ZEROOPERA: LOW-RESOURCE PEKING OPERA SINGING VOICE SYNTHESIS WITH ZERO-SHOT TIMBRE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent singing voice synthesis (SVS) systems have achieved strong performance, but Peking Opera synthesis remains difficult because it requires domain-specific pronunciation rules and has only limited annotated data. We propose **ZeroOpera**, a low-resource Peking Opera SVS framework that supports zero-shot timbre transfer from an acoustic prompt. Instead of fine-tuning the foundation model, ZeroOpera keeps it frozen and uses its prompt-conditioned mel-spectrogram as an acoustic prior. A trainable decoder reconstructs regions that require opera-specific pronunciations and enhances the full spectrogram. We further introduce Technique-driven Phonological Modulation (TPM). TPM converts lexicon-derived phonological category labels into continuous feature-level controls. By separating Standard Mandarin pinyin from opera-specific pronunciation categories, TPM can handle pinyin–category combinations that are not observed during adaptation. This design does not require an expanded pinyin vocabulary or manually aligned opera-specific phoneme annotations. For evaluation, we develop an automatic G2P pipeline and a Peking Opera ASR model that recognizes pinyin syllables. On target passages from a Peking Opera singer excluded from opera-domain adaptation, ZeroOpera reduces Phoneme Error Rate by 19.4% relative to the frozen foundation model. It also improves timbre-similarity MOS by 0.43 when using unseen utterances from singers included in adaptation, and by 0.30 when using cross-domain prompts from 50 pop singers excluded from opera-domain adaptation. These results show that ZeroOpera improves pronunciation and operatic expressiveness while preserving the timbre specified by the prompt under low-resource domain adaptation.

1 INTRODUCTION

Deep learning has advanced singing voice synthesis (SVS) (Pan et al., 2025), yet Peking Opera synthesis remains challenging for two reasons. First, its domain-specific phonological rules, including *Jian–Tuan* and *Shang-kou-zi*, differ from Standard Mandarin and complicate text-to-acoustic modeling. Second, building such a corpus requires expert knowledge, limiting the availability of high-quality annotated data.

Existing Peking Opera SVS systems often rely on domain-specific or singer-dependent training, as in OperaSinger (Zhou et al., 2023). Limited opera data may restrict their prompt-based timbre transfer to unseen singers. Opera-specific pronunciations are also commonly represented by dedicated phoneme symbols, which are sparsely observed and may make it difficult to generalize to unseen combinations. In particular, a model may encounter an unseen pairing of a Mandarin pinyin syllable and an opera-specific pronunciation category, even if each has been observed separately. We refer to such pairings as *pinyin–category combinations*.

To improve generalization, we explore adapting zero-shot pop SVS systems to Peking Opera synthesis (Zheng et al., 2025; Zhang et al., 2024a;b; Qian et al., 2026; Zhang et al., 2025). Trained on substantially larger and more diverse corpora, recent models such as SoulX-Singer (Qian et al., 2026) provide strong prompt-based timbre cloning and stable acoustic priors. However, their predominantly pop-oriented training does not capture the domain-specific phonological rules of Peking

