YingMusic-Singer: Zero-shot Singing Voice Synthesis and Editing with Annotation-free Melody Guidance

Junjie Zheng[1] **Chunbo Hao**[1,2] **Guobin Ma**[1,2] **Xiaoyu Zhang**[1,3]
Gongyu Chen[1] **Chaofan Ding**[1] **Zihao Chen**[1] **Lei Xie**[2]
[1]AI Lab, GiantNetwork [2]ASLP Lab, Northwestern Polytechnical University
[3]University College London

Abstract

Singing Voice Synthesis (SVS) faces significant challenges in real-world applications due to its heavy reliance on precise phoneme-level alignment and melody annotations, which are costly and limit scalability. To address these limitations, we propose a novel melody-driven SVS model that enables arbitrary lyrics to be synthesized with any reference melody without requiring phoneme-level alignment. Our approach leverages a Diffusion Transformer (DiT) based generative model, incorporating a pre-trained melody extraction module to derive melody information directly from reference audio. We introduce an implicit guidance mechanism that utilizes similarity distribution constraints to enhance the stability and coherence of melodies. Furthermore, we optimize duration modeling using weakly-annotation song data and employ Flow-GRPO reinforcement learning with a multi-objective reward function to improve pronunciation clarity, melodic accuracy, and musicality. Experimental results demonstrate that our model significantly outperforms existing methods in both objective metrics and subjective evaluations, particularly in zero-shot and lyric replacement scenarios, while maintaining high auditory quality without manual annotations. This work provides a practical and scalable solution for democratizing singing voice synthesis. To encourage reproducibility and further exploration, we release the inference code and model checkpoints. Code and weights are available at: `https://github.com/GiantAILab/YingMusic-Singer`

1 Introduction

The digital entertainment industry has been continuously evolving, driven by advancements in audio and music technologies. Among these, Singing Voice Synthesis (SVS) stands out as a pivotal research direction with substantial application potential in music production, virtual singers, personal creative endeavors, and interactive media Chen et al. [2020], Zhang et al. [2022a,b], Hong et al. [2023]. Compared to general speech synthesis, SVS must simultaneously satisfy requirements for clear speech content, accurate melodic pitch, and natural singing expression, rendering it considerably more complex than speech synthesis modelCho et al. [2021]. Conventional SVS methods have long relied on precise phoneme-level duration and pitch annotations during both training and inference stages Lu et al. [2020]. This dependency not only necessitates specialized data production pipelines but also impedes the acquisition of large-scale training data, thereby significantly hindering the widespread adoption and industrial deployment of SVS technology. Recently, there has been a growing demand in the industry for melody-controlled singing voice synthesis, where a vocal is generated to match a given reference melody. This approach requires only the lyric and a reference melody audio as inputs, enabling users without professional musical expertise to participate in creative activities—rather than restricting music creation to trained professionals.

However, existing singing voice synthesis methods still exhibit considerable limitations, resulting in a notable performance gap between real-world applications and expectations. On one hand, the vast majority of systems depend on manual labor or alignment tools to obtain precise MIDI rhythms and phoneme-level duration annotations Zhang et al. [2022c], Huang et al. [2021], Wang et al. [2022], Hong et al. [2024]. Such annotations are prohibitively expensive and difficult to scale to songs of arbitrary styles or languages. On the other hand, current approaches typically support only fixed lyric-melody pairs as seen during training. When users attempt to substitute lyrics, mix languages, or alter musical syntactic structures, the inherent mismatch between phoneme counts and melodic beats often leads to issues such as robotic pronunciation, rhythmic misalignment, and unnatural phrasing, substantially degrading the auditory experience. Furthermore, most contemporary methods lack zero-shot capability; their performance deteriorates significantly when encountering unseen text or prosodic structures, which contradicts practical application needs. Thus, transitioning from "usable" to "easy-to-use and practical" remains a critical bottleneck for SVS deployment.

To address these challenges, we propose a singing voice synthesis system that eliminates phoneme-level duration and pitch annotations, and enables free combination of arbitrary lyrics with any reference melody. We design a generative model based on the Diffusion Transformer (DiT) Dhariwal and Nichol [2021], incorporating a melody extraction module to directly derive melody information from reference songs, which is then used as a melodic condition during synthesis to avoid reliance on manual annotations. Recognizing that merely conditioning on melody may not ensure structural adherence to the reference track's overall melodic progression, we further introduce a implicit guidance mechanism based on similarity distribution constraints. Specifically, we compute similarity distribution matrices for both the reference song's MIDI and the model's acoustic flow representations derived via flow-matching, progressively minimizing their discrepancy during training to enable the model to more accurately follow the structural characteristics of the reference melody. This design yields significantly improved melodic stability and singing coherence compared to traditional conditional control methods.

To tackle the issue of duration mismatch between lyrics and melody, we optimize the model using training datas with just sentence-level timestamps Ning et al. [2025], allowing it to automatically infer reasonable duration allocations without phoneme alignment, thereby mitigating problems such as lyric squeezing, beat drift, and abrupt phrasing. Additionally, we pioneer the integration of reinforcement learning into the SVS task. Through Flow-GRPO Liu et al. [2025] policy fine-tuning, we construct a multi-objective reward function that incorporates pronunciation clarity, melodic accuracy, and musical aesthetics, leading to simultaneous improvements in both objective metrics and subjective listening quality.

