SoulX-Singer: Towards High-Quality Zero-Shot Singing Voice Synthesis

Jiale Qian[1]* **Hao Meng**[1,3]* **Tian Zheng**[1] **Pengcheng Zhu**[2] **Haopeng Lin**[1]
Yuhang Dai[1,4] **Hanke Xie**[1,4] **Wenxiao Cao**[1] **Ruixuan Shang**[1] **Jun Wu**[1]
Hongmei Liu[1] **Hanlin Wen**[1] **Jian Zhao**[2] **Zhonglin Jiang**[2] **Yong Chen**[2]
Shunshun Yin[1] **Ming Tao**[1] **Jianguo Wei**[3] **Lei Xie**[4] **Xinsheng Wang**[1]†

[1]Soul AI Lab, China
[2]AI Center, Geely Automobile Research Institute (Ningbo) Co., Ltd., Ningbo, China
[3]Audio-Visual Cognitive Computing Team, Tianjin University, Tianjin, China
[4]Audio, Speech and Language Processing Group (ASLP@NPU),
Northwestern Polytechnical University, Xi'an, China

Abstract

While recent years have witnessed rapid progress in speech synthesis, open-source singing voice synthesis (SVS) systems still face significant barriers to industrial deployment, particularly in terms of robustness and zero-shot generalization. In this report, we introduce SoulX-Singer, a high-quality open-source SVS system designed with practical deployment considerations in mind. SoulX-Singer supports controllable singing generation conditioned on either symbolic musical scores (MIDI) or melodic representations, enabling flexible and expressive control in real-world production workflows. Trained on more than 42,000 hours of vocal data, the system supports Mandarin Chinese, English, and Cantonese, and consistently achieves state-of-the-art synthesis quality across languages under diverse musical conditions. Furthermore, to enable reliable evaluation of zero-shot SVS performance in practical scenarios, we construct SoulX-Singer-Eval, a dedicated benchmark with strict training–test disentanglement, facilitating systematic assessment in zero-shot settings.

Demo page: https://soul-ailab.github.io/soulx-singer
Source code: https://github.com/Soul-AILab/SoulX-Singer

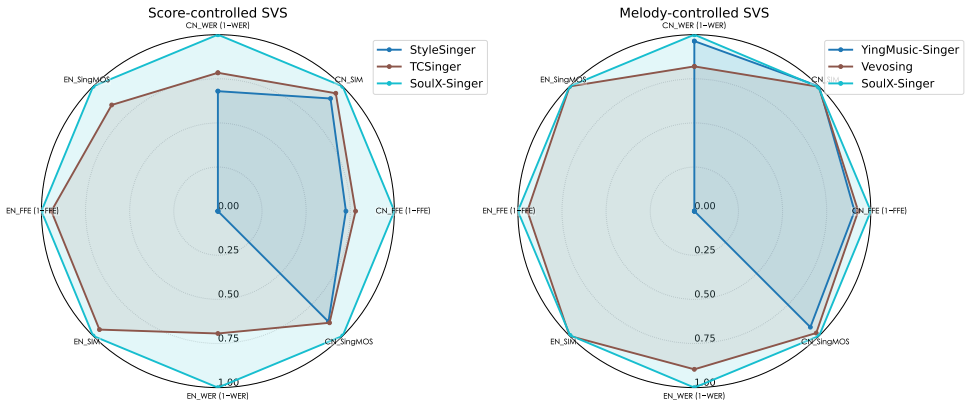

Figure 1: Performance of SoulX-Singer.

*Equal contribution.
†Corresponding author. qianjiale@soulapp.cn, menghao@soulapp.cn, wangxinsheng@soulapp.cn

1 Introduction

Singing Voice Synthesis (SVS) aims to generate expressive human vocals conditioned on lyrics and musical scores. Although significant progress has recently been made in both speech synthesis [1], [2], [3], [4], [5], [6], [7] and music generation [8], [9], [10], high-quality singing voice generation with flexible control in a zero-shot setting remains largely unavailable. To address this gap, we introduce SoulX-Singer, a singing voice synthesis system capable of high-quality zero-shot generation across multiple languages.

Early work on SVS primarily focused on synthesizing singing voices from speakers or singers observed during training [11], [12], [13]. Representative systems such as DiffSinger [11] were trained on relatively small-scale datasets with carefully curated annotations, and therefore lacked the ability to generalize to unseen singers. This limitation significantly restricts the applicability of such models in real-world scenarios. Subsequent studies, including StyleSinger [14] and the TCSinger series [15], [16], began to explore zero-shot SVS. However, these methods were still trained on datasets consisting of only a few hundred hours of singing data from a limited number of singers, making it difficult to achieve robust zero-shot generalization in practice.

