

# S2M-Inject: Prompt-Guided Zero-Shot Voice Cloning for Music Generation

Teng Ma<sup>1,2,\*</sup> Chunyu Qiang<sup>2,\*,†</sup> Zhongliang Liu<sup>2</sup> Xiaopeng Wang<sup>2</sup> Kang Yin<sup>2</sup>  
Ziyu Zhang<sup>2</sup> Yuzhe Liang<sup>2</sup> Yushen Chen<sup>2</sup> Yuxin Guo<sup>2</sup> Feiyu Chen<sup>1,†</sup>

<sup>1</sup>National Center for Applied Mathematics in Chongqing, Chongqing Normal University

<sup>2</sup>Kling AI

qiangchunyu@tju.edu.cn fchen\_cqnu@163.com

## Abstract

Despite advancements in controllable text-to-music (TTM) generation, existing models struggle with cross-domain voice cloning for singing due to inadequate modeling of speech-derived vocal characteristics, inherently limiting precise timbre transfer. To achieve precise and controllable music generation conditioned on the timbre of reference speech, we propose **S2M-Inject**, a cross-domain framework driven by reference speech. Incorporating a fixed timbre embedding strategy, S2M-Inject injects speaker representations into an instruction-guided generation pipeline. This effectively preserves target vocal attributes while enabling comprehensive natural-language control over lyrics, genre, instrumentation, rhythm, melody, and mood. Extensive experiments demonstrate that the framework achieves effective cross-domain voice cloning, significantly enhancing vocal similarity while maintaining high-fidelity audio quality in both Chinese and English tasks. This work extends voice cloning into the musical domain, offering a rigorous paradigm for personalized music generation. Audio samples are available at <https://anonymous.4open.science/w/S2M-Inject-975>.

## 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.

From a practical perspective, music generation conditioned on reference speech is highly valuable, since speech data is easier to collect and can provide stable and reliable speaker identity information, making it a natural conditioning source for personalized music generation. However, existing music generation models mainly rely on textual instructions or music-domain audio references for control, while lacking the ability to explicitly model and preserve speaker timbre characteristics from reference speech (Copet et al., 2023; Liu et al., 2023; Agostinelli et al., 2023; Gong et al., 2025).

In contrast, speech-based voice cloning has been extensively studied (Wang et al., 2023a; Du et al., 2024; Chen et al., 2025b), largely benefiting from the availability of large-scale speech data and the stability of speaker representations. Despite the potential of speech-conditioned music generation, directly transferring speaker representations extracted from speech into music generation remains highly challenging due to the significant cross-domain gap between speech and music. Compared with ordinary speech, singing and music exhibit substantially different pitch distributions, rhythmic structures, and expressive dynamics, making speech-derived speaker representations difficult to generalize to the music domain. In addition, existing TTM models mainly focus on semantic, stylistic, or compositional control rather than speaker identity preservation, which makes controllable cross-domain voice cloning particularly difficult. Effectively bridging the domain gap between speech and music is therefore a key challenge for enabling controllable singing generation driven by reference speech.

To address these challenges, we propose **S2M-Inject**, a cross-domain music generation framework that supports voice cloning based on reference speech. By injecting speaker representations

\*Equal contribution. Work done during internship at Kling AI. †Corresponding author.

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 contributions are summarized as follows:

- We study cross-domain voice cloning from speech to music, extending voice cloning beyond conventional speech synthesis;
- We propose a fixed timbre injection strategy that integrates speaker representations into instruction-guided music generation;
- Experiments demonstrate that S2M-Inject improves voice consistency while maintaining competitive music generation quality, indicating its effectiveness for generation.

## 2 Method

### 2.1 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-VAE encoder, and Mel-VAE decoder are employed as pretrained modules and remain frozen, without participating in the optimization of the diffusion backbone. The instruction embedding and the lyric representation fused with speaker timbre information are used as conditioning signals, while 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.

### 2.2 Model Architecture

S2M-Inject is built upon a MM-DiT backbone and integrates heterogeneous inputs from natural-language instructions, lyrics, reference timbre, and latent audio representations.