Our principal contributions are summarized as follows:

- **End-to-End Melody-Driven SVS System for Real-World Applications:** We propose a system that can synthesize arbitrary lyrics with any reference melody without requiring precise phoneme-level duration or pitch annotations. The model automatically learns to align lyrics with melody, substantially reducing both data production costs and usage barriers while improving real-world applicability.
- **Annotation-free Melodic Guidance Based on DiT and Weak Alignment Optimization:** To enable effective lyric–melody alignment, we utilize a pre-trained MIDI extraction module to derive melodic information and introduce a similarity distribution matrix constraint. This ensures that the generated acoustic flow structurally aligns with the reference melody, thereby enhancing pitch stability and vocal naturalness. Moreover, by leveraging weakly annotated songs containing sentence-level timestamps for duration modeling, we mitigate common issues arising from melody–lyric misalignment, such as word crowding, fragmented phrasing, and rhythmic drift.
- **Reinforcement Learning Post-Training via Flow-GRPO and Comprehensive Evaluation:** We develop a reward model integrating content accuracy and melodic similarity, and enhance synthesis quality through policy optimization. Our approach achieves superior results over existing systems on public benchmarks, particularly excelling in zero-shot and lyric editing scenarios, thereby validating the effectiveness and generalization capability of the proposed method.

Experimental results demonstrate that our method significantly outperforms existing systems across multiple public benchmarks and real-song scenarios, exhibiting more stable melodic control and more natural singing expressiveness, especially in lyric replacement and zero-shot synthesis settings.

Moreover, even without manual alignment annotations, our approach maintains auditory quality comparable to — or even better than — models trained with precise alignment, confirming the practicality of our solution for real-world applications.

2 Related Work

2.1 Singing Voice Synthesis (SVS)

Singing Voice Synthesis (SVS) aims to generate natural, fluent, and pitch-accurate singing voices based on lyrics and melody. Early systems (e.g., XiaoiceSing Lu et al. [2020], VISinger Zhang et al. [2022d]) relied on precise phoneme alignment and manually annotated MIDI information, achieving high-quality singing synthesis through a two-stage acoustic model-vocoder structure, but suffering from high training costs and difficult data production. With the introduction of diffusion models, works like DiffSinger Liu et al. [2022] and SmoothSinger Sui et al. [2025] have made significant improvements in sound quality and stability, making end-to-end diffusion-based SVS a mainstream direction. In recent years, the research focus has gradually shifted towards zero-shot and cross-lingual capabilities. TCSinger 2 Zhang et al. [2025a] achieves zero-shot style transfer (not voice cloning) through a fuzzy boundary content encoder and Flow-Transformer structure, supporting multilingual and multi-style controllable singing; CoMeLSinger Zhao et al. [2025] models lyrics and pitch using discrete tokens and achieves structured melody control through contrastive learning; Transinger Shen et al. [2025], based on an IPA phonetic decomposition strategy, shows good generalization in unseen language scenarios. Additionally, RMSSinger He et al. [2023] and N-Singer Lee et al. [2022] attempt to reduce alignment dependency and improve efficiency under real musical scores and non-autoregressive frameworks. Although these methods have made progress in sound quality, stability, and multilingual generalization, they still generally rely on manually annotated training data, lack zero-shot voice cloning capability, and require external conditional inputs such as MIDI or pitch sequences during inference. This makes it difficult for systems to be widely used in non-professional scenarios. Therefore, we propose an architecture that does not belong to the traditional SVS paradigm: it can generate natural singing voices without inputting precise phoneme-level duration or pitch annotations, supports zero-shot voice cloning, and is compatible with traditional SVS tasks, achieving a leap from "usable" to "easy-to-use and effective".

2.2 Reinforcement Learning (RL)

In recent years, reinforcement learning (RL) has been widely used for alignment and generation quality optimization in large models. Since RLHF was proposed for summarization and instruction-following tasks Stiennon et al. [2020], Ouyang et al. [2022], preference modeling based on PPO became the standard paradigm Stiennon et al. [2020], Ouyang et al. [2022], Schulman et al. [2017]. Subsequently, Group Relative Policy Optimization (GRPO) significantly reduced the complexity of RL fine-tuning by using within-group relative scores instead of an explicit value function Shao et al. [2024], and has been applied to Flow Matching and Rectified Flow models, such as Flow-GRPO and Dance-GRPO, effectively improving compliance and aesthetic quality in text-to-image generation tasks Liu et al. [2025], Xue et al. [2025]. In the speech domain, existing work has introduced RLHF into emotional or diffusion-based speech synthesis tasks (e.g., i-ETTS Liu et al. [2021], DLPO Chen et al. [2024]). The recent F5R-TTS Sun et al. [2025a] further applied GRPO policy optimization to a DiT backbone for TTS tasks, achieving end-to-end reward-driven speech generation Sun et al. [2025b]. However, there is still a lack of research systematically applying reinforcement learning to SVS. Existing methods generally rely on manual annotations and external melodic input, making it difficult to directly optimize singing quality through reward signals. To this end, we propose a multi-objective reward model integrating pronunciation clarity, melodic accuracy, and musical aesthetic quality, and improve synthesis performance based on Flow-GRPO policy optimization. Across multiple benchmarks and real song tests, our model outperforms existing systems, particularly excelling in zero-shot and lyric replacement scenarios.