More recently, Vevo2 [17] and YingMusic-Singer [18] leveraged the scaling capabilities of the Transformer [19] and Diffusion Transformer (DiT) [20] architecture, respectively, expanding singing datasets to the scale of several thousand hours. Despite these advances, both systems adopt a melody-driven synthesis paradigm and do not provide note-level duration control, which leads to two key limitations. First, melody extraction from existing songs is required, preventing song generation purely from musical scores and lyrics. Second, the absence of explicit note duration modeling makes syllable-level timing uncontrollable, resulting in temporal misalignment between the synthesized vocals and the original accompaniment. This severely limits practical usage in music production workflows, particularly for mixing and arrangement.

To address these challenges, we propose **SoulX-Singer**. In contrast to prior work relying on at most several thousand hours of training data, we construct a large-scale vocal dataset comprising over **42,000 hours** of singing audio, substantially enhancing the model's zero-shot generalization capability. From an algorithmic perspective, SoulX-Singer is designed to support both **score-based** controllable singing generation and **melody-conditioned** synthesis, enabling flexible control under diverse real-world scenarios. This dual-control design allows SoulX-Singer to better accommodate music creation workflows based on symbolic musical scores as well as scenarios requiring melody-guided generation from existing songs.

The main contributions of this work can be summarized as follows:

- **A High-Quality Zero-shot SVS Model**. We propose **SoulX-Singer**, a high-quality singing voice synthesis model designed for zero-shot scenarios. SoulX-Singer supports both **music score-based inputs** and **melody-based inputs** within a unified framework, enabling high-fidelity timbre cloning, reference-based singing style transfer, and flexible editing of both musical scores and lyrics. This design makes the system suitable for a wide range of real-world music production workflows.
- **A Large-Scale Singing Voice Data Processing Pipeline**. We introduce a large-scale data processing pipeline that takes songs with background music as input and automatically produces clean vocal recordings paired with aligned lyrics and musical scores. Using this pipeline, we construct a multilingual singing dataset comprising over 42,000 hours of vocal recordings in **Mandarin Chinese**, **English**, and **Cantonese**, representing an order-of-magnitude increase compared to the data scale used in existing SVS studies.
- **A Dedicated Evaluation Benchmark for Zero-shot SVS**. To facilitate systematic and reproducible evaluation of zero-shot SVS performance, we introduce a dedicated benchmark dataset featuring **50 unseen singers** in Mandarin Chinese and English, accompanied by **fine-grained, note-level score annotations**. This benchmark provides a standardized evaluation protocol for assessing synthesis quality, controllability, and generalization, and serves as a reliable testbed for future SVS research.

In the following sections, we introduce the data construction and model architecture of SoulX-Singer, followed by a comprehensive evaluation of its performance on singing voice synthesis tasks.

2 Method

Clean vocal recordings annotated with aligned MIDI scores and lyrics are essential for training SVS models with fine-grained, score-based controllability. In this section, we first describe the data processing pipeline adopted in this work, including vocal extraction from mixed songs and the annotation of MIDI information and lyric transcriptions. We then present the overall architecture and core algorithms of SoulX-Singer.

2.1 Data Processing

To obtain aligned vocal audio, MIDI, and lyric annotations, as illustrated in Figure 2, the overall data processing workflow consists of vocal separation, lyric transcription, and note transcription. In addition, to enable melody-controlled SVS, the fundamental frequency (F0) is extracted from the vocal recordings.

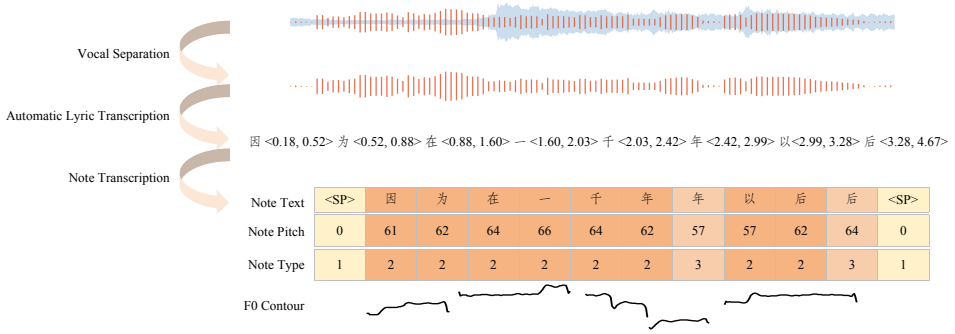

Figure 2: Pipeline for large-scale singing data curation: from raw audio extraction to time-aligned MIDI and text formulation.

