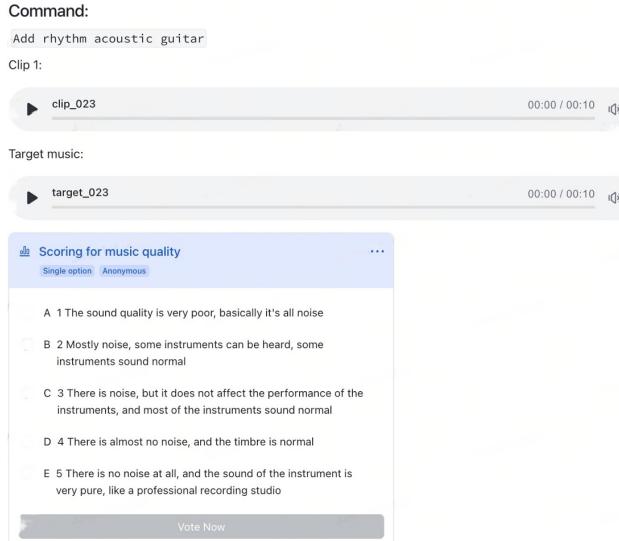


Figure 6: User interface of subjective evaluation.

## Model Details

**AutoEncoder** Unlike image and text modalities, audio often has extremely low information density and ultra long sequences. For example, in 24khz sample rate, music with 10 seconds contains 240,000 samples, which is approximately 234 images in 32x32. To lower the computational demands of training diffusion models towards efficient music editing, we employ an autoencoder to learn a low dimension latent space which is perceptually equivalent to the waveform space, but offers significantly reduced computational complexity, like latent diffusion (Rombach et al. 2022).

Our autoencoder in this work is built based on a Variational AutoEncoder (VAE) model which consists of an encoder  $\mathcal{E}$ , a decoder  $\mathcal{D}$  and a discriminator with 4 stacked ResNet-style (He et al. 2016) convolutional 1D blocks. Encoder  $\mathcal{E}$  compresses the waveform  $x \in \mathbb{R}^{T \times 1}$  into a lower dimension latent space  $z \in \mathbb{R}^{\frac{T}{r} \times C \times 1}$ , where  $r/C = 96/4 = 24$  can represent the compression ratio. Decoder  $\mathcal{D}$  can reconstructs the waveform  $\hat{x}$  from the latent space  $z$ . The discriminator is used in training process to distinguish between real and reconstructed waveform. Note that our autoencoder can reconstruct waveform directly, so there is no need for another vocoder like (Wang et al. 2023; Liu et al. 2023a). To train the VAE, we adopt a mean-square-error (MSE) based reconstruction loss  $\mathcal{L}_{rec}$ , an adversarial loss  $\mathcal{L}_{adv}$ , and a Kullback-Leibler loss  $\mathcal{L}_{kl}$  to regularize the latent representation  $z$ . Then, the total training loss of VAE can be formulated as:

$$\mathcal{L}_{VAE} = \mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{kl} \quad (11)$$

The detail configuration of autoencoder are listed in Table 7. We pretrain autoencoder on the dataset we generated before for 2M steps with a learning rate of 2e-4 and batch size 16. And the optimizer we used is Adam.

**Text Encoder** To encode the text commands, we employ T5-large (Raffel et al. 2020), as the text encoder  $\mathcal{T}$  of In-

AutoEncoder Configuration	
Number of Parameters	69.8M
In/Out Channels	1
Number of Down/Up Blocks	4
Cross Attention Dimension	1024
Down sample out channels	[16, 32, 64, 128]
Up sample out channels	[384, 192, 96, 48]
Down sample Rate	[4, 4, 3, 2]
Up sample Rate	[2, 3, 4, 4]

Table 7: Configuration of AutoEncoder

structME, similar to (Wang et al. 2023).  $\mathcal{T}$  is a pre-trained language model which is based on transformer and can convert text-based command into text embedding sequence for specifying the music editing task. It is important to note that the parameters of  $\mathcal{T}$  is pre-trained and frozen during the training and sampling process. The detail is shown in Table 8.

T5 Large Configuration	
Number of Parameters	737.7M
Output Channels	1024

Table 8: T5 Large

**U-Net** We adopt the U-Net backbone of StableDiffusion (Rombach et al. 2022) as the basic architecture for InstructME. And the detail configuration is shown in Table 9.

All diffusion models in the experiments are trained using 16 NVIDIA TESLA V100 32GB GPUs with a batch size of 1 per GPU for 200k steps. We optimize them with the AdamW optimizer and initial learning rate 1e-4.

Diffusion U-Net Configuration	
Number of Parameters	898.9M
In Channels	1
Out Channels	2
Learn Sigma Variance	True
Number of Down/Up Blocks	3
Number of ResBlocks	[4, 4, 4]
Latent channels	[512, 1024, 1024]
Down-Sample Rate	[2, 2, 2]
Up-Sample Rate	[2, 2, 2]
Number of Head Channels	64
Cross Attention Dimension	1024
Chunk Transformer Size	200
Diffusion Steps	1000
Noise Schedule	Linear
Sampling Strategy	DDIM
Sampling Steps	200

Table 9: Configuration of Diffusion U-Net