Prompt-Guided Zero-Shot Voice Cloning for Music Generation

Anonymous ACL submission

Abstract

Existing text-to-music (TTM) models rarely address voice cloning for singing, and most mainstream approaches do not support cross-domain voice cloning based on reference speech. However, speech-to-music voice transfer is highly valuable in practical applications, as speech data is easier to collect and provides more stable speaker representations.In this paper, we propose **S2M-Inject**, a cross-domain music generation framework that enables voice cloning based on reference speech. By injecting speaker representations extracted from speech into the music generation process, S2M-Inject produces music that preserves the target voice characteristics of the input speech. Experimental results demonstrate that S2M-Inject can effectively perform cross-domain voice cloning while maintaining reasonable music generation quality, and supports both Chinese and English music generation.

1 Introduction

In recent years, TTM models have achieved remarkable progress in controllable music generation (Copet et al., 2023; Liu et al., 2023; Gong et al., 2025; Chen et al., 2025a). However, voice cloning for singing remains a relatively underexplored problem, and most existing TTM approaches do not support cross-domain voice cloning based on reference speech (Copet et al., 2023; Liu et al., 2023; Gong et al., 2025; Chen et al., 2025a). Current models typically focus on controlling textual or musical attributes, such as style, emotion, or instrumentation, while exhibiting limited capability in modeling singer timbre, making precise and controllable timbre transfer difficult to achieve.

In contrast, speech-based voice cloning has been extensively studied (Wang et al., 2023; Du et al., 2024; Chen et al., 2025b), largely due to the large availability of speech data and the stability of speaker representations. From a practical perspective, speech-to-music voice transfer is highly desirable: speech data is easier to collect and provides reliable speaker identity information, making it a natural and scalable conditioning source for personalized music generation. Despite its potential, how to effectively incorporate speech-based voice characteristics into music generation remains insufficiently explored, as existing music generation models either rely on textual conditioning or music-intrinsic audio references, without explicitly modeling speech-derived timbre information (Copet et al., 2023; Liu et al., 2023; Agostinelli et al., 2023; Gong et al., 2025).

The lack of cross-domain voice cloning capability highlights a fundamental gap between speech generation and music generation. Although speech and music differ significantly in structure and expressive patterns, they share common acoustic factors related to speaker or singer timbre. Bridging this gap is crucial for enabling controllable singing generation driven by reference speech.

To address this challenge, we propose S2M-Inject, a cross-domain music generation framework that supports voice cloning based on reference speech. By injecting speaker representations extracted from speech into the music generation process, S2M-Inject enables music generation that preserves voice characteristics consistent with the reference speech. Experimental results demonstrate that S2M-Inject can effectively achieve cross-domain voice cloning while maintaining reasonable music generation quality, and supports both Chinese and English music generation.

Our main contributions are summarized as follows: (1) We study cross-domain voice cloning from speech to music, extending voice cloning beyond conventional speech synthesis; (2) We propose a fixed timbre injection strategy that integrates speech-based speaker representations into instruction-guided music generation; (3) Experiments show improved voice consistency with com-

petitive music generation quality.

2 Related Works

2.1 Zero-shot Voice Cloning

With the advancement of neural speech codecs and large-scale generative models, zero-shot voice cloning has achieved substantial progress in generation quality, stability, and generalization (Shen et al., 2023).

A prominent line of research is based on neural codec representations. VALL-E (Wang et al., 2023) represents speech using discrete codec tokens and adopts a cascaded autoregressive and non-autoregressive generation scheme, inheriting strong contextual modeling capability from language models while enabling high-quality zero-shot speech synthesis. NaturalSpeech 2 (Shen et al., 2023) instead employs continuous latent representations and introduces in-context learning into a diffusion-based acoustic modeling framework. Building upon this line of work, NaturalSpeech 3 (Shen et al., 2023) proposes a factorized diffusion-based zero-shot TTS system, improving naturalness through decoupled codec representations, although its codec prediction objectives remain constrained by identical textual content conditions.Another category of methods introduces intermediate semantic representations to bridge the modeling gap between text and acoustic features. SpearTTS (Kharitonov et al., 2023) and Make-a-Voice (Huang et al., 2023) leverage semantic tokens as an intermediate layer, enabling more robust zero-shot voice transfer across speakers and speaking styles. VoiceBox (Le et al., 2023) adopts a non-autoregressive flow-matching framework to perform speech infilling conditioned on audio context and text, demonstrating strong generation stability under partial conditioning. The Mega-TTS series (Jiang et al., 2023) employs mel-spectrograms as generation targets and explicitly disentangles timbre and prosody, modeling prosody in an autoregressive manner to improve timbre preservation. Similarly, P-Flow (Kim et al., 2023) applies flow-based models to zero-shot speech synthesis and achieves robust performance. In addition, some studies focus on improving inference efficiency and parallel generation capability. SoundStorm (Borsos et al., 2023b) and MobileSpeech (Ji et al., 2024) adopt masked non-autoregressive iterative generation strategies, achieving a favorable trade-off between synthesis speed and audio quality, which is important for large-scale deployment scenarios.