**Overview** S2M-Inject models three types of conditional information together with an audio latent stream. The first textual condition is the natural-language instruction, which describes global musical attributes such as genre, instrumentation, rhythm, melody, and mood. The second textual condition is the lyric text, which provides the linguistic content to be sung. The timbre condition is provided by a speaker embedding extracted from reference speech or clean vocal signals, and is used to control the target vocal identity.

On the audio side, the target waveform is encoded into a continuous latent representation by a pretrained Mel-VAE. During diffusion training, this latent is linearly interpolated with Gaussian noise to form the noisy acoustic state used by the audio generation branch. The instruction condition, the lyric condition fused with timbre information, and the noisy audio latent are then jointly modeled by the MM-DiT backbone. The model finally predicts the target Mel-VAE latent, which is decoded into waveform audio by the Mel-VAE decoder.

**Instruction and Lyric Encoding** For instruction-level control, we feed structured audio descriptions generated by Gemini (Team et al., 2023) into a frozen Qwen2.5-7B (Yang et al., 2024) and extract dense hidden states as instruction embeddings. These descriptions contain global musical attributes such as genre, instrumentation, rhythm, melody, and mood, providing conditioning signals for the overall musical style and expressive properties through natural language.

For lyric modeling, the lyric text is first converted into phoneme sequences using a Grapheme-to-Phoneme (G2P) model (Qiang et al., 2022). Following the design principles of ZipVoice (Zhu et al., 2025) and M3-TTS (Wang et al., 2025), we adopt Zipformer (Yao et al., 2023) as the textual encoder.

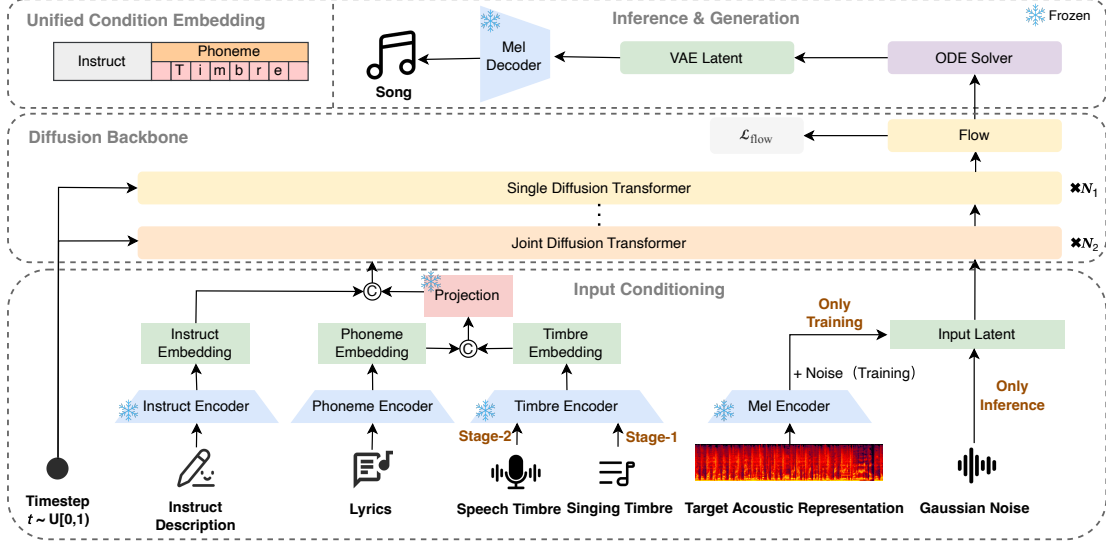


Figure 1: S2M-Inject is built upon a multimodal diffusion transformer (MM-DiT) for music generation, supporting multi-attribute control and voice cloning conditioned on natural language instructions and reference speech timbre. Audio is represented using continuous latent variables extracted from a pretrained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

The input text is tokenized using Byte Pair Encoding (BPE) (Sennrich et al., 2016) and then processed by Zipformer to obtain lyric/text embeddings, which provide singing content and pronunciation constraints for generation.