2.1.1 Processing Workflow

Vocal Separation. To obtain high-purity dry vocals devoid of backing harmonies, we employ a two-stage vocal extraction process. Specifically, a pretrained lead vocal separation model[3] is adopted to isolate the lead vocal and suppress accompanying harmonies. Since commercial recordings often contain reverberation introduced during mixing, a pretrained vocal de-reverberation model[4] is subsequently applied. Both models are based on Mel-Band Roformer [21], providing a clean acoustic foundation that supports high-fidelity synthesis and large-scale training.

Automatic Lyric Transcription. To ensure accurate alignment of lyrics with vocals, we first perform robust language identification. Specifically, SenseVoiceSmall[5] [22] is adopted as a pretrained backbone and fine-tuned on a singing dataset labeled with Mandarin, Cantonese, and English. Once the language is determined, we extract lyrics and word-level timestamps using language-specific ASR models: Paraformer[6] [23] for Mandarin and Cantonese, and Parakeet-TDT-0.6B-V2[7] for English. To guarantee data quality, extracted lyrics are compared with reference lyrics annotated at the sentence level. Samples containing insertion or deletion errors are discarded, and substitution errors are corrected using the reference words. This procedure ensures the linguistic accuracy required for robust large-scale training.

Note Transcription. To generate note-level representations aligned with the lyrics, we employ the ROSVOT model [24] for pitch-class estimation and note boundary detection. Specifically, timestamped lyrics from the previous stage are used as input, producing a sequence of note-level tokens—including text, pitch, note type, and duration—precisely aligned along the temporal axis.

[3] https://huggingface.co/becruily/mel-band-roformer-karaoke

[4] https://huggingface.co/anvuew/dereverb_mel_band_roformer

[5] https://www.modelscope.cn/models/iic/SenseVoiceSmall

[6] https://modelscope.cn/models/iic/speech_seaco_paraformer_large_asr_nat-zh-cn-16k-common-vocab8404-pytorch

[7] https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2

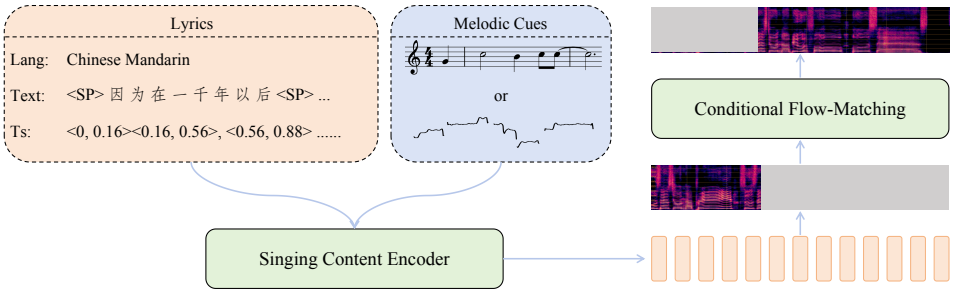

Figure 3: Overview of SoulX-Singer.

These aligned tokens provide the necessary foundation for controllable and expressive singing voice synthesis.

2.1.2 Corpus Overview

Using the aforementioned processing methods, we obtained approximately 42,000 hours of high-quality vocal recordings, with roughly 20k hours each in Mandarin and English, and about 2k hours in Cantonese. To enable precise modeling of singing voice, the dataset is organized at the musical note level, which serves as the fundamental unit for synthesis. Specifically, each note is represented as a tuple consisting of its corresponding textual token, pitch class, and note type. The note type is defined as a categorical attribute, with values 1, 2, and 3 corresponding to rest, lyric, and slur notes, respectively. This note-level representation ensures high-precision alignment between linguistic and melodic structures, providing a robust foundation for controllable and expressive singing voice synthesis.

2.2 SoulX-Singer

SoulX-Singer is a non-autoregressive (NAR) singing voice synthesis model based on flow matching, designed for high-quality and controllable singing voice generation. The core backbone of the model is a flow matching decoder built with Diffusion Transformer (DiT) [25], which takes lyrics and melody cues as input and predicts mel-spectrograms. The predicted mel-spectrograms are subsequently converted into waveforms using a neural vocoder.