Despite the significant progress of zero-shot voice cloning in the speech domain, existing approaches are largely restricted to speech-to-speech scenarios. Most methods assume that both the reference and target signals belong to the speech domain, and cross-domain voice cloning from speech to music or singing remains largely unexplored. To address this limitation, we propose S2M-Inject, which investigates speech-based cross-domain voice cloning by injecting speaker representations extracted from speech into the music generation process, enabling consistent voice modeling from speech to music.

2.2 Music Generation

Early automatic music generation methods primarily relied on rule-based systems and symbolic representations, such as MIDI, combined with sequence models including RNNs, LSTMs, and Transformers (Hadjeres et al., 2017; Huang et al., 2018). While these approaches are relatively mature in modeling musical structure, they heavily depend on large-scale, well-annotated MIDI datasets and exhibit clear limitations in audio quality and realism (Huang et al., 2018), making them difficult to scale to diverse and complex musical scenarios.

With the advancement of deep learning and large-scale generative models, research has gradually shifted toward direct audio-level music generation. AudioLM (Borsos et al., 2023a) introduced the language modeling paradigm to audio generation, and MusicLM (Agostinelli et al., 2023) further incorporated pretrained models such as MuLan and w2v-BERT to achieve high-quality TTM synthesis. MusicGen (Copet et al., 2023) adopts a single-stage Transformer architecture with an efficient token interleaving strategy, striking a favorable balance among generation quality, efficiency, and textual controllability.

In parallel, diffusion-based approaches have demonstrated strong performance in music generation. Methods such as AudioLDM (Liu et al., 2023), Noise2Music(Huang et al., 2023), and MusicLDM (Liu et al., 2023) employ latent diffusion frameworks to enable text-guided high-fidelity music generation, and progressively extend to multi-track music and arrangement modeling.

Subsequent works, including JEN-1(Yao et al., 2025b) and Mustango(Melechovsky et al., 2024), introduce two-stage diffusion architectures or explicit musical priors (e.g., rhythm, harmony, and

tonality) to further enhance structural modeling. In recent studies, MusicGen, AudioLDM, and MusicLDM represent mainstream text-guided music generation approaches, while DiffRhythm+ (Chen et al., 2025a) and ACE-Step (Gong et al., 2025) further target long-form song generation and achieve notable progress in musical coherence and structure modeling. However, the conditioning mechanisms of these methods are still largely restricted to textual descriptions or music-intrinsic attributes, and they generally do not support cross-domain voice cloning based on reference speech (Copet et al., 2023; Liu et al., 2023; Chen et al., 2024; Gong et al., 2025).

Overall, music generation research has evolved from symbolic modeling to end-to-end audio generation frameworks that integrate language models and diffusion models. Nevertheless, existing approaches predominantly focus on single-modality conditions, such as text or audio, and remain limited in cross-modal or cross-domain conditioning. To address this gap, we propose S2M-Inject, which explores speech-conditioned cross-domain music generation by injecting voice representations extracted from speech into the music generation process, enabling consistent voice modeling from speech to music and providing a new perspective for multi-source conditioned music generation.

3 Method

3.1 Overview

As illustrated in Fig. 1, S2M-Inject is built upon an MM-DiT architecture, drawing inspiration from the design of MM-Audio (Cheng et al., 2025) and targeting the modeling of Voice Cloning in cross-domain scenarios. Specifically, the proposed framework enables singer timbre modeling through cross-domain injection of *speech-derived timbre representations* for music generation. The model consists of two core components: joint diffusion transformer layers and single diffusion transformer layers (Esser et al., 2024).

During training, the instruction encoder, mel encoder, and mel decoder are employed as pretrained modules and remain frozen, without participating in the optimization of the diffusion backbone. In the first joint diffusion transformer layer, the diffusion timestep is introduced, and the instruction embedding, text or lyric embedding, and speech-derived timbre embedding (Desplanques et al., 2020) are temporally concatenated as conditioning signals. Meanwhile, a noise-perturbed Mel-VAE latent is used as the audio modality input (Liu et al., 2023).

Unlike existing non-autoregressive text-to-audio architectures (Chen et al., 2025b; Zhu et al., 2025), S2M-Inject does not require explicit upsampling of textual representations to align with audio features, thereby simplifying the overall modeling pipeline.