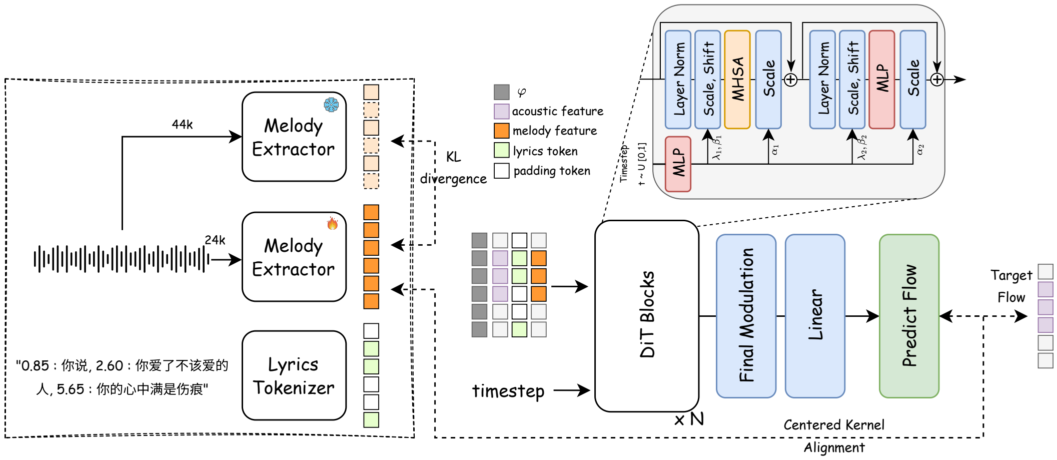

Figure 1: Architecture overview of SVS model

3 Method

3.1 Overview

This paper proposes an end-to-end singing voice generation framework, whose core objective is to synthesize a singing voice that is accurate in content and highly consistent in melody with the input, given text lyrics and a raw audio melody. To achieve this goal, this work abandons the paradigm of mutually independent melody extraction and song synthesis in traditional pipeline models, and instead adopts a design philosophy of synergy and mutual enhancement. As shown in Figure 1, the framework mainly consists of two tightly coupled modules: (1) Online Melody Extraction Module: A parameterized melody extractor responsible for extracting frame-level melodic representations from the raw input audio. (2) Diffusion Transformer-based SVS Module: A denoising diffusion model conditioned on the audio prompt, lyric text and the aforementioned melodic representation, responsible for generating the final singing voice audio.

First, we introduce a distillation-based joint optimization strategy. The melody extraction module is trained end-to-end together with the singing synthesis module. Simultaneously, we utilize a pre-trained and frozen teacher melody extraction model to provide stable supervision signals for the online learning extractor by minimizing the divergence. This design enables the melody extractor to adaptively optimize its representations to be more beneficial for the downstream synthesis task, thereby achieving mutual performance enhancement between the two modules. Second, to ensure the effective utilization of melodic conditions during the generation process, we introduce a representation layer alignment constraint based on Centered Kernel Alignment (CKA) Davari et al. [2022]. This constraint explicitly maximizes the correlation between the extracted melodic representation and the internal feature representations of the singing synthesis model, thereby strengthening the role of melodic guidance within the model and ensuring high consistency between the generated singing and the input melody.

Through the two core designs of online joint optimization and internal representation alignment, this method collectively ensures the accurate and efficient flow of melodic information from input to output, ultimately achieving high-quality and highly robust singing voice synthesis.

3.2 Pre-training

Online Melody Learning and Joint Optimization. Traditional singing synthesis pipelines often treat melody extraction as an independent, fixed pre-processing step, which can lead to error propagation through the pipeline, and the extracted melodic features may deviate from the learning objective of the downstream synthesis model. To solve this problem, we design an online learning melody extractor and perform joint optimization with the singing synthesis model. Specifically, our melody extractor E_ϕ is an encoder network with parameters ϕ. It takes preprocessed raw audio x as input and outputs a frame-level melody representation sequence $m_e = E_\phi(x)$, where $m_e \in \mathbb{R}^{T \times D_m}$, T is the number of time frames, and D_m is the dimension of the melody representation.

To ensure that this learning extractor captures realistic and effective melody information, we introduce a distillation constraint based on KL divergence. We employ a teacher model[1] $E_{teacher}$, pre-trained on a small-scale music dataset with accurate MIDI annotations and then frozen, to provide stable melody supervision openvpi [2022]. The melody representation extracted by this teacher model, $m_{teacher} = E_{teacher}(x)$, is treated as a "soft label". We guide the learning process of the student extractor E_ϕ by minimizing the Kullback-Leibler divergence Kullback and Leibler [1951] between the output of the student extractor E_ϕ and the teacher output $m_{teacher}$. This constraint loss function is defined as follows:

$$\mathcal{L}_{KD} = D_{KL}(\text{Proj}(m_e) \| m_{teacher})$$

where $\text{Proj}(\cdot)$ is a projection layer used to align the dimension of m_e with that of $m_{teacher}$. This loss function ensures that the student model, while maintaining flexibility, produces melodic semantics consistent with a relatively accurate pre-trained model.

During joint training, the parameters (ϕ, θ) of the melody extractor E_ϕ and the singing synthesis model G_θ are updated together. The training loss of the singing synthesis model (the denoising loss of the diffusion model) provides direct gradient feedback to E_ϕ regarding "what kind of melodic features are beneficial for the synthesis task". This design enables the melody extractor to adaptively optimize its representations, no longer merely pursuing general melody extraction accuracy but focusing on the melodic features that most contribute to singing generation, thereby enhancing both its own performance and the end-to-end performance of the entire system.

Melody-Content Alignment Constraint Based on CKA. In conditional generation models, ensuring high correlation between the generated content and the given condition is crucial. To further strengthen the guiding role of the melodic condition for the generated song, we introduce a CKA loss Davari et al. [2022] to explicitly constrain the correlation between the internal representations during song synthesis and the input melody representation. We use a flow matching-based model as the backbone of the song synthesis model G_θ, which learns a highly structured latent space when processing data. Let z_l be the feature representation of an intermediate layer in the flow model, which encodes the semantic and acoustic information of the song being generated.