To effectively encode the diverse and multimodal information required for SVS, i.e, lyrics, musical scores, note types, and F0, SoulX-Singer incorporates a Singing Content Encode. This encoder transforms the input features into rich, temporally aligned representations, providing the decoder with a structured and informative latent space for accurate mel-spectrogram generation. By combining flow-based NAR modeling with specialized content encoding, SoulX-Singer achieves both expressive control over vocal performance and efficient, high-fidelity synthesis.

2.2.1 Feature Representation

Textual Representation. For Mandarin and Cantonese, the modeling units are defined as character-level pinyin, whereas English is represented using phonemes. To explicitly distinguish phoneme boundaries across different words, the phoneme sequence corresponding to each English word is wrapped with special boundary tokens, starting with <BOW> and ending with <EOW>. To further disambiguate pinyin representations between Mandarin and Cantonese, language-specific tags are appended to the pinyin tokens, enabling the model to distinguish pronunciation patterns across languages. The text embeddings are obtained via a linear embedding layer.

Melodic Representation. The melodic input consists of discrete note pitch sequences and continuous F0 sequences. Both sequences are first processed through binary gating layers, followed by a linear projection to produce note pitch embeddings and F0 embeddings, respectively. The gating mechanism regulates the contribution of each prosodic feature: during training, either the note pitch or F0 input is randomly dropped to encourage robust feature extraction from a single modality; during inference, the model can flexibly enable the corresponding gate to perform melody-based or score-based generation.

Length Regulation and Feature Integration. To unify heterogeneous representations into a shared temporal resolution before decoding, we adopt a length regulator as the core feature integration mechanism. Specifically, the embeddings corresponding to note type, note pitch, and text tokens are expanded according to the duration of each musical note, aligning all note-level representations to the mel-spectrogram time scale. After length expansion, these embeddings share identical temporal and feature dimensions and are combined through element-wise addition, producing a unified conditioning sequence that is subsequently fed into the decoder. By explicitly enforcing note-to-mel alignment, the length regulator enables precise synchronization between linguistic content and melodic structure, which is essential for controllable and expressive singing voice synthesis.

2.2.2 Training

The training of SoulX-Singer is conducted in two stages to progressively enhance model robustness and long-form generation capability. In the first stage, the model is trained on relatively short audio segments, with durations ranging from 2 s to 16 s. During this stage, the prompt mel-spectrogram is deliberately sampled from a non-adjacent segment of the target audio. This design encourages the model to rely less on local acoustic continuity and more on the provided linguistic and musical conditions, thereby improving robustness and generalization under diverse prompt conditions.

In the second stage, the training strategy is shifted toward long-form singing voice modeling. Adjacent audio segments are concatenated to construct longer training samples with durations between 30 s and 90 s, enabling the model to capture long-range temporal dependencies in singing performances. To further strengthen the model's prompt-following capability, the prompt audio in this stage is sampled from the immediately preceding adjacent segment of the target audio.

By adopting this two-stage training strategy—progressing from short, non-adjacent prompting to long, contextually adjacent prompting—SoulX-Singer achieves both robust conditional generation and effective modeling of long-duration singing audio.

2.2.3 Inference

During inference, SoulX-Singer offers high flexibility by supporting two complementary generation modes, which can be seamlessly switched according to the available prosodic control signals.

Melody-control Mode. This mode is intended for scenarios where a target melody is available. In this configuration the model takes the target lyrics together with a continuous F0 contour extracted from a reference audio as primary inputs. This design enables faithful preservation of fine-grained melodic details and expressive vocal techniques, while still allowing flexible modification of lyrics or transformation of timbre.

Score-control Mode. This mode is designed for creative synthesis scenarios. In this mode, the inputs consist solely of MIDI score information and the target lyrics. The model autonomously predicts naturalistic acoustic features constrained by the provided score. By eliminating the dependency on pre-recorded melody lines, this mode provides creators with greater artistic freedom for high-fidelity zero-shot SVS.

3 Performance of SoulX-Singer

3.1 Evaluation Dataset

Due to the absence of a widely adopted and standardized benchmark for singing voice synthesis (SVS), we construct two complementary evaluation datasets to comprehensively assess the performance of SoulX-Singer under both open-source and zero-shot conditions: GMO-SVS and SoulX-Singer-Eval.

GMO-SVS. The GMO-SVS dataset is built upon multiple publicly available SVS corpora, including **G**TSinger [26], **M**4Singer [27], and **O**pencpop [28]. For M4Singer and Opencpop, we directly adopt their official test splits. Specifically, M4Singer contains 8 Mandarin songs performed by 4 singers spanning four vocal ranges, while Opencpop provides 5 Mandarin songs sung by a single professional female singer. GTSinger contributes 25 songs in both English and Mandarin from 5 singers, covering a diverse set of vocal techniques and expressive styles.