For the TTM task, S2M-Inject leverages natural language instructions together with cross-domain injected speech timbre conditions to achieve multi-attribute controllable music generation, including singer timbre injection, melodic and emotional expression, music genre, and instrumentation control, while supporting both Chinese and English singing generation.

3.2 Model Architecture

S2M-Inject is built upon a MM-DiT backbone and integrates specially designed encoders to process heterogeneous input streams from different modalities.

Conditional Input Representation To enable conditional modeling for singing generation, we propose a unified conditional input representation that supports target audio generation under multiple heterogeneous constraints. The model inputs include natural language audio descriptions, lyric content, speaker timbre conditions, latent audio representations, and stochastic Gaussian noise. Lyric text is first converted into phoneme sequences using a Grapheme-to-Phoneme (G2P) model (Qiang et al., 2022), providing a unified content representation. The target audio waveform is encoded into a latent representation via a Mel-VAE, which is then linearly interpolated with Gaussian noise under a diffusion timestep sampled from $t \sim \mathcal{U}(0, 1)$ (Lipman et al., 2022), forming the noisy acoustic state required for diffusion modeling. This unified conditional representation provides a stable conditioning foundation for the subsequent generation process.

Semantic and Textual Encoders For semantic control, we feed structured audio descriptions generated by Gemini (Team et al., 2023) into Qwen3-Omni (Yang et al., 2025) and extract its dense hidden states as semantic embeddings $\mathbf{E}_{\text{inst}}$, which characterize the global semantic attributes and expressive properties of the target audio. For content modeling, following the design principles of ZipVoice (Zhu et al., 2025) and M3-TTS (Wang et al., 2025), we adopt Zipformer (Yao et al., 2023)

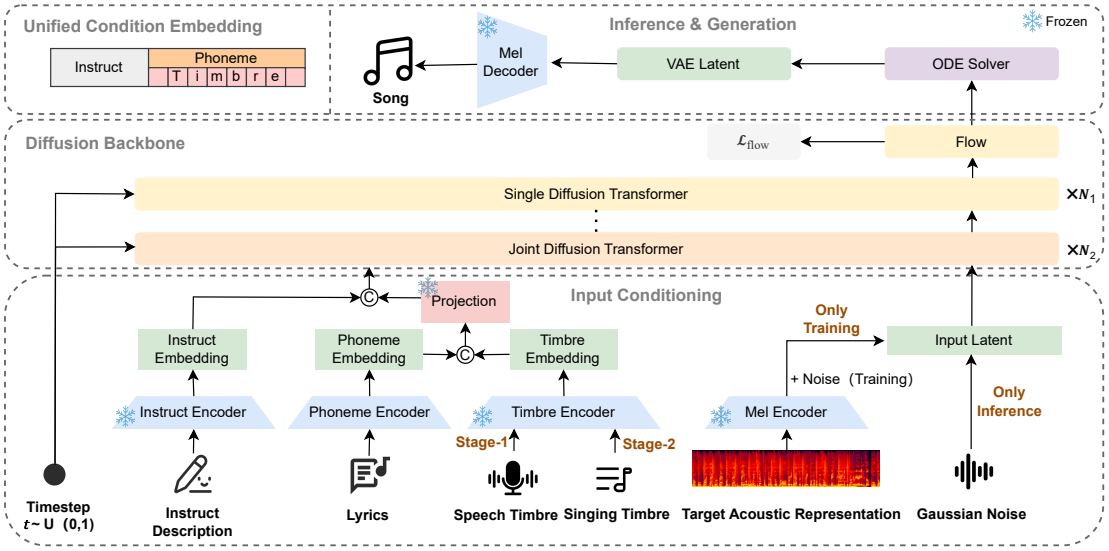

Figure 1: S2M-Inject is built upon a multimodal diffusion transformer (MM-DiT) for music generation, supporting multi-attribute control and voice cloning conditioned on natural language instructions and reference speech timbre. Audio is represented using continuous latent variables extracted from a pretrained Mel-VAE. During inference, the target music latent is obtained by solving an ordinary differential equation (ODE) and decoded into waveform audio.

as the textual encoder. The input text is tokenized using Byte Pair Encoding (BPE) (Sennrich et al., 2016) and then processed by the Zipformer to efficiently extract temporally aligned feature representations, enabling accurate modeling of singing content.

Acoustic Conditioning and Timbre Injection The proposed framework adopts a densely connected time-delay neural network (D-TDNN) (Yu and Li, 2020) to extract highly discriminative speaker timbre representations from short speech segments. To mitigate interference from expressive singing variations, we adopt a two-stage timbre conditioning strategy that uses singing-derived timbre information in Stage 1 and speech-derived timbre information from constructed speech–singing pairs in Stage 2, with data construction detailed in Section 3.3.