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

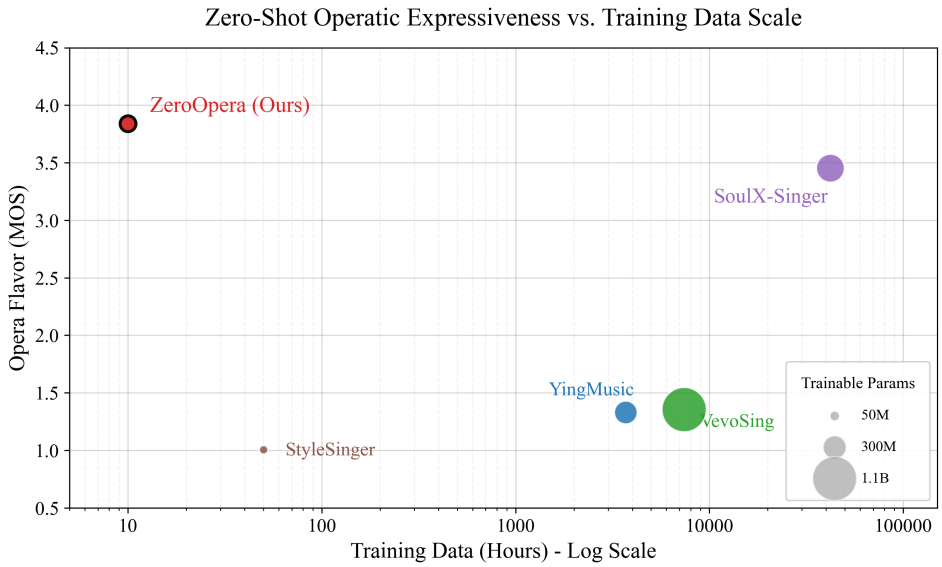


Figure 1: Opera-domain adaptation data and subjective Opera Flavor (OF) under cross-domain unseen-singer prompts. ZeroOpera achieves the highest OF among the evaluated systems using 7.58 hours of opera adaptation data. Results are descriptive rather than a controlled data-efficiency comparison because the systems differ in architecture, pre-training, control interface, and vocoder.

Opera. Directly adapting these models creates a trade-off. Full fine-tuning may weaken the model's ability to preserve the timbre specified by the prompt under low-resource adaptation, whereas parameter-efficient updates may better preserve the pre-trained model but do not explicitly provide localized control over opera-specific pronunciation regions. In our experiments, the evaluated full-fine-tuning and LoRA configurations exhibit different trade-offs between pronunciation adaptation and prompt-timbre preservation. We therefore investigate an alternative approach that adapts the generated acoustic representation while leaving the foundation model unchanged.

To address these challenges, we propose **ZeroOpera**, a low-resource Peking Opera adaptation framework that supports zero-shot prompt-timbre transfer through prior-guided acoustic inpainting and enhancement. Here, zero-shot refers to transferring the timbre of a prompt singer excluded from opera-domain adaptation, rather than adapting to the Peking Opera domain without opera recordings. ZeroOpera contains two core designs. First, rather than updating the foundation model, ZeroOpera uses its generated mel-spectrogram as a structural prior. A trainable inpainting stream reconstructs regions requiring opera-specific pronunciations while enhancing the unmasked acoustic context, thereby adapting phonological realization while retaining the prompt-based timbre capability of the frozen model. Second, we introduce **Technique-driven Phonological Modulation (TPM)**, which maps automatically generated character-level phonological labels into continuous control signals that modulate the base linguistic representations. By modeling technique effects as compositional feature modulation rather than expanding the base pinyin-token inventory, TPM is designed to transfer category-conditioned transformations to base-pinyin–category combinations that are absent from the adaptation set. It does so without requiring manually aligned opera-specific phoneme annotations. As shown in Figure 1, ZeroOpera achieves strong zero-shot operatic expressiveness using only 7.58 hours of domain-specific training data.

To support rigorous evaluation, we curate an 8.35-hour high-fidelity Peking Opera corpus from the Mei-school *Qingyi* tradition, together with expert-verified opera-pronunciation transcriptions. These transcriptions serve only as evaluation references and are not used as Pinyin-syllable supervision for ZeroOpera. We also develop an automated G2P pipeline and a Pinyin-syllable Peking Opera ASR model for objective pronunciation assessment. Our main contributions are:

- We introduce **ZeroOpera**, a low-resource Peking Opera SVS framework with zero-shot prompt-timbre transfer. It adapts the output of a frozen pop SVS model through prior-guided acoustic inpainting and enhancement, avoiding direct updates to the foundation model.

- We propose **Technique-driven Phonological Modulation (TPM)**, which factorizes opera-specific pronunciation control into base linguistic representations and continuous category-conditioned feature modulation. This design provides a factorized alternative to expanding the base pinyin-token inventory. Experiments on a held-out singer provide preliminary evidence of improved generalization to base-pinyin–category combinations that are absent from the adaptation set.

- We establish a Peking Opera evaluation benchmark with expert-verified G2P annotations, an automated G2P pipeline, and a Pinyin-syllable ASR model for objective pronunciation evaluation.