In total, GMO-SVS consists of 802 samples. For each song, the first sentence is used as the acoustic prompt, while the remaining content is synthesized by the evaluated models. Importantly, GMO-SVS preserves the ground-truth recordings of the prompt singers, enabling a comprehensive evaluation of pronunciation accuracy, prosodic consistency, and overall synthesis quality. It is worth emphasizing that none of the above open-source datasets are used during the training of SoulX-Singer, ensuring a fair and unbiased evaluation.

To further evaluate the model's performance in singing voice editing scenarios, we create modified versions of the target lyrics. Specifically, the original lyrics of each song are rewritten using the DeepSeek-V large language model, while strictly maintaining the same number of words as the original. This setup allows us to assess how well the model can adapt to lyric modifications while preserving melodic and expressive characteristics.

SoulX-Singer-Eval. This dataset is a newly collected dataset designed to evaluate zero-shot generalization on unseen speakers. It contains 100 singing segments from 50 distinct individuals (25 Mandarin and 25 English speakers), with 2 segments per speaker. The Mandarin data were collected from recruited professional and amateur singers who consented to open-source their voice data for academic purposes. The English segments were sliced and filtered from the multitrack *Mixing Secrets* dataset [29]. Crucially, all segments in SoulX-Singer-Eval underwent precise manual melody annotation to meet the prompt input requirements of various zero-shot SVS models. The target lyrics and melody for synthesis are randomly selected from 15 Mandarin and 15 English tracks in GMO-SVS. This set introduces speakers unseen by any baseline models, providing a rigorous benchmark for timbre cloning and style transfer capabilities.

3.2 Evaluation Metrics

To comprehensively assess the performance of singing voice synthesis, we evaluate SoulX-Singer across multiple key dimensions, including melodic accuracy, timbre similarity, intelligibility, and overall singing quality. Each metric is carefully selected to quantify different aspects of synthesized vocal fidelity and expressiveness.

Melodic Accuracy. We measure pitch precision using F0 Frame Error (FFE), defined as the proportion of frames in which the predicted pitch deviates by more than 20% from the ground-truth F0.

Timbre Similarity. To evaluate the model's ability to reproduce the timbre of the acoustic prompt in zero-shot scenarios, we compute cosine similarity (SIM) between speaker embeddings of the synthesized audio and the corresponding prompt. The embeddings are extracted using a WavLM-based speaker verification model [30].

Intelligibility. We assess pronunciation accuracy via Word Error Rate (WER), calculated by comparing the target lyrics with automatic speech recognition (ASR) transcriptions of the synthesized audio. For Mandarin samples, character-level error rate (CER) is the standard metric; however, for simplicity and consistency across languages, we report all results using WER. Paraformer [23] is used for Mandarin, and Whisper-large-v3 [31] is used for English samples.

Singing Quality. Overall perceptual quality is evaluated using two learned objective metrics. SingMOS [32] is a singing-specific quality metric trained to align closely with human perception. Sheet-SSQA (Sheet) [33], derived from the MOS-Bench framework, is used to assess perceptual quality in zero-shot scenarios and demonstrates strong generalization across unseen speakers.

3.3 Results

3.3.1 Performance on the GMO-SVS

Comparisons of SoulX-Singer and other baseline methods on GMO-SVS are presented in Table 1, where Singing Voice Editing indicates that the lyrics are rewritten. For the SVS task, SoulX-Singer outperforms all baseline models in both Mandarin and English. When controlled with continuous F0 contours (melody-control mode), the model achieves the lowest FFE, significantly surpassing the best baseline, YingMusic-Singer. This demonstrates that explicit acoustic features guide the flow-matching decoder to generate accurate pitch trajectories. In contrast, when driven by discrete MIDI notes (score-control mode), SoulX-Singer attains the lowest WER, outperforming Vevosing and

TCSinger. This indicates that MIDI-based timing constraints help stabilize pronunciation and rhythm, particularly for complex phonemes. Across both modes, SoulX-Singer also achieves state-of-the-art performance on SingMOS and SIM, confirming the robustness and generalization capability of the architecture under both pitch- and score-conditioned synthesis scenarios.