**Speech Timbre Conditioning** To enable reference-based cross-domain voice cloning, S2M-Inject uses different timbre sources during training and inference. During training, speaker embeddings are extracted from either original clean singing vocals or constructed speech counterparts and used as sample-level timbre conditions. During inference, the model directly extracts the speaker embedding from user-provided reference speech, enabling speech-conditioned timbre control for music generation. In practice, we use a speaker encoder based on CAM++ (Wang et al., 2023b) to extract a 192-dimensional global speaker embedding from short audio segments.

The two-stage training strategy is designed to make this timbre condition usable across domains. In Stage 1, the model uses in-domain clean singing vocals as the timbre source, allowing it to learn the correspondence among instruction conditions, lyric content, speaker timbre, and the target Mel-VAE latent in the music domain. This stage establishes basic music generation ability and encourages the model to associate different condition components with their roles in the generated audio: instruc-

tions control global musical attributes, lyrics constrain singing content, and speaker embeddings provide vocal timbre information. In Stage 2, speech-derived timbre conditions are introduced through constructed speech–singing pairs, encouraging the model to transfer speaker identity information from the speech domain to music generation and reducing the domain mismatch between reference speech and generated vocals.

**Conditional Fusion** After obtaining instruction embeddings, lyric embeddings, and speaker embeddings, S2M-Inject organizes them into two conditional streams. The instruction stream is formed by projecting the Qwen2.5-7B instruction embeddings into the MM-DiT hidden dimension:

$$\mathbf{E}_{\text{inst}} \in \mathbb{R}^{B \times L_1 \times D}, \quad (1)$$

where  $B$  is the batch size,  $L_1$  is the instruction length, and  $D$  is the MM-DiT hidden dimension.

The lyric–timbre stream is formed by injecting the speaker embedding into the lyric sequence. The Zipformer lyric embeddings are denoted as

$$\mathbf{E}_{\text{text}} \in \mathbb{R}^{B \times L_2 \times D_t}, \quad (2)$$

where  $L_2$  is the lyric sequence length and  $D_t$  is the text encoder dimension. The speaker embedding is a global vector:

$$\mathbf{e}_{\text{spk}} \in \mathbb{R}^{B \times 1 \times D_s}, \quad D_s = 192. \quad (3)$$

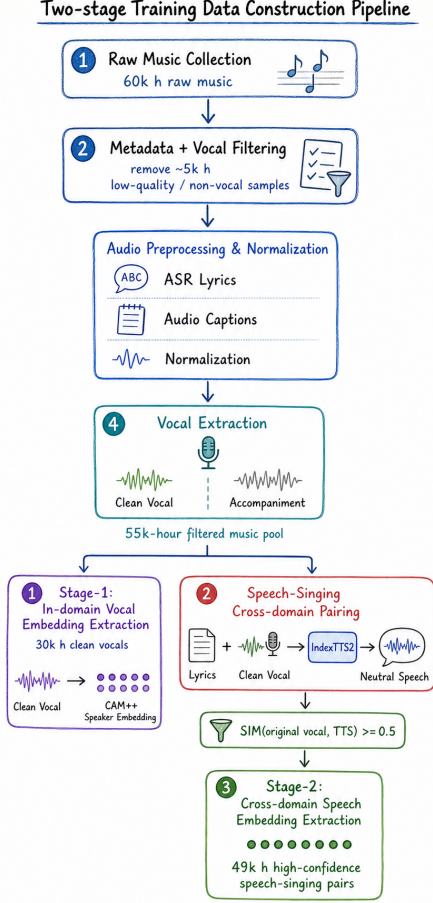


Figure 2: Overall Two-stage Training Data Construction Pipeline.

It is repeated along the temporal dimension to match the lyric length:

$$\mathbf{E}_{\text{spk}} = \text{Repeat}_{L_2}(\mathbf{e}_{\text{spk}}) \in \mathbb{R}^{B \times L_2 \times D_s}. \quad (4)$$

The repeated speaker embedding and the lyric embedding are then concatenated along the feature dimension:

$$\tilde{\mathbf{E}}_{\text{text}} = [\mathbf{E}_{\text{text}} \parallel \mathbf{E}_{\text{spk}}] \in \mathbb{R}^{B \times L_2 \times (D_t + D_s)}. \quad (5)$$

A trainable linear projection maps the concatenated representation to the unified MM-DiT hidden dimension:

$$\hat{\mathbf{E}}_{\text{text}} = \text{Linear}(\tilde{\mathbf{E}}_{\text{text}}) \in \mathbb{R}^{B \times L_2 \times D}. \quad (6)$$

This projection layer is not frozen; it is trained jointly with the diffusion backbone and initialized using the standard Xavier uniform initialization for linear layers.