To enable cross-domain voice cloning, speaker timbre information is derived from the original speech, emotional attributes are neutralized using a predefined neutral embedding, and neutral speech for the target text is synthesized with IndexTTS2 as the timbre reference.

$$\mathbf{E}_{\text{spk}} \in \mathbb{R}^{B\times D}. \tag{1}$$

In the joint diffusion layers, the instruction embeddings and phoneme embeddings are used as textual conditioning inputs. Specifically, the instruction embeddings are denoted as

$$\mathbf{E}_{\text{inst}} \in \mathbb{R}^{B\times L_1\times D}, \tag{2}$$

and the phoneme embeddings are represented as

$$\mathbf{E}_{\text{phn}} \in \mathbb{R}^{B\times L_2\times D}. \tag{3}$$

Here, B denotes the batch size, D is the shared embedding dimension, and L_1 and L_2 represent the sequence lengths of instruction tokens and phoneme tokens, respectively.

The speaker timbre embedding $\mathbf{E}_{\text{spk}}$, derived from the stage-specific timbre condition e_{spk} (i.e., $e_{\text{spk}}^{\text{sing}}$ in Stage 1 and $e_{\text{spk}}^{\text{cross}}$ in Stage 2), is first projected and expanded along the temporal dimension to align with the phoneme sequence, and then concatenated with the phoneme embeddings along the feature dimension:

$$\tilde{\mathbf{E}}_{\text{phn}} = [\mathbf{E}_{\text{phn}} \parallel \mathbf{E}_{\text{spk}}] \in \mathbb{R}^{B\times L_2\times 2D}. \tag{4}$$

A linear projection is subsequently applied to restore the unified feature dimension:

$$\hat{\mathbf{E}}_{\text{phn}} \in \mathbb{R}^{B\times L_2\times D}. \tag{5}$$

Finally, the instruction embeddings and the fused phoneme–timbre representations are concatenated

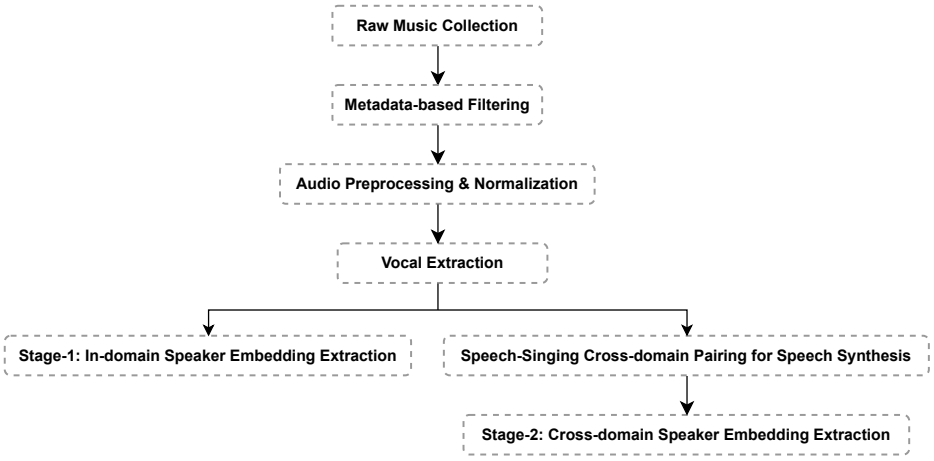

Figure 2: Overall Two-stage Training Data Construction Pipeline.

along the token (temporal) dimension to form the unified textual conditioning input:

$$\mathbf{C}_{\text{text}} = \left[\mathbf{E}_{\text{inst}} \parallel_T \hat{\mathbf{E}}_{\text{phn}}\right] \in \mathbb{R}^{B\times(L_1+L_2)\times D}. \quad (6)$$

Audio generation is performed in the latent space of a Mel-VAE, which compresses 44.1 kHz waveforms into low-dimensional continuous representations, effectively reducing computational cost while preserving essential acoustic characteristics.

MM-DiT Backbone Our generative model adopts a MM-DiT architecture, whose overall design is based on Stable Diffusion 3 (Esser et al., 2024) and trained with Conditional Flow Matching (CFM) (Lipman et al., 2022). The backbone consists of N_2 Joint Diffusion Transformer layers and N_1 Single Diffusion Transformer layers.