The results under Singing Voice Editing reflect the performance of different models when the lyrics are modified. As shown in the table, melody-based methods exhibit a noticeable drop in intelligibility compared to the original lyrics. This decline arises because the original melody and lyrics are inherently correlated—for example, the pitch contours are aligned with the original words—so modifying the lyrics introduces a certain degree of mismatch. In contrast, SoulX-Singer, which supports MIDI score-based control, maintains better performance under lyric modification. By relying on the explicit score rather than the original acoustic melody, the model can effectively accommodate rewritten lyrics without sacrificing pronunciation accuracy or rhythmic consistency.

Table 1: Performance of different models on GMO-SVS (↑ larger is better, ↓ smaller is better; bold indicates the best result).

Model	Control	Chinese					English				
		WER ↓	SIM ↑	FFE ↓	SingMOS ↑	Sheet ↑	WER ↓	SIM ↑	FFE ↓	SingMOS ↑	Sheet ↑
Singing Voice Synthesis											
GroundTruth	-	0.074	-	-	4.624	4.334	0.197	-	-	4.441	3.825
StyleSinger	Score	0.367	0.817	0.363	3.938	3.819	-	-	-	-	-
TCSinger	Score	0.270	0.855	0.315	3.977	3.843	0.410	0.879	0.211	3.662	3.482
YingMusic-Singer	Melody	0.099	0.902	0.132	4.145	3.620	-	-	-	-	-
Vevosing	Melody	0.233	0.899	0.112	4.355	4.005	0.239	0.922	0.088	4.321	3.699
SoulX-Singer	Melody	**0.065**	0.897	**0.044**	**4.458**	**4.110**	0.151	0.918	**0.036**	**4.323**	**3.751**
SoulX-Singer	Score	0.069	**0.905**	0.122	4.445	4.107	**0.149**	**0.926**	0.164	4.303	3.705
Singing Voice Editing											
YingMusic-Singer	Melody	0.146	**0.904**	0.180	4.107	3.612	-	-	-	-	-
Vevosing	Melody	0.308	0.897	0.148	4.328	4.011	0.484	**0.924**	0.108	4.267	**3.691**
SoulX-Singer	Melody	0.212	0.895	**0.057**	4.325	4.024	0.444	0.911	**0.047**	4.141	3.615
SoulX-Singer	Score	**0.089**	**0.904**	0.181	**4.447**	**4.125**	**0.213**	0.920	0.200	**4.269**	3.678

3.3.2 Performance on the SoulX-Singer-Eval

To evaluate the model's capability in timbre cloning and style transfer, we conduct experiments on SoulX-Singer-Eval, which contains speakers that are completely unseen during training for all compared methods. As shown in Table 2, SoulX-Singer consistently outperforms baseline models even with unseen target speakers. The SoulX-Singer controlled with music scores achieves the highest SIM scores (0.922 for Mandarin and 0.914 for English), demonstrating robust zero-shot voice cloning.

Table 2: Performance of different models on SoulX-Singer-Eval (↑ larger is better, ↓ smaller is better; bold indicates the best result).

Model	Control	**Chinese**				**English**			
		WER ↓	SIM ↑	SingMOS ↑	Sheet ↑	WER ↓	SIM ↑	SingMOS ↑	Sheet ↑
GroundTruth	-	0.089	-	4.450	4.237	0.208	-	4.569	3.815
StyleSinger	Score	0.388	0.808	3.892	3.833	-	-	-	-
TCSinger	Score	0.227	0.822	3.939	4.009	0.460	0.729	3.652	3.685
YingMusic-Singer	Melody	0.117	0.913	4.068	3.599	-	-	-	-
Vevosing	Melody	0.233	0.908	4.265	3.814	0.256	0.888	4.184	3.582
SoulX-Singer	Melody	**0.069**	0.902	**4.394**	**4.053**	0.155	0.870	4.223	**3.703**
SoulX-Singer	Score	**0.069**	**0.922**	4.370	4.010	**0.129**	**0.914**	**4.255**	3.690

3.3.3 Cross-Lingual Synthesis Evaluation

To further evaluate the model's ability in cross-lingual synthesis (e.g., using a Mandarin prompt to synthesize English singing), which requires strict disentanglement of speaker identity from linguistic content, we conducted experiments reported in Table 3. The baseline model Vevosing exhibits severe

intelligibility degradation, with a WER as high as 0.717, indicating leakage of linguistic patterns from the prompt into the generated output. In contrast, SoulX-Singer achieves a WER of 0.110 while maintaining a high SIM of 0.898, demonstrating strong preservation of speaker identity across languages. These results highlight the effectiveness of the singing content encoder, which successfully separates language-independent timbre features from language-dependent linguistic content, enabling high-fidelity cross-lingual style transfer without compromising pronunciation or vocal identity.