Finally, the instruction stream and the lyric-timbre stream are concatenated along the token dimension to form the final conditional input:

$$\mathbf{C}_{\text{text}} = [\mathbf{E}_{\text{inst}} \parallel_T \hat{\mathbf{E}}_{\text{text}}] \in \mathbb{R}^{B \times (L_1 + L_2) \times D}. \quad (7)$$

**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.

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.

In the joint diffusion layers, the conditional input  $\mathbf{C}_{\text{text}}$  and a latent acoustic state are used as the two interacting modalities.

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

$$\mathcal{L}_{\text{flow}}(\theta) = \mathbb{E} \|v_\theta(t, \mathbf{C}_{\text{text}}, \mathbf{x}_t) - u(t, \mathbf{x}_t)\|^2, \quad (8)$$

where  $v_\theta$  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.

### 2.3 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 first collect approximately 60k hours of raw music data from the Internet. After metadata- and vocal-presence-based filtering, approximately 5k hours of low-quality or non-vocal samples are removed, resulting in a 55k-hour filtered music pool. All retained samples are

---

**Algorithm 1:** Two-Stage Cross-Domain Timbre Training Strategy

---

**Input:**  
Stage-1 speaker embedding  $e_{\text{spk}}^{\text{sing}}$ ;  
Stage-2 speaker embedding  $e_{\text{spk}}^{\text{speech}}$ ;  
Instruction embedding  $E_{\text{inst}}$  and lyric embedding  $E_{\text{text}}$ ;  
Stage transition step  $s_{\text{stage1}}$   
**for each training step  $s$  do**  
    **if  $s \leq s_{\text{stage1}}$  then**  
        Stage 1: In-domain timbre injection;  
        Construct  $C_{\text{text}}^{(1)}$  using  $e_{\text{spk}}^{\text{sing}}$  according to the conditional fusion formulation;  
        Optimize model with  $\mathcal{L}_{\text{flow}}(\mathbf{x}_t, t, C_{\text{text}}^{(1)})$ ;  
    **else**  
        Stage 2: Cross-domain timbre injection;  
        Construct  $C_{\text{text}}^{(2)}$  using  $e_{\text{spk}}^{\text{speech}}$  according to the conditional fusion formulation;  
        Optimize model with  $\mathcal{L}_{\text{flow}}(\mathbf{x}_t, t, C_{\text{text}}^{(2)})$ ;  
    **end if**  
**end for**

---

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, we select 30k hours from the filtered music pool and perform vocal separation to obtain clean singing vocals. CAM++ (Wang et al., 2023b) is then used to extract speaker embeddings from the clean vocals, which serve as in-domain timbre conditions for initial music-domain learning. In Stage 2, we construct speech–singing cross-domain pairs from the full 55k-hour filtered pool. Specifically, IndexTTS2 (Zhou et al., 2025) is used to synthesize neutral speech according to the corresponding lyrics and target timbre. We then compute the speaker similarity between the synthesized speech and the original clean vocal using CAM++. Pairs with speaker similarity below 0.5 are discarded, leaving approximately 49k hours of high-confidence speech–singing pairs for cross-domain timbre training. The synthesized neutral speech is used as a training-time bridge for cross-domain adaptation, while the evaluation reference speech clips are constructed from held-out identities and are not included in the training pairs.

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.

### 3 Experiments

#### 3.1 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., 2024), 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 and Ba) with an initial learning rate of  $1e-4$ , following common practice in diffusion-based audio generation models.