In the joint diffusion layers, a unified textual conditioning input and a latent acoustic state are used as the two interacting modalities. Following the timbre injection formulation in Sec. 3.2, the instruction embedding $\mathbf{E}_{\text{inst}}$ and the fused phoneme–timbre representation $\hat{\mathbf{E}}_{\text{phn}}$ are concatenated along the token (temporal) dimension to form $\mathbf{C}_{\text{text}}$.

For the audio modality, the latent acoustic state $\mathbf{x}_t$ is constructed via linear interpolation between Gaussian noise and the Mel-VAE latent under a diffusion timestep sampled from $t \sim \mathcal{U}(0, 1)$. The two modalities interact through joint attention, where queries, keys, and values from different modalities are concatenated and processed by scaled dot-product attention to enable cross-modal information fusion. The output preserves the input dimensionality and is subsequently split back into the corresponding modalities.

In the single diffusion transformer layers, the model operates solely on the latent acoustic state to further refine audio generation quality, where the joint attention mechanism degenerates into standard self-attention over the audio modality. During training, We define the CFM loss as

$$\mathcal{L}_{\text{flow}}(\theta) = \mathbb{E}\left\|v_\theta(t, \mathbf{C}_{\text{text}}, \mathbf{x}_t) - u(t, \mathbf{x}_t)\right\|^2. \quad (7)$$

where v_θ denotes the learned conditional velocity field, u is the target conditional vector field, and t is the randomly sampled timestep. During inference, the target Mel-VAE latent is obtained by solving the corresponding ODE and finally decoded into waveform audio.

3.3 Two-Stage Data Construction and Training Strategy

We construct a two-stage dataset and adopt a corresponding training strategy to enable stable speech-to-singing timbre injection. The data construction pipeline is illustrated in Fig. 2. We collect approximately 5,000 hours of raw music data and apply metadata-based filtering (e.g., duration, sampling rate, compression), resulting in about 4,000 hours of curated data (around 1M clips). All samples are processed with unified preprocessing and normalization, including ASR-based lyric transcription, natural language audio captions, and vocal extraction using Demucs (Défossez et al., 2019).

Timbre supervision is constructed in two stages. In Stage 1, the original singing vocals from 500k clips provide in-domain timbre information that is acoustically consistent with the target singing audio. In Stage 2, original speech recordings provide speaker timbre information, while emotional

Algorithm 1: Two-Stage Cross-Style Training Strategy

Input:
Stage-1 condition $C_{\text{text}}^{(1)} = [C_{\text{text}}, e_{\text{spk}}^{\text{sing}}]$;
Stage-2 condition $C_{\text{text}}^{(2)} = [C_{\text{text}}, e_{\text{spk}}^{\text{cross}}]$;
Stage transition step s_{stage1}
for *each training step* s **do**
  if $s \leq s_{\text{stage1}}$ **then**
    Stage 1: In-domain timbre injection;
    Use condition $C_{\text{text}}^{(1)}$;
    Optimize model with $\mathcal{L}_{\text{flow}}(x_t, t, C_{\text{text}}^{(1)})$;
  else
    Stage 2: Cross-style timbre injection;
    Use condition $C_{\text{text}}^{(2)}$;
    Optimize model with $\mathcal{L}_{\text{flow}}(x_t, t, C_{\text{text}}^{(2)})$;
  end if
end for

factors are removed by using a predefined neutral emotion embedding. Neutral speech corresponding to the target texts is synthesized using IndexTTS2 (Zhou et al., 2025), and the synthesized speech signals serve as cross-domain timbre conditions aligned with the target singing voice.

Based on the constructed data, we adopt a two-stage training strategy summarized in Algorithm 1. The model is first trained with in-domain singing timbre conditions and then fine-tuned with speech-derived timbre conditions, using the same CFM objective in both stages.

4 Experiments

4.1 Model Details and Datasets

Following the overall architectural design described in Sec. 3.2, the proposed S2M-Inject model contains 1.34 billion parameters, with a feed-forward network dimension of 1024. The model consists of 14 Joint Diffusion Transformer layers and 6 Single Diffusion Transformer layers, and employs Rotary Positional Embedding (RoPE) (Su et al., 2024) to enhance sequence modeling capability.

For textual and instruction modeling, we adopt a Zipformer-based phoneme encoder (Yao et al., 2023) with a feed-forward dimension of 512. The instruction encoder is instantiated using Qwen2.5-7B (Yang et al., 2025), which maps natural-language instructions into high-dimensional semantic representations. The mel encoder takes raw waveforms sampled at 44.1 kHz as input and produces latent embeddings at approximately 43 Hz, corresponding to a temporal downsampling factor of about $1024\times$ relative to the input sampling rate. This design significantly reduces computational cost while preserving sufficient representational capacity.