Table 3: Cross-Lingual Synthesis Performance on SoulX-Singer-Eval (↑ larger is better, ↓ smaller is better; bold indicates the best result).

Model	Control	**Cross-Lingual**			
		WER ↓	SIM ↑	SingMOS ↑	Sheet ↑
GroundTruth	-	0.148	-	4.510	4.026
TCSinger	Score	0.333	0.789	3.817	3.830
Vevosing	Melody	0.717	0.877	4.133	3.631
SoulX-Singer	Melody	0.122	0.866	**4.342**	**3.939**
SoulX-Singer	Score	**0.110**	**0.898**	4.337	3.883

4 Conclusions

In this work, we introduced SoulX-Singer, a high-fidelity, zero-shot singing voice synthesis model. Leveraging a note-level aligned dataset of over 42,000 hours, the model enables flexible control over both melody and score, while faithfully reproducing diverse timbres. Extensive evaluations across multiple benchmarks demonstrate that SoulX-Singer consistently outperforms state-of-the-art baselines in pitch accuracy, intelligibility, timbre similarity, and overall singing quality. By combining explicit acoustic and symbolic cues with large-scale data, SoulX-Singer provides a robust foundation for creative applications, including personalized singing synthesis, music production, and multilingual vocal content generation. This work paves the way for future research in zero-shot expressive SVS and singing voice editing.

5 Ethics Statement

The development and release of SoulX-Singer raise important ethical considerations. As a zero-shot singing voice synthesis system, the model is capable of generating realistic singing voices conditioned on a short reference prompt, which may introduce risks related to voice impersonation and misuse. Users of SoulX-Singer are therefore strongly encouraged to respect intellectual property, privacy, and personal consent when generating singing content. The system should not be used to impersonate individuals without authorization, nor to produce deceptive or misleading audio content.

6 Acknowledgments

We would like to express our sincere gratitude to Rizhao Youle Studio and the following individuals for contributing high-quality vocal recordings and providing valuable assistance in the construction of the SoulX-Singer-Eval benchmark: Zhengyu Chen, Hao Fan, Jialan Kuang, Haotian Luo, Linfei Ma, Hao Meng, Rui Shi, Wenbin Shi, Yucun Sui, Xuewen Sun, Jiayu Wang, Weiliang Wang, Shuhua Weng, Minghao Xu, Dan Yang, Baojin Yi, Youpeng Yuan, Meng Zhang, Peng Zhao, and Zhiyi Zheng.

References

[1] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, and X. Chen, "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)*, 2025, pp. 6255–6271.

[2] X. Wang, M. Jiang, Z. Ma, Z. Zhang, S. Liu, L. Li, Z. Liang, Q. Zheng, R. Wang, X. Feng, et al., "Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens," *arXiv preprint arXiv:2503.01710*, 2025.

[3] S. Zhou, Y. Zhou, Y. He, X. Zhou, J. Wang, W. Deng, and J. Shu, "Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech," *arXiv preprint arXiv:2506.21619*, 2025.

[4] Z. Du, C. Gao, Y. Wang, F. Yu, T. Zhao, H. Wang, X. Lv, H. Wang, C. Ni, X. Shi, et al., "Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training," *arXiv preprint arXiv:2505.17589*, 2025.

[5] K. Xie, F. Shen, J. Li, F. Xie, X. Tang, and Y. Hu, "Fireredtts-2: Towards long conversational speech generation for podcast and chatbot," *arXiv preprint arXiv:2509.02020*, 2025.

[6] Y. Zhou, G. Zeng, X. Liu, X. Li, R. Yu, Z. Wang, R. Ye, W. Sun, J. Gui, K. Li, et al., "Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning," *arXiv preprint arXiv:2509.24650*, 2025.

[7] H. Hu, X. Zhu, T. He, D. Guo, B. Zhang, X. Wang, Z. Guo, Z. Jiang, H. Hao, Z. Guo, et al., "Qwen3-tts technical report," *arXiv preprint arXiv:2601.15621*, 2026.

[8] Y. Jiang, H. Chen, Z. Ning, J. Yao, Z. Han, D. Wu, M. Meng, J. Luan, Z. Fu, and L. Xie, "Diffrhythm 2: Efficient and high fidelity song generation via block flow matching," *arXiv preprint arXiv:2510.22950*, 2025.

[9] R. Yuan, H. Lin, S. Guo, G. Zhang, J. Pan, Y. Zang, H. Liu, Y. Liang, W. Ma, X. Du, et al., "Yue: Scaling open foundation models for long-form music generation," *arXiv preprint arXiv:2503.08638*, 2025.