For music generation experiments, we collect approximately 60k hours of raw music data from the Internet and curate a 55k-hour filtered music pool after removing low-quality or non-vocal samples. We 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, 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 train-

Table 1: Cross-domain voice cloning results evaluated by speaker similarity, WER, and subjective evaluation. Dashes indicate unavailable metrics under the corresponding evaluation setting.

| Model                        | Task                      | SIM $\uparrow$ | WER (%) $\downarrow$ | SIM-MOS $\uparrow$              |
|------------------------------|---------------------------|----------------|----------------------|---------------------------------|
| Ground Truth                 | –                         | 1.00           | 0.00                 | –                               |
| CosyVoice2 (Du et al., 2024) | Speech→Speech             | <b>0.68</b>    | 2.01                 | <b>3.75<math>\pm</math>0.13</b> |
| HQ-SVC (Bai et al., 2026)    | Vocal SVC                 | 0.53           | 2.24                 | 3.12 $\pm$ 0.21                 |
| NeuCoSVC (Sha et al., 2024)  | Vocal SVC                 | 0.44           | 1.89                 | 3.05 $\pm$ 0.18                 |
| Vevo2 (Zhang et al., 2025)   | Speech/Singing Generation | 0.58           | 1.63                 | 3.24 $\pm$ 0.20                 |
| <b>S2M-Inject</b>            | Speech→Music              | 0.60           | <b>1.56</b>          | 3.32 $\pm$ 0.34                 |

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

| Model                              | Params | SongEval $\uparrow$ |             |             |             |             | MOS $\uparrow$                    |                                   |
|------------------------------------|--------|---------------------|-------------|-------------|-------------|-------------|-----------------------------------|-----------------------------------|
|                                    |        | 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 $\pm$ 0.46                   | 2.79 $\pm$ 0.54                   |
| ACE-Step (Gong et al., 2025)       | 3B     | 2.89                | 2.87        | 2.83        | 2.77        | 2.71        | <b>3.30 <math>\pm</math> 0.28</b> | 2.88 $\pm$ 0.20                   |
| InstructAudio (Qiang et al., 2025) | 1.3B   | <b>3.08</b>         | 2.98        | 3.00        | 2.89        | <b>2.82</b> | 2.82 $\pm$ 0.26                   | <b>2.91 <math>\pm</math> 0.35</b> |
| <b>S2M-Inject</b>                  | 1.3B   | 3.06                | <b>2.99</b> | <b>3.03</b> | <b>2.95</b> | 2.77        | 2.92 $\pm$ 0.33                   | 2.90 $\pm$ 0.37                   |

ing and evaluation. For evaluation, we construct a 500-sample held-out test set that is disjoint from the training data at both audio and speaker levels. Each test sample pairs a held-out reference speech clip with a target music clip from the same speaker identity, enabling zero-shot speech-to-music voice cloning evaluation on unseen speaker conditions.

### 3.2 Comparisons and Evaluation Metrics

Many existing TTM models are not explicitly designed for 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 relatively scarce, with many approaches focusing on coarse-grained singer attributes rather than strict Voice Cloning.

Since no publicly available baseline exactly matches the reference-speech-conditioned full music generation setting, we include representative task-adjacent systems from speech voice cloning, vocal SVC, speech/singing generation, and TTM as reference comparisons. HQ-SVC (Bai et al., 2026) and NeuCoSVC (Sha et al., 2024) are included as vocal-level SVC baselines, since they focus on singing voice timbre conversion rather than full music generation. Vevo2 (Zhang et al., 2025) is included as a closely related speech and singing voice generation system; however, it is not directly comparable under our instruction-guided speech-to-music setting. Vevo2 unifies control-

lable speech and singing voice generation through prosody modeling. In contrast, S2M-Inject targets full music generation with accompaniment and instruction-level controllability conditioned on reference speech. 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.

Since no publicly available baseline exactly matches the reference-speech-conditioned full music generation setting, we include representative task-adjacent systems from speech voice cloning, vocal SVC, speech/singing generation, and TTM as reference comparisons.

For the Voice Cloning task, we compare S2M-Inject with CosyVoice2 (Du et al., 2024), HQ-SVC (Bai et al., 2026), NeuCoSVC (Sha et al., 2024), and Vevo2 (Zhang et al., 2025), 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 further report the ablation results of the two-stage training strategy in Table 3.