Model training is conducted on 32 NVIDIA Tesla A800 GPUs with 80 GB memory each, using a per-GPU batch size of 16. We employ the Adam optimizer (Kingma, 2014) with an initial learning rate of 1e-4, following common practice in diffusion-based audio generation models.

For music generation experiments, we collect and curate approximately 5,000 hours of music data from the Internet, and automatically generate aligned natural-language instruction descriptions and lyric annotations through an internal data processing pipeline. The instruction descriptions cover a wide range of musical and vocal attributes, including music genre, instrumentation, singer gender and age, rhythmic characteristics, and overall atmosphere. Audio clips range from 2 to 20 seconds in duration, and the dataset maintains balanced distributions in both language (Chinese/English) and singer gender at a 1:1 ratio. The dataset primarily consists of single-singer segments to reduce the interference of multi-voice overlap on timbre modeling, and covers diverse musical styles, accompaniment patterns, tempo variations, and accompaniment complexity. All audio samples are uniformly resampled to 44.1 kHz for both training and evaluation.

4.2 Comparisons and Evaluation Metrics

Most existing TTM models do not support cross-domain Voice Cloning based on reference speech, such as speech-to-singing timbre transfer. Even in music generation scenarios, models explicitly designed for singer Voice Cloning remain scarce, with the majority of approaches providing only coarse-grained singer attribute control rather than strict Voice Cloning.

To comprehensively evaluate the performance of S2M-Inject in both Voice Cloning and music generation, we conduct comparative experiments under Voice Cloning and TTM settings against representative baseline methods. All baseline models are deployed using their local inference implementations and evaluated on the same held-out test set

Table 1: Cross-domain voice cloning results evaluated by speaker similarity, WER, and subjective evaluation.

Model	Task	SIM ↑	WER ↓	SIM-MOS ↑
Ground Truth	–	1.00	0.00	–
CosyVoice2 (Du et al., 2024)	Speech→Speech	**0.68**	2.01	$\mathbf{3.75 \pm 0.13}$
Stage-1 S2M-Inject	Speech→Music	0.34	1.77	2.35 ± 0.22
Stage-2 S2M-Inject	Speech→Music	0.60	**1.56**	3.32 ± 0.34

Table 2: Music generation results evaluated by SongEval and subjective evaluation.

Model	Params	SongEval ↑					MOS ↑	
		Coh.	Mus.	Mem.	Clar.	Nat.	QMOS	MMOS
Ground Truth	–	3.60	3.52	3.56	3.40	3.34	–	–
DiffRhythm+ (Chen et al., 2025a)	1B	2.68	2.61	2.57	2.48	2.37	3.04 ± 0.46	2.79 ± 0.54
ACE-Step (Gong et al., 2025)	3B	2.89	2.87	2.83	2.77	2.71	$\mathbf{3.30 \pm 0.28}$	2.88 ± 0.20
InstructAudio (Qiang et al., 2025)	1.3B	**3.08**	2.98	3.00	2.89	**2.82**	2.82 ± 0.26	$\mathbf{2.91 \pm 0.35}$
Stage-1 S2M-Inject	1.3B	2.88	2.79	2.69	2.64	2.66	2.73 ± 0.36	2.78 ± 0.42
Stage-2 S2M-Inject	1.3B	3.06	**2.99**	**3.03**	**2.95**	2.77	2.92 ± 0.33	2.90 ± 0.37

consisting of 500 samples, ensuring fair and consistent comparison. Since DiffRhythm+ (Chen et al., 2025a) does not natively support generating music clips shorter than 90 seconds, we first generate full-length sequences and subsequently crop them to the target duration; this implementation difference may introduce minor bias in certain evaluation metrics.

For the Voice Cloning task, we compare S2M-Inject with CosyVoice2 (Du et al., 2024), and the results are summarized in Table 1. For music generation, we compare S2M-Inject with DiffRhythm+ (Chen et al., 2025a), ACE-Step (Gong et al., 2025), and InstructAudio (Qiang et al., 2025), with quantitative results summarized in Table 2.

We adopt a combination of objective and subjective evaluation metrics. For Voice Cloning, objective metrics include Speaker Similarity(SIM) and Word Error Rate (WER) (Anastassiou et al., 2024). For music generation, we employ the SongEval benchmark (Yao et al., 2025a), which evaluates Coherence, Musicality, Memorability, Clarity, and Naturalness. Subjective evaluations are conducted using QMOS, MMOS, and SIM-MOS.

4.3 Results and Analysis

Voice Cloning Performance Analysis Table 1 reports both objective and subjective evaluation results on the voice cloning task. We focus on analyzing the performance variation of S2M-Inject across the two training stages (Stage 1 and Stage 2).