CKA is a reliable metric for measuring the similarity between two different representation spaces. We use linear CKA to measure the correlation between the melody representation m_e and the flow model's internal feature z_l. Given two feature sets m_e and z_l, their linear CKA is calculated as follows: (1) Compute covariance matrices: $K = m_e m_e^T$ and $L = z_l z_l^T$. (2) Center the covariance matrices. (3) The CKA value is given by the normalized form of the Hilbert-Schmidt Independence Criterion:

$$\text{CKA}(K, L) = \frac{\|K^T L\|_F^2}{\|K^T K\|_F \|L^T L\|_F}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Our goal is to maximize the CKA value between m_e and z_l, i.e., to minimize the following CKA loss:

$$\mathcal{L}_{CKA} = 1 - \text{CKA}(m_e, z_l)$$

The introduction of this loss function encourages the song synthesis model, during the generation process, to maintain high structural consistency between its internal data flow (corresponding to the timbre, rhythm, etc., of the song) and the externally provided melody condition. This is equivalent to imposing a correlation inductive bias within the model, effectively preventing the problem of the generated result deviating from the input melody, thereby significantly improving the accuracy and robustness of melodic guidance.

[1] https://github.com/openvpi/SOME

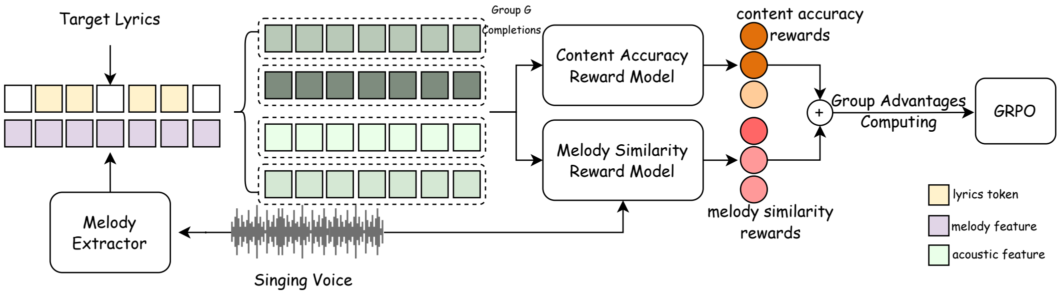

Figure 2: Multi-objective alignment for both intelligibility and melody similarity. This figure demonstrates how we utilize singing voice as melody prompts during post-training of YingMusic-Singer.

Overall Training Objective. The total training loss of our model is the weighted sum of the aforementioned losses:

$$\mathcal{L}_{Total} = \mathcal{L}_{Diffusion} + \lambda_{KD} \cdot \mathcal{L}_{KD} + \lambda_{CKA} \cdot \mathcal{L}_{CKA}$$

where $\mathcal{L}_{Diffusion}$ is the standard denoising score matching loss (mean squared error loss) of the diffusion model, and λ_{KD} and λ_{CKA} are hyperparameters used to balance the importance of each task.

3.3 Post Training

Motivation. After pre-training, we observed that the DiT model exhibited strong generalization ability on datasets of different styles. However, its stability remains insufficient, especially in fundamental capabilities such as melody consistency and lyric alignment, where further improvement is still possible. Furthermore, in downstream applications, the model may encounter input distributions not covered during training—for example, in the context of SVS, the model receives melody markers extracted from MIDI-rendered singing voices (as shown in Figure 2) and uses them in conjunction with rhythmic cues from speech input. Based on these considerations, this study introduces a post-training stage aimed at improving the model's controllability over lyrics and melody, and enhancing its robustness to diverse audio data types. Training employs the reinforcement learning-based GRPO method to improve the performance of the singing voice conversion model (Figure 2). This method focuses on strategy optimization. By designing and introducing reward functions for lyric alignment and melody consistency, the model can iteratively update toward better singing performance during the post-training stage. This enables the model to preserve semantic clarity while maintaining consistent melodic structure, thereby producing higher-quality and more controllable singing outputs.

In this stage, we refine the model using reinforcement learning to directly optimize non-differentiable perceptual objectives. To enable stochastic policy optimization, we reinterpret the deterministic flow dynamics as a stochastic policy by injecting a small amount of noise into the ODE trajectory, similar to techniques used in Flow-GRPO Liu et al. [2025]:

$$dx_t = v_\theta(x_t, t)\, dt + \sigma_t\, dw_t, \tag{1}$$

where w_t is a standard Wiener process and σ_t controls the injected stochasticity. We adopt a simple monotonic schedule $\sigma_t = a\sqrt{t/(1-t)}$, where a determines the overall noise level. To avoid the credit-assignment ambiguity associated with full SDE sampling, we inject randomness at a single uniformly sampled timestep while keeping all remaining steps deterministic Team et al. [2025].

For each prompt c, G completions are generated and normalized rewards are used to compute the advantage $A^{(i)}$, leading to the training objective

$$J(\theta) = \mathbb{E}_{c,\, t',\, \{x^i\}_{i=1}^{G} \sim \pi_{\text{old}}} \left[\frac{1}{G} \sum_{i=1}^{G} \left(r^i_{t'}(\theta)\, \hat{A}^i - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})_{t'} \right) \right], \tag{2}$$

where $r^i_{t'}(\theta)$ is the policy ratio correcting for off-policy sampling, t' denotes the uniformly sampled timestep at which stochasticity is injected, and π_θ and π_{ref} represent the current policy and the frozen reference policy, respectively. This selective-noise scheme preserves exploration while substantially improving optimization stability.