2 Related Work

Zero-Shot Singing Voice Synthesis. Zero-shot SVS synthesizes vocals for unseen singers using short reference prompts. Earlier systems, such as StyleSinger (Zhang et al., 2024a) and TCSinger (Zhang et al., 2024b), focused on disentangling timbre from linguistic and prosodic factors. More recent systems, including YingMusic-Singer (Zheng et al., 2025), Vevo (Zhang et al., 2025), and SoulX-Singer (Qian et al., 2026), leverage larger and more diverse training corpora together with autoregressive or flow-based generative architectures (Lipman et al., 2022). Although these models provide strong prompt-based timbre cloning, their predominantly general-purpose or pop-oriented training does not explicitly model Peking Opera phonology. ZeroOpera instead preserves a pretrained model as a frozen acoustic prior and performs opera-specific adaptation through conditional inpainting and phonological modulation.

Chinese Opera Synthesis. Chinese opera synthesis involves flexible rhythm, prolonged vowels, and genre-specific articulation. Prior work addresses score–performance mismatch and melody-aware modeling in Peking Opera (Wu et al., 2020; Zhou et al., 2023), combines speech pre-training with manual annotations for low-resource Gezi Opera synthesis (Zheng et al., 2024), and improves local acoustic modeling on Chinese pop and Hokkien Gezi Opera data (Bai et al., 2023). UniVoice (Zhou et al., 2026) further unifies speech, singing, and Hokkien Gezi Opera synthesis through factorized content, melody, and timbre conditioning. However, these approaches do not specifically address prompt-based zero-shot timbre transfer with explicit Peking Opera phonological control. In contrast, ZeroOpera adapts a frozen zero-shot SVS model using Standard Mandarin pinyin together with automatically derived phonological-category labels, without introducing opera-surface pinyin tokens or requiring manually aligned opera-specific phoneme annotations.

3 Preliminaries

Prompt-based Zero-Shot SVS. Given musical conditions $C = \{p, f_0, n_{\text{type}}\}$ and an acoustic prompt Mel_{prompt}, prompt-based SVS generates a mel-spectrogram that follows the linguistic and melodic conditions while preserving the prompt timbre. Here, p, f_0, and n_{type} denote the tone-bearing Standard Mandarin pinyin tokens, target pitch contour, and note attributes, respectively. Throughout this paper, p is obtained from the lyric characters using a Standard Mandarin G2P conversion and is shared by the frozen SoulX-Singer prior model and the ZeroOpera content encoder. In particular, neither model receives opera-surface pinyin forms such as *zian*, *bo*, or *guo* as linguistic input tokens. Opera-specific pronunciation information is instead provided explicitly through the character-level phonological-category labels described below. In this work, *zero-shot* refers specifically to prompt-based timbre transfer when the prompt singer is excluded from opera-domain adaptation. It does not mean that the system is trained or adapted without Peking Opera data. Accordingly, the cross-domain pop-prompt condition is an unseen-singer zero-shot setting, whereas the in-domain opera-prompt condition uses seen singers but unseen prompt utterances.

Conditional Flow Matching. Our acoustic decoder follows conditional flow matching (CFM) (Lipman et al., 2022), which learns a conditional vector field that transports samples from a simple prior to the target acoustic distribution. ZeroOpera determines the flow endpoint through stochastic routing (Section 4.1) and extends the standard CFM objective with region-aware weighting (Section 4.3).

Peking Opera Phonological Categories. Peking Opera contains domain-specific pronunciation patterns, including the historical *Jian–Tuan* distinction and *Shang-kou-zi* vowel variants. Rather than expanding the pinyin-token inventory, we assign each lyric character a label $l_i \in \{0, 1, 2, 3\}$, denoting Standard, *Jian*, *Tuan*, and *Shang-kou-zi*, respectively. The labels are automatically de-