From the SIM metric, we observe that Stage 1 achieves a similarity score of 0.34, while Stage 2 significantly improves this score to 0.60 after introducing the speech–music cross-pair based training strategy. This improvement indicates that incorporating reference speech timbre constraints in the second stage effectively enhances the model's ability to capture and transfer target speaker characteristics. In contrast, Stage 1, which relies solely on in-domain singing data, tends to depend more on music-domain statistics, resulting in limited performance for cross-domain voice alignment.

Compared with the speech-to-speech voice cloning model CosyVoice2, Stage 2 still exhibits a performance gap in terms of SIM. It is important to note that CosyVoice2 is specifically designed for in-domain speech voice cloning, whereas S2M-Inject targets the more challenging speech-to-music cross-domain voice cloning scenario. Given the fundamental differences in task difficulty and input conditions, this performance gap is within reasonable expectations.

In terms of intelligibility, Stage 2 achieves a lower WER of 1.56 compared to 1.77 for Stage 1, indicating that introducing cross-domain voice conditioning does not degrade linguistic content. Subjective evaluation results are consistent with these observations: Stage 2 attains a SIM-MOS score of

3.32, significantly higher than Stage 1, demonstrating improved voice similarity and overall perceptual quality.

Music Generation Quality Analysis Table 2 reports both objective and subjective evaluation results for music generation. We focus on whether S2M-Inject can maintain reasonable music generation quality while introducing cross-domain speech-to-singing timbre conditioning.

Comparing the two training stages, Stage 2 consistently outperforms Stage 1 across all SongEval metrics, with clear improvements in coherence, musicality, memorability, clarity, and naturalness. This indicates that the second-stage training successfully incorporates speech-derived timbre information without degrading musical structure or perceptual consistency, and in fact improves overall music generation quality under short-segment settings.These gains suggest that the introduced cross-domain timbre conditioning is well aligned with the underlying music generation objectives.

When compared with existing TTM models, S2M-Inject (Stage 2) achieves competitive SongEval performance. It outperforms DiffRhythm+ across all SongEval dimensions and is comparable to ACE-Step, while remaining slightly below InstructAudio on several metrics. We note that these comparisons are conducted on short audio clips ranging from 5 to 20 seconds, whereas models such as ACE-Step and DiffRhythm+ are primarily optimized for long-form music generation. Therefore, the reported results mainly reflect performance under short-segment generation scenarios.

In terms of subjective evaluation, S2M-Inject (Stage 2) shows consistent improvements over Stage 1 on both QMOS and MMOS, indicating that introducing cross-domain timbre conditioning does not negatively affect perceived audio quality. Although its QMOS remains lower than ACE-Step and InstructAudio, S2M-Inject achieves comparable MMOS, suggesting that the generated music remains musically acceptable while prioritizing cross-domain voice cloning capability. Overall, these results demonstrate that S2M-Inject preserves stable and competitive music generation quality while enabling speech-driven singing voice cloning.

5 Discussion

This work investigates speech-conditioned cross-domain voice cloning for music generation under a unified instruction-guided framework. Using speech as the timbre source provides stable and scalable identity cues, which contributes to the consistent improvement from Stage 1 to Stage 2; however, the intrinsic acoustic and expressive mismatch between speech and singing still limits the achievable upper bound of timbre similarity compared to in-domain speech-to-speech cloning. Moreover, while instruction-based conditioning enables unified control across speech and music generation, it inevitably compresses acoustic information and introduces one-to-many mappings, leading to averaged realizations and a remaining gap in perceptual quality relative to reference-audio-based methods. These results indicate that cross-domain timbre transfer remains challenging, especially when balancing controllability and perceptual quality. To ensure stable joint modeling, music generation is restricted to short segments, which facilitates fair cross-domain evaluation aligned with speech durations but does not fully reflect long-form music generation capability. Overall, S2M-Inject demonstrates the feasibility of speech-driven cross-domain voice cloning for music generation and serves as an initial step toward more flexible conditioning mechanisms across audio domains.

6 Conclusions and Future Work

In this paper, we propose S2M-Inject, a cross-domain voice cloning framework based on the MM-DiT architecture, which enables timbre injection and cloning from speech to music generation. Experimental results demonstrate that the proposed method can effectively achieve cross-domain timbre transfer, validating both the feasibility and effectiveness of the framework.The main contributions of this work can be summarized as follows. (1) We explore the problem of cross-domain voice cloning from speech to music, extending voice cloning beyond conventional speech synthesis scenarios to music generation. (2) We propose a fixed timbre injection strategy that integrates speech-based speaker representations into instruction-guided music generation. (3) Extensive experiments show that the proposed approach improves timbre consistency while maintaining reasonable music generation quality.For future work, we plan to further investigate long-form music generation, enhance cross-domain voice cloning capability, and extend voice cloning applications to a broader range of audio domains.