Content Accuracy Reward. To assess the articulation clarity and content accuracy of the converted singing, we first employ an ASR model to compute a reward grounded in the word error rate (WER). Specifically, given the transcription $\hat{y}$ of the generated singing and the corresponding reference text y, we calculate the WER as follows:

$$\text{WER} = \frac{S + D + I}{N}, \tag{3}$$

Specifically, S denotes the number of substitution errors, D denotes the number of deletion errors, and I denotes the number of insertion errors, while N is the total number of words in the reference text. To ensure that the reward correlates positively with better recognition outcomes, we define the content accuracy reward R_{con} as follows:

$$R_{\text{con}} = 1 - \text{WER}. \tag{4}$$

Melodic Similarity Reward. We use the Pearson correlation coefficient between the generated pitch contour and the reference pitch contour (F0) as the melodic similarity reward. Specifically, we first extract the pitch trajectories of the two audio segments. Then, we compute the similarity only on voiced frames (i.e., frames with non-zero F0) to avoid interference from silence and invalid regions. For each sample, we denote the generated pitch sequence as $f^{(g)} = \{f_1^{(g)}, f_2^{(g)}, \ldots, f_N^{(g)}\}$ and the target pitch sequence as $f^{(t)} = \{f_1^{(t)}, f_2^{(t)}, \ldots, f_N^{(t)}\}$. The melodic similarity reward R_{mel} is then defined as the Pearson correlation between the two sequences:

$$R_{\text{mel}}(f^{(g)}, f^{(t)}) = \frac{\sum_{i=1}^{N} \left(f_i^{(g)} - \bar{f}^{(g)}\right)\left(f_i^{(t)} - \bar{f}^{(t)}\right)}{\sqrt{\sum_{i=1}^{N} \left(f_i^{(g)} - \bar{f}^{(g)}\right)^2 \sum_{i=1}^{N} \left(f_i^{(t)} - \bar{f}^{(t)}\right)^2}}. \tag{5}$$

The reward reflects the consistency of the melodic direction between the two audio segments. The higher the correlation coefficient, the closer the pitch change trend of the generated speech is to the target, thus obtaining a higher melodic similarity reward.

Multi-Objective Optimization. The final multi-objective reward for the sample i-th is:

$$R^i = \sum_k w_k R_k^i, \tag{6}$$

and the group-relative advantage is:

$$\hat{A}^i = \frac{R^i - \mu}{\sigma}, \tag{7}$$

where μ and σ denote the mean and standard deviation over the group for the same conditioning prompt, and w_k denotes the weighting coefficient associated with the k-th reward term in the multi-objective formulation. This multi-objective RL framework enables direct optimization of lyric alignment and melody consistency, complementing the supervised objectives in preceding stages. In our implementation, the weights for both the melodic similarity reward and the content accuracy reward are set to 1.

4 Experiments

4.1 Implementation

We initialize the parameters of our DiT-based decoder from an existing DiT-based TTS model, F5-TTS Chen et al. [2025], to expedite model convergence and improve generalization. During training,

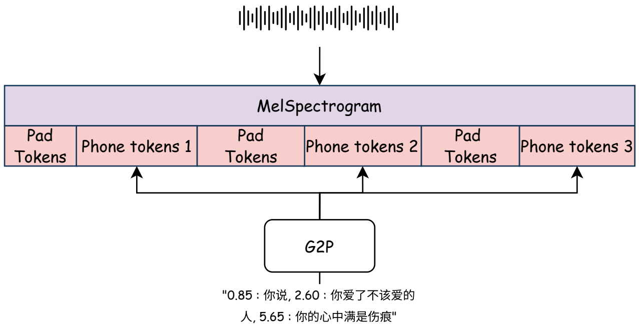

Figure 3: Lyrics go through G2P and are placed at the positions corresponding to their timestamps

Table 1: Comparison of Singing Voice Synthesis and Editing tasks. Best results are highlighted in **bold**.

Task	Method	Objective Metrics			Aesthetic Scores			
		WER (%)↓	SIM (%)↑	FPC (%)↑	CE↑	CU↑	PC↑	PQ↑
Zero-Shot Singing Voice Synthesis	TCSinger Zhang et al. [2024a]	3.47	**94.41**	77.79	5.56	6.20	1.77	7.36
	Vevo Zhang et al. [2025b]	9.83	93.51	**87.96**	6.42	**6.76**	**1.84**	**7.60**
	Ours	**1.28**	93.95	81.28	**6.57**	6.68	1.72	7.58
Singing Voice Editing	***Lyrics Editing***							
	Vevo Zhang et al. [2025b]	29.89	**95.87**	83.47	6.28	**6.66**	**1.78**	**7.54**
	Ours	**16.58**	95.36	**89.53**	**6.31**	6.52	1.73	7.50
	Structural Editing							
	Vevo Zhang et al. [2025b]	30.63	**96.11**	89.27	6.28	**6.65**	**1.80**	**7.54**
	Ours	**18.44**	95.47	**90.34**	**6.38**	6.47	1.75	7.50
Zero-Shot Singing Voice Editing	***Lyrics Editing***							
	Vevo Zhang et al. [2025b]	67.31	93.46	**83.91**	6.32	6.64	**1.83**	7.54
	Ours	**15.18**	**93.75**	82.84	**6.78**	**6.71**	1.77	**7.59**
	Structural Editing							
	Vevo Zhang et al. [2025b]	73.97	93.53	**82.52**	6.32	**6.67**	**1.83**	7.53
	Ours	**12.62**	**93.77**	81.19	**6.54**	6.66	1.75	**7.57**