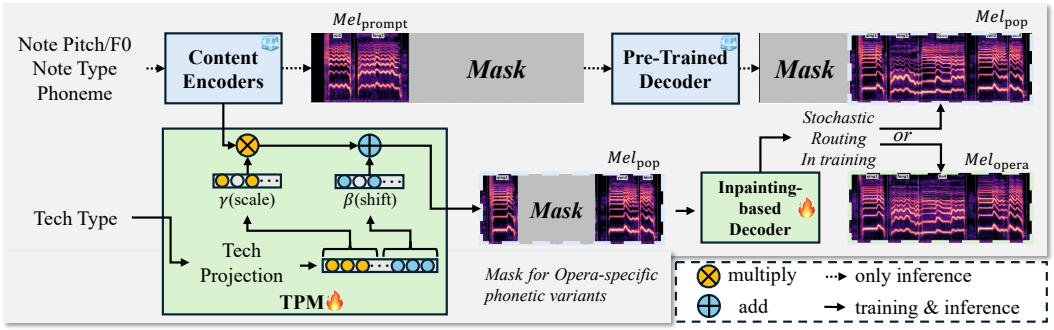


Figure 2: Overview of ZeroOpera. A frozen SVS model generates a prompt-conditioned acoustic prior. Opera-specific phonological regions are masked and reconstructed by a trainable flow decoder conditioned on TPM-modulated linguistic features. Stochastic routing alternates between pop-to-opera and masked identity reconstruction during training; only the pop-to-opera route is used at inference.

rived from the input lyrics using an expert-verified pronunciation lexicon during both training and inference; no recording-derived oracle labels are used at test time. Importantly, the lexical lookup supplies only the four-way category label to ZeroOpera: the linguistic token sequence remains the Standard Mandarin pinyin sequence p. Thus, opera-surface pronunciation forms are not added to the pinyin vocabulary and are not provided as input tokens to either SoulX-Singer or ZeroOpera. Appendix A provides the automatic annotation rules and linguistic examples.

4 Methodology

Figure 2 illustrates ZeroOpera. A frozen zero-shot SVS model first generates a prompt-conditioned pop-style mel-spectrogram as an acoustic prior. A trainable inpainting decoder then reconstructs regions associated with opera-specific phonological categories and enhances the complete acoustic output. Technique-driven Phonological Modulation (TPM) injects character-level category labels into the linguistic representation without expanding the pinyin-token inventory.

4.1 Prior-Guided Acoustic Inpainting

Acoustic Prior and Phonology-Aware Masking. Given musical conditions C and an acoustic prompt, the frozen foundation model generates

$$Mel_{\text{pop}} = \mathcal{F}_{\text{base}}\left(C, Mel_{\text{prompt}}\right).$$

The generated prior provides prompt-dependent timbre, pitch, and temporal structure. It is treated as a fixed condition, and gradients are not propagated through $\mathcal{F}_{\text{base}}$.

Because the pinyin tokens, phonological labels, and character durations are aligned, we define the character-level mask as $m_i = \mathbb{I}[l_i \neq 0]$ and expand it to the frame level:

$$M = \text{Expand}\left(\{m_i\}_{i=1}^{N}, d\right).$$

Thus, frames belonging to *Jian*, *Tuan*, or *Shang-kou-zi* characters are masked, whereas Standard regions remain visible. The masked prior is

$$\widetilde{Mel}_{\text{pop}} = Mel_{\text{pop}} \odot (\mathbf{1} - M).$$

Reference character durations are used during training, whereas durations predicted by the frozen model are used at inference. No reference alignment or recording-derived phonological annotation is therefore required at inference.

The decoder receives the masked prior, binary mask, and linguistic conditions. Although the mask focuses editing on category-specific regions, the decoder predicts the complete spectrogram and can therefore enhance the global acoustic context.

Stochastic Routing. Always reconstructing masked regions toward an opera target could cause the decoder to associate masking itself with phonological modification. We therefore sample $r \sim \mathrm{Bernoulli}(0.5)$ and set the flow endpoint to

$$Y_r = \begin{cases} Mel_{\mathrm{opera}}, & r = 1, \\ Mel_{\mathrm{pop}}, & r = 0. \end{cases}$$

Both routes use the same masked prior. When $r = 1$, TPM is activated and the decoder learns pop-to-opera reconstruction. When $r = 0$, TPM is disabled and the decoder performs masked identity reconstruction