Limitations

Despite the promising results in unified speech and music generation, this work has several limitations.

First, our timbre control relies on reference speech signals, and cross-domain transfer from speech to singing/music may introduce a domain mismatch, which can affect timbre faithfulness and perceptual consistency in some cases.

Second, to accommodate the joint modeling of speech and music, music generation is restricted to short clips of 5–20 seconds, which limits the model's ability to capture long-term musical structure.

Finally, unified speech and music generation still lacks mature and comprehensive evaluation benchmarks, and existing metrics may not fully reflect long-term musical coherence or creative diversity. We also note that AI-based tools were used for proofreading and language polishing of the manuscript, while all technical content and conclusions were determined by the authors.

Ethical Statement

This work investigates instruction-guided speech and music generation with timbre control, which may introduce potential ethical risks. We outline the following considerations:

- **Voice misuse and impersonation.** The proposed method enables timbre-controlled generation, which could be misused for impersonation. This work is intended solely for research purposes and does not aim to replicate or impersonate any specific individual without authorization.

- **Data usage and privacy.** All training data are collected from publicly available or properly licensed sources and processed in accordance with applicable data usage and privacy policies.

- **Responsible deployment.** We encourage responsible use of the proposed method and emphasize that appropriate safeguards, consent mechanisms, and usage restrictions should be considered in real-world applications to prevent malicious or deceptive use.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023a. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

Huakang Chen, Yuepeng Jiang, Guobin Ma, Chunbo Hao, Shuai Wang, Jixun Yao, Ziqian Ning, Meng Meng, Jian Luan, and Lei Xie. 2025a. Diffrhythm+: Controllable and flexible full-length song generation with preference optimization. *arXiv preprint arXiv:2507.12890*.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1206–1210. IEEE.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025b. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. 2025. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28901–28911.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. 2019. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. 2025. Ace-step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*.

Gaëtan Hadjeres, François Pachet, and Frank Nielsen. 2017. Deepbach: a steerable model for bach chorales generation. In *International conference on machine learning*, pages 1362–1371. PMLR.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music transformer. *arXiv preprint arXiv:1809.04281*.

Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, and 1 others. 2023. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*.

Shengpeng Ji, Ziyue Jiang, Hanting Wang, Jialong Zuo, and Zhou Zhao. 2024. Mobilespeech: A fast and high-fidelity framework for mobile zero-shot text-to-speech. *arXiv preprint arXiv:2402.09378*.

Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, and 1 others. 2023. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2306.03509*.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11:1703–1718.

Sungwon Kim, Kevin Shih, Joao Felipe Santos, Evelina Bakhturina, Mikyas Desta, Rafael Valle, Sungroh Yoon, Bryan Catanzaro, and 1 others. 2023. P-flow: A fast and data-efficient zero-shot tts through speech prompting. *Advances in Neural Information Processing Systems*, 36:74213–74228.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and 1 others. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8293–8316.

Chunyu Qiang, Peng Yang, Hao Che, Jinba Xiao, Xiaorui Wang, and Zhongyuan Wang. 2022. Back-translation-style data augmentation for mandarin chinese polyphone disambiguation. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1915–1919. IEEE.

Chunyu Qiang, Kang Yin, Xiaopeng Wang, Yuzhe Liang, Jiahui Zhao, Ruibo Fu, Tianrui Wang, Cheng Gong, Chen Zhang, Longbiao Wang, and 1 others. 2025. Instructaudio: Unified speech and music generation with natural language instruction. *arXiv preprint arXiv:2511.18487*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1715–1725.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Xiaopeng Wang, Chunyu Qiang, Ruibo Fu, Zhengqi Wen, Xuefei Liu, Yukun Liu, Yuzhe Liang, Kang Yin, Yuankun Xie, Heng Xie, and 1 others. 2025. M3-tts: Multi-modal dit alignment & mel-latent for zero-shot high-fidelity speech synthesis. *arXiv preprint arXiv:2512.04720*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, and 1 others. 2025a. Songeval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*.

Yao Yao, Peike Li, Boyu Chen, and Alex Wang. 2025b. Jen-1 composer: A unified framework for high-fidelity multi-track music generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14459–14467.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2023. Zipformer: A faster and better encoder for automatic speech recognition. *arXiv preprint arXiv:2310.11230*.

Ya-Qi Yu and Wu-Jun Li. 2020. Densely connected time delay neural network for speaker verification. In *Interspeech*, pages 921–925.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. 2025. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. *arXiv preprint arXiv:2506.13053*.