the lyrics are padded following the DiffRhythm Ning et al. [2025] methodology (illustrated in Fig. 3). At inference, detailed timestamps are not employed; instead, a single timestamp is used to separate the prompt from the generated content. Our DiT architecture follows that of F5-TTS, consisting of 12 decoder layers with a hidden size of 1024 and 16-head self-attention mechanisms (each head of 64 dimensions), amounting to a total of 0.3B parameters. To facilitate classifier-free guidance (CFG), we apply 20% dropout independently to both lyrics and audio prompts. The diffusion process employs an Euler ODE solver with 32 sampling steps and a CFG scale of 2 during inference. In the post-training stage, we use the FireRedASR model Xu et al. [2025] to compute the content accuracy reward and the RMVPE model Wei et al. [2023] to extract the pitch trajectories and compute the melodic similarity reward. All models are trained using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The learning rate is set to 1×10^{-4}, with a linear warm-up over the first 2k steps followed by a linear decay for the remaining training. Training is conducted on 8 A800 80GB GPUs with a batch size of 116,000 audio frames (approximately 0.35 hours) for a total of 110k steps.

4.2 Training Data

For singing voice training, we employed the data preparation pipeline introduced in DiffRhythm Ning et al. [2025] to curate a dataset consisting of 3.7K hours of Mandarin singing vocals. These vocals were extracted via source separation from publicly available songs collected from the Internet. All audio signals were converted to mono channels at a sampling rate of 24 kHz and segmented into clips of approximately 30 seconds each. (Note that, as the melody extraction teacher model requires a 44 kHz input, the audio supplied to this model was resampled accordingly.) To facilitate the proposed multi-objective alignment task, the dataset was further refined by filtering approximately 500 hours

Table 2: Performance comparison on Zero-Shot Singing Voice Editing on subjective evaluation.

Model	Zero-Shot Singing Voice Editing	
	N-CMOS	*Melody-MOS*
Vevo	-0.75 ± 0.12	1.62 ± 0.38
Ours	**0.00 ± 0.00**	**1.76 ± 0.35**

Table 3: Ablation study results. Best results are highlighted in **bold**.

Model	Objective Metrics			Aesthetic Scores			
	WER (%)↓	SIM (%)↑	FPC (%)↑	CE↑	CU↑	PC↑	PQ↑
Full	**15.18**	**93.75**	**82.84**	**6.78**	6.71	1.77	7.59
w/o post-training	16.75	93.31	76.64	6.42	6.51	**1.78**	7.46
w/o cka-alignment	16.49	93.41	82.73	6.60	**6.73**	1.73	**7.60**

of high-quality audio based on a combination of metrics, including DNSMOS Reddy et al. [2021], word error rate (WER), and aesthetic score Tjandra et al. [2025].

4.3 Evaluation Data

We construct the evaluation set under various settings. For singing voice synthesis, we randomly selected 60 audio clips from GTSinger Zhang et al. [2024b], which covers a broad spectrum of singing techniques, styles, and timbre types. For the zero-shot setting, we randomly chose five speakers from in-the-wild singing data that were not included in the training set. To evaluate singing voice editing, we built two dedicated datasets: one for modifying lyrics while preserving the original lyrical structure and character count, and another for modifying lyrics with changes to both structure and character count. All lyric modifications were generated using DeepSeek-V3.2[2], with each original song containing at least three variations.

4.4 Evaluation Metrics

Objective Metrics For objective evaluation, we employ several metrics to assess key aspects of the generated singing voices: intelligibility via Word Error Rate (WER, ↓), speaker similarity (SIM, ↑), and F0 correlation (FPC, ↑) Huang et al. [2023], Zhang et al. [2024c, 2025b,c]. WER is computed using the FireRedASR[3] model Xu et al. [2025]. For SIM, we measure the cosine similarity between WavLM-TDNN[4] speaker embeddings Chen et al. [2022] extracted from the generated samples and the corresponding reference audio.

Subjective Metrics For subjective evaluation, we employ the Comparative Mean Opinion Score (CMOS, scaled from -2 to 2, ↑) to measure several perceptual attributes of the generated samples: naturalness (N-CMOS). We also utilize the Melody-MOS metric introduced in Vevo2 Zhang et al. [2025c] (ranging from 1 to 3) to specifically assess melody-following capability, with the following scoring criteria: 1 indicates "unable to follow the melody," 2 corresponds to "roughly following the melody contour," and 3 represents "accurately following all melodic details."

4.5 Controllability in Zero-shot Singing Voice Synthesis and Editing

To evaluate the controllability of YingMusic-Singer over content and melody, we conducted a series of tasks including zero-shot Singing Voice Synthesis (SVS) and singing voice editing (encompassing both structural and lyrical modifications). In the zero-shot SVS task—specifically defined here as zero-shot timbre transfer—the model is conditioned on target lyrics, melodic notation (e.g., MIDI Zhang et al. [2024a]), and a reference audio waveform. The objective is to generate a singing voice that adheres to the target content and notes while preserving the reference timbre. Uniquely,

[2] https://chat.deepseek.com/

[3] https://huggingface.co/FireRedTeam/FireRedASR-AED-L

[4] https://huggingface.co/microsoft/wavlm-base-sv

YingMusic-Singer renders MIDI-like information into a reference singing melody, effectively treating the task as a melody-to-singing synthesis.