We adopt a combination of objective and subjective evaluation metrics. For Voice Cloning, ob-

Table 3: Ablation study on the two-stage training strategy.

| Model Variant              | SIM $\uparrow$ | WER (%) $\downarrow$ | SIM-MOS $\uparrow$                | QMOS $\uparrow$                   | MMOS $\uparrow$                   |
|----------------------------|----------------|----------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| S2M-Inject w/ Stage 1 only | 0.34           | 1.77                 | $2.35 \pm 0.22$                   | $2.73 \pm 0.36$                   | $2.78 \pm 0.42$                   |
| S2M-Inject w/ Stage 1+2    | <b>0.60</b>    | <b>1.56</b>          | <b><math>3.32 \pm 0.34</math></b> | <b><math>2.92 \pm 0.33</math></b> | <b><math>2.90 \pm 0.37</math></b> |

jective metrics include Speaker Similarity(SIM) and Word Error Rate (WER) (Anastassiou et al., 2024). SIM is computed as the cosine similarity between speaker-verification embeddings extracted from the reference speech and the generated vocal signal. For generated music, the vocal signal is first extracted using the same vocal separation pipeline as in data preprocessing. WER is computed between the recognized transcription of the generated vocal or speech signal and the target lyric/text. 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. We invite 10 professional audio evaluators, each rating 20 audio samples on a five-point scale for audio quality (QMOS), musicality (MMOS), and perceived voice similarity to the reference speech (SIM-MOS), and report the mean and standard deviation across all ratings. To further evaluate instruction adherence, we caption the generated audio using Qwen3-Omni and automatically match five target attributes between the generated caption and the original instruction: genre, instrumentation, rhythm, melody, and mood.

### 3.3 Results and Analysis

**Voice Cloning Performance Analysis** Table 1 reports both objective and subjective evaluation results on the voice cloning task.

From the SIM metric, S2M-Inject achieves a similarity score of 0.60 under the speech-to-music setting, indicating that the proposed cross-domain timbre injection strategy can effectively capture and transfer target speaker characteristics from reference speech.

Compared with vocal-level SVC baselines, S2M-Inject achieves higher speaker similarity while performing instruction-guided full-music generation, demonstrating effective speech-conditioned timbre injection in a more challenging setting.

Compared with Vevo2, S2M-Inject obtains slightly higher SIM and lower WER, suggesting competitive timbre preservation and intelligibility

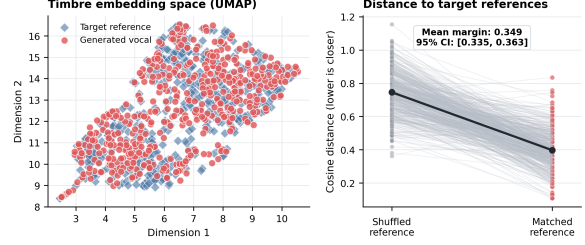


Figure 3: Timbre embedding analysis on the test set. Left: UMAP visualization of speaker embeddings extracted from reference speech and generated dry vocals. Right: cosine-distance comparison between generated dry vocals and matched or randomly shuffled reference targets.

while additionally generating accompaniment.

Although S2M-Inject still lags behind the speech-to-speech voice cloning model CosyVoice2 in SIM, this gap is expected because CosyVoice2 is specifically designed for in-domain speech cloning, whereas S2M-Inject addresses cross-domain speech-to-music voice cloning.

In terms of intelligibility, S2M-Inject achieves a WER of 1.56%, indicating that speech-derived timbre conditioning does not substantially interfere with lyric intelligibility. Subjective results are consistent with this observation: S2M-Inject attains a SIM-MOS score of 3.32, showing that listeners can perceive the transferred vocal identity in the generated music.

Figure 3 further supports the quantitative SIM results from the embedding-space perspective. The generated dry vocals are consistently closer to the corresponding reference timbres than to mismatched targets, indicating effective speech-conditioned timbre transfer. This matched-versus-shuffled comparison suggests that the learned timbre space captures speaker-specific cues rather than only generic vocal characteristics.

**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.

Table 4: Instruction-adherence accuracy of S2M-Inject.

| Musical Attribute | Accuracy |
|-------------------|----------|
| Genre             | 78.80%   |
| Instrumentation   | 80.00%   |
| Rhythm            | 94.40%   |
| Melody            | 81.60%   |
| Mood              | 87.60%   |

S2M-Inject achieves competitive results across SongEval metrics, especially in musicality, memorability, and clarity, indicating that speech-derived timbre conditioning can be incorporated without degrading musical structure under short-segment settings.