[10] C. Yang, S. Wang, H. Chen, W. Tan, J. Yu, and H. Li, "Songbloom: Coherent song generation via interleaved autoregressive sketching and diffusion refinement," *arXiv preprint arXiv:2506.07634*, 2025.

[11] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, pp. 11 020–11 028.

[12] H. Xue, X. Wang, Y. Zhang, L. Xie, P. Zhu, and M. Bi, "Learn2sing 2.0: Diffusion and mutual information-based target speaker svs by learning from singing teacher," *arXiv preprint arXiv:2203.16408*, 2022.

[13] Y. Lei, S. Yang, X. Wang, Q. Xie, J. Yao, L. Xie, and D. Su, "Unisyn: An end-to-end unified model for text-to-speech and singing voice synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 13 025–13 033.

[14] Y. Zhang, R. Huang, R. Li, J. He, Y. Xia, F. Chen, X. Duan, B. Huai, and Z. Zhao, "Stylesinger: Style transfer for out-of-domain singing voice synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 19 597–19 605.

[15] Y. Zhang, Z. Jiang, R. Li, C. Pan, J. He, R. Huang, C. Wang, and Z. Zhao, "Tcsinger: Zero-shot singing voice synthesis with style transfer and multi-level style control," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 1960–1975.

[16] Y. Zhang, W. Guo, C. Pan, D. Yao, Z. Zhu, Z. Jiang, Y. Wang, T. Jin, and Z. Zhao, "Tcsinger 2: Customizable multilingual zero-shot singing voice synthesis," *arXiv preprint arXiv:2505.14910*, 2025.

[17] X. Zhang, J. Zhang, Y. Wang, C. Wang, Y. Chen, D. Jia, Z. Chen, and Z. Wu, "Vevo2: Bridging controllable speech and singing voice generation via unified prosody learning," *arXiv e-prints*, arXiv–2508, 2025.

[18] J. Zheng, C. Hao, G. Ma, X. Zhang, G. Chen, C. Ding, Z. Chen, and L. Xie, "Yingmusic-singer: Zero-shot singing voice synthesis and editing with annotation-free melody guidance," *arXiv preprint arXiv:2512.04779*, 2025.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[21] J.-C. Wang, W.-T. Lu, and M. Won, "Mel-band roformer for music source separation," in *Ismir 2023 Hybrid Conference*, 2023.

[22] K. An, Q. Chen, C. Deng, Z. Du, C. Gao, Z. Gao, Y. Gu, T. He, H. Hu, K. Hu, et al., "Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms," *arXiv preprint arXiv:2407.04051*, 2024.

[23] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition," in *Interspeech 2022*, 2022, pp. 2063–2067. DOI: 10.21437/Interspeech.2022-9996.

[24] R. Li, Y. Zhang, Y. Wang, Z. Hong, R. Huang, and Z. Zhao, "Robust singing voice transcription serves synthesis," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9751–9766.

[25] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[26] Y. Zhang, C. Pan, W. Guo, R. Li, Z. Zhu, J. Wang, W. Xu, J. Lu, Z. Hong, C. Wang, et al., "Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks," *Advances in Neural Information Processing Systems*, vol. 37, pp. 1117–1140, 2024.

[27] L. Zhang, R. Li, S. Wang, L. Deng, J. Liu, Y. Ren, J. He, R. Huang, J. Zhu, X. Chen, et al., "M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6914–6926, 2022.

[28] Y. Wang, X. Wang, P. Zhu, J. Wu, H. Li, H. Xue, Y. Zhang, L. Xie, and M. Bi, "Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis," in *Interspeech 2022*, 2022, pp. 4242–4246. DOI: 10.21437/Interspeech.2022-48.

[29] S. Gururani and A. Lerch, "Mixing secrets: A multi-track dataset for instrument recognition in polyphonic music," *Proc. ISMIR-LBD*, pp. 1–2, 2017.

[30] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.

[32] Y. Tang, L. Liu, W. Feng, Y. Zhao, J. Han, Y. Yu, J. Shi, and Q. Jin, "Singmos-pro: An comprehensive benchmark for singing quality assessment," *arXiv preprint arXiv:2510.01812*, 2025.

[33] W.-C. Huang, E. Cooper, and T. Toda, "Mos-bench: Benchmarking generalization abilities of subjective speech quality assessment models," *arXiv preprint arXiv:2411.03715*, 2024.