Table 1 summarizes the comparison between our model and two baseline systems, TCSinger Zhang et al. [2024a] and Vevo Zhang et al. [2025d]. Across nearly all tasks, the post-trained YingMusic-Singer demonstrates superior performance in lyric and melody transcription, achieving the lowest Word Error Rate (WER) and competitive F0 Pearson Correlation (FPC). Subjective evaluations (Table 2) further indicate that YingMusic-Singer achieves higher N-CMOS scores than Vevo, reflecting superior naturalness. Although the FPC is slightly lower than that of Vevo, it still exceeds the 80 % level, indicating a strong ability to follow the target melody. Furthermore, the model achieves a significantly lower WER, indicating that it successfully adheres to the melodic contour while maintaining superior content fidelity. This performance underscores an effective post-training strategy that balances the often-competing objectives of content accuracy, naturalness, and melody adherence.

Furthermore, in the structural editing task—where lyrics and overall sentence structures are significantly altered—both Vevo and YingMusic-Singer maintain low WER and strong F0 correlation. This suggests that in singing voice editing, preserving the melodic direction of phonetic units is more critical than strictly maintaining the original lyrical count or sentence structure.

4.6 Effectiveness of Training Strategies

To evaluate the contribution of different modules in YingMusic-Singer, we conducted ablation studies focusing on the cka-alignment mechanism and the post-training procedure. Specifically, beginning with the pre-trained YingMusic-Singer-base model, we introduced a post-trained version optimized using an intelligibility reward (measured by WER) and a Melody Similarity Reward (measured by FPC Score). For the zero-shot singing voice editing task, we designed two subjective evaluation metrics: given target lyrics, a target melody (extracted from another singing voice), and a generated singing voice, participants assessed whether the generated output accurately followed the lyrics (lyrics accuracy) and the melody (melody accuracy). Experimental results indicate that each module contributes positively to the final outcome, as summarized in Table 3. Removing the post-training stage led to a decline in nearly all metrics (WER: 15.18% → 16.75%; FPC: 82.84% → 76.64%). Additional observations include that cka-alignment facilitates faster convergence to melody-related guidance during the early training phase. However, the corresponding loss weights must be gradually reduced during training; otherwise, although a strong F0 correlation is achieved, it results in increased WER. Notably, when both rewards are applied together, the model demonstrates not only improved melody-following capability but also enhanced text-following performance. We hypothesize that this improvement stems from reinforced melody modeling, which in turn promotes clearer pronunciation and higher intelligibility. These findings further validate the advantages of our proposed multi-objective post-training strategy. Overall, the model achieves a better balance between content accuracy and melodic adherence.

5 Conclusion

In this paper, we presented a novel framework for melody-driven Singing Voice Synthesis that addresses scalability challenges by eliminating the dependence on precise phoneme alignment and manual annotations. By integrating a Diffusion Transformer with automated melody extraction and Flow-GRPO reinforcement learning, our system achieves superior melodic stability and pronunciation clarity using only weakly-aligned data. Experimental results confirm that our approach outperforms existing baselines in both objective metrics and subjective quality, particularly in zero-shot lyric replacement scenarios, offering a practical solution for democratized content creation. Building on these findings, our future work will focus on three key advancements. First, we aim to enhance multilingual support by adopting unified phoneme representations to enable high-quality cross-lingual synthesis. Second, we will prioritize sound quality improvement by leveraging advanced neural vocoders and latent diffusion techniques to achieve high-sampling-rate, studio-grade fidelity. Finally, we plan to improve generalization and expressiveness by scaling training to diverse, "in-the-wild" datasets and disentangling vocal attributes, allowing for fine-grained control over emotion and style across different musical contexts.

References

Jiawei Chen, Xu Tan, Jian Luan, Tao Qin, and Tie-Yan Liu. Hifisinger: Towards high-fidelity neural singing voice synthesis. *arXiv preprint arXiv:2009.01776*, 2020.

Zewang Zhang, Yibin Zheng, Xinhui Li, and Li Lu. Wesinger: Data-augmented singing voice synthesis with auxiliary losses. *arXiv preprint arXiv:2203.10750*, 2022a.

Yongmao Zhang, Jian Cong, Heyang Xue, Lei Xie, Pengcheng Zhu, and Mengxiao Bi. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7237–7241. IEEE, 2022b.

Zhiqing Hong, Chenye Cui, Rongjie Huang, Lichao Zhang, Jinglin Liu, Jinzheng He, and Zhou Zhao. Unisinger: Unified end-to-end singing voice synthesis with cross-modality information matching. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7569–7579, 2023.

Yin-Ping Cho, Fu-Rong Yang, Yung-Chuan Chang, Ching-Ting Cheng, Xiao-Han Wang, and Yi-Wen Liu. A survey on recent deep learning-driven singing voice synthesis systems, 2021. URL https://arxiv.org/abs/2110.02511.

Peiling Lu, Jie Wu, Jian Luan, Xu Tan, and Li Zhou. Xiaoicesing: A high-quality and integrated singing voice synthesis system. *arXiv preprint arXiv:2006.06261*, 2020.

Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems*, 35: 6914–6926, 2022c.

Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3945–3954, 2021.

Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. *arXiv preprint arXiv:2201.07429*, 2022.