When compared with existing TTM models, S2M-Inject 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.

Table 4 further reports the instruction-adherence accuracy of S2M-Inject across five high-level musical attributes. The model achieves high accuracy on rhythm and mood, while genre, instrumentation, and melody remain relatively more challenging due to the ambiguity of short music clips.

In terms of subjective evaluation, S2M-Inject obtains competitive QMOS and MMOS scores. 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 the more constrained goal of matching reference-speech timbre.

**Ablation Analysis** Table 3 shows that adding the second-stage speech–music cross-domain training substantially improves speaker similarity, reducing WER from 1.77% to 1.56% and increasing SIM from 0.34 to 0.60. The subjective scores also improve consistently, confirming that speech-derived timbre supervision strengthens cross-domain voice alignment while maintaining music generation quality. This improvement suggests that Stage 1

mainly establishes in-domain singing timbre modeling, while Stage 2 exposes the model to speech-derived timbre conditions and reduces the mismatch between reference speech and generated singing vocals. As a result, the model learns to associate speech speaker embeddings with the corresponding singing timbre more effectively.

## 4 Discussion

Using speech as the timbre source provides stable and scalable identity cues, and the proposed two-stage training strategy further improves cross-domain voice alignment; 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. Since our training corpus primarily consists of short music clips ranging from 2 to 20 seconds, this work focuses on short-segment speech-to-music voice cloning and controllable music generation. Long-form music generation, which requires modeling verse–chorus structure, long-range accompaniment consistency, and temporal coherence across sections, is beyond the current scope of this study.

## 5 Conclusion and Future Work

In this paper, we presented S2M-Inject, an MM-DiT-based framework for speech-conditioned cross-domain voice cloning in music generation. By injecting speaker representations extracted from reference speech into an instruction-guided music generation process, S2M-Inject enables speech-to-music timbre transfer while supporting controllable music generation. Experimental results demonstrate improved timbre consistency and competitive music generation quality under short-segment settings. In future work, we will explore long-form music generation, stronger cross-domain timbre modeling, and broader voice cloning applications across 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 music may introduce a domain mismatch, which can affect timbre faithfulness and perceptual consistency in some cases. In the current evaluation, reference speech conditions are constructed for held-out identities and excluded from training, but robustness to fully in-the-wild natural speech recordings requires further study.

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. This work adopts a lyric-level timbre fusion strategy, but does not exhaustively compare alternative timbre injection locations or fusion mechanisms, such as independent timbre streams or attention-bias-based conditioning. 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. Musi-clm: 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*.
- Bingsong Bai, Yizhong Geng, Fengping Wang, Cong Wang, Puyuan Guo, Yingming Gao, and Ya Li. 2026. Hq-svc: Towards high-quality zero-shot singing voice conversion in low-resource scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30013–30021.
- Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023a. Audioldm: 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. Diffrrhythm+: 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. Musioldm: 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.
- Shuqi Dai, Yunyun Wang, Roger B. Dannenberg, and Zeyu Jin. 2025. Everyone-can-sing: Zero-shot singing voice synthesis and conversion with speech reference. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- 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*.
- 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 Dinulescu, 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. 2023a. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*.
- Rongjie Huang, Chunlei Zhang, Yongqi Wang, Dongchao Yang, Luping Liu, Zhenhui Ye, Ziyue Jiang, Chao Weng, Zhou Zhao, and Dong Yu. 2023b. Make-a-voice: Unified voice synthesis with discrete representation. *arXiv preprint arXiv:2305.19269*.
- 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*.
- Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, and 1 others. 2024. Naturspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.
- 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 and Jimmy Ba. Adam: A method for stochastic optimization.
- 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.
- Binzhu Sha, Xu Li, Zhiyong Wu, Ying Shan, and Helen Meng. 2024. Neural concatenative singing voice conversion: Rethinking concatenation-based approach for one-shot singing voice conversion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12577–12581. IEEE.
- 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. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.
- Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023b. CAM++: A fast and efficient network for speaker verification using context-aware masking. In *Interspeech 2023*, pages 5301–5305.
- 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, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- 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*.
- Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu. 2025. Vevo2: Bridging controllable speech and singing voice generation via unified prosody learning. *arXiv preprint arXiv:2508.16332*.
- Siqi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. In-dextts2: 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, Weiwei 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*.