Zhiqing Hong, Rongjie Huang, Xize Cheng, Yongqi Wang, Ruiqi Li, Fuming You, Zhou Zhao, and Zhimeng Zhang. Text-to-song: Towards controllable music generation incorporating vocals and accompaniment, 2024. URL https://arxiv.org/abs/2404.09313.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Ziqian Ning, Huakang Chen, Yuepeng Jiang, Chunbo Hao, Guobin Ma, Shuai Wang, Jixun Yao, and Lei Xie. Diffrhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. *arXiv preprint*, abs/2503.01183, 2025.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

Yongmao Zhang, Jian Cong, Heyang Xue, Lei Xie, Pengcheng Zhu, and Mengxiao Bi. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7237–7241. IEEE, 2022d.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11020–11028, 2022.

Kehan Sui, Jinxu Xiang, and Fang Jin. Smoothsinger: A conditional diffusion model for singing voice synthesis with multi-resolution architecture. *arXiv preprint arXiv:2506.21478*, 2025.

Yu Zhang, Wenxiang Guo, Changhao Pan, Dongyu Yao, Zhiyuan Zhu, Ziyue Jiang, Yuhan Wang, Tao Jin, and Zhou Zhao. Tcsinger 2: Customizable multilingual zero-shot singing voice synthesis. *arXiv preprint arXiv:2505.14910*, 2025a.

Junchuan Zhao, Wei Zeng, Tianle Lyu, and Ye Wang. Comelsinger: Discrete token-based zero-shot singing synthesis with structured melody control and guidance. *arXiv preprint arXiv:2509.19883*, 2025.

Chen Shen, Lu Zhao, Cejin Fu, Bote Gan, and Zhenlong Du. Transinger: Cross-lingual singing voice synthesis via ipa-based phonetic alignment. *Sensors*, 25(13):3973, 2025.

Jinzheng He, Jinglin Liu, Zhenhui Ye, Rongjie Huang, Chenye Cui, Huadai Liu, and Zhou Zhao. Rmssinger: Realistic-music-score based singing voice synthesis. *arXiv preprint arXiv:2305.10686*, 2023.

Gyeong-Hoon Lee, Tae-Woo Kim, Hanbin Bae, Min-Ji Lee, Young-Ik Kim, and Hoon-Young Cho. N-singer: A non-autoregressive korean singing voice synthesis system for pronunciation enhancement, 2022. URL https://arxiv.org/abs/2106.15205.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Rui Liu, Berrak Sisman, and Haizhou Li. Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability. *arXiv preprint arXiv:2104.01408*, 2021.

Jingyi Chen, Ju-Seung Byun, Micha Elsner, and Andrew Perrault. Dlpo: Diffusion model loss-guided reinforcement learning for fine-tuning text-to-speech diffusion models. *arXiv preprint arXiv:2405.14632*, 2024.

Xiaohui Sun, Ruitong Xiao, Jianye Mo, Bowen Wu, Qun Yu, and Baoxun Wang. F5r-tts: Improving flow-matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*, 2025a.

Xiaohui Sun, Ruitong Xiao, Jianye Mo, Bowen Wu, Qun Yu, and Baoxun Wang. F5r-tts: Improving flow-matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*, 2025b.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of cka as a similarity measure in deep learning, 2022. URL https://arxiv.org/abs/2210.16156.

openvpi. Some: Singing-oriented midi extractor. https://github.com/openvpi/SOME, 2022. Accessed: [2025-11-26].

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

Yu Zhang, Ziyue Jiang, Ruiqi Li, Changhao Pan, Jinzheng He, Rongjie Huang, Chuxin Wang, and Zhou Zhao. Tcsinger: Zero-shot singing voice synthesis with style transfer and multi-level style control. In *EMNLP*, pages 1960–1975. Association for Computational Linguistics, 2024a.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, and Mingbo Ma. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *ICLR*. OpenReview.net, 2025b.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *ACL (1)*, pages 6255–6271. Association for Computational Linguistics, 2025.

Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration, 2025. URL https://arxiv.org/abs/2501.14350.

Haojie Wei, Xueke Cao, Tangpeng Dan, and Yueguo Chen. Rmvpe: A robust model for vocal pitch estimation in polyphonic music. In *INTERSPEECH 2023*, interspeech$_2$023, $page$5421$\breve{}$5425.$ISCA, August$2023. doi : . URL http://dx.doi.org/10.21437/Interspeech.2023-528.

Chandan K A Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors, 2021. URL https://arxiv.org/abs/2010.15258.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound, 2025. URL https://arxiv.org/abs/2502.05139.

Yu Zhang, Changhao Pan, Wenxiang Guo, Ruiqi Li, Zhiyuan Zhu, Jialei Wang, Wenhao Xu, Jingyu Lu, Zhiqing Hong, Chuxin Wang, Lichao Zhang, Jinzheng He, Ziyue Jiang, Yuxin Chen, Chen Yang, Jiecheng Zhou, Xinyu Cheng, and Zhou Zhao. Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. In *NeurIPS*, 2024b.

Wen-Chin Huang, Lester Phillip Violeta, Songxiang Liu, Jiatong Shi, and Tomoki Toda. The singing voice conversion challenge 2023. In *ASRU*, pages 1–8. IEEE, 2023.

Xueyao Zhang, Zihao Fang, Yicheng Gu, Haopeng Chen, Lexiao Zou, Junan Zhang, Liumeng Xue, and Zhizheng Wu. Leveraging diverse semantic-based audio pretrained models for singing voice conversion. In *SLT*. IEEE, 2024c.

Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu. Vevo2: Bridging controllable speech and singing voice generation via unified prosody learning, 2025c. URL https://arxiv.org/abs/2508.16332.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, and Mingbo Ma. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *The Thirteenth International Conference on Learning Representations*, 2025d. URL https://openreview.net/forum?id=anQDiQZhDP.