## A Related Works

### A.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., 2023a) 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 (Ju et al., 2024) proposes a factorized codec and diffusion framework that disentangles content, prosody, timbre, and acoustic details for zero-shot TTS. 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., 2023b) leverage semantic or discrete voice representations as intermediate layers, enabling flexible voice synthesis and 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) explore efficient 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 ap-

proaches are largely restricted to speech-to-speech scenarios. Most methods assume that both the reference and target signals belong to the speech domain, while cross-domain voice cloning from speech to full music generation remains comparatively less explored. Recent singing voice systems have started to explore speech-referenced singing generation. Everyone-Can-Sing (Dai et al., 2025) studies zero-shot singing voice synthesis and conversion with a speech reference, using lyrics and musical scores for vocal-level content and prosody control. Vevo2 (Zhang et al., 2025) further bridges controllable speech and singing voice generation through unified prosody learning, supporting synthesis, conversion, and editing tasks for human voice. Different from these voice-level systems, S2M-Inject targets instruction-guided full music generation with accompaniment rather than score-conditioned vocal synthesis, and injects speech-derived timbre representations into a text-to-music generation framework to support speech-to-music voice cloning. To further address this gap, 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.

### A.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., 2023a), and MusicLDM (Chen et al., 2024) employ text-conditioned diffusion or latent diffusion frameworks for audio/music generation, with MusicLDM further adapting AudioLDM-style modeling to the music domain and improving novelty through beat-synchronous mixup strategies.

Subsequent works, including JEN-1 (Yao et al., 2025b) and Mustango (Melechovsky et al., 2024), introduce multi-track modeling or explicit musical priors (e.g., rhythm, chord, beat, and key conditions) to further enhance structural controllability. 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, these methods primarily focus on textual descriptions, melody/style control, or music-intrinsic attributes, while explicit cross-domain voice cloning from reference speech remains less explored (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 mainly emphasize text, melody, style, or music-intrinsic audio conditions, while speech-derived timbre conditioning for full music generation remains less explored. 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.

## B Dataset Attribute Distributions

Figure 4 summarizes the attribute distributions of the training and test splits used in our experiments. The test set is constructed to maintain balanced language and singer-gender coverage, while the training set preserves broader in-the-wild diversity across genres, instrumentation, tempo, singer

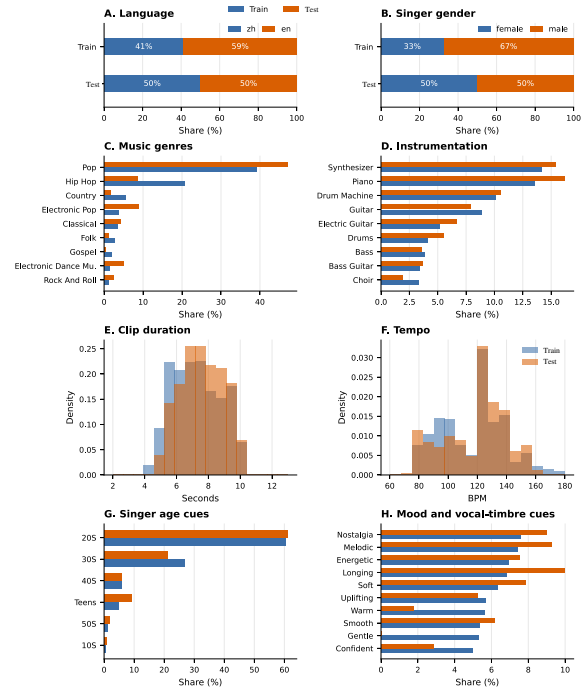


Figure 4: Attribute distributions of the training and test datasets.

age cues, and mood or vocal-timbre descriptors. These distributions provide additional context for interpreting the short-segment cross-domain voice cloning and music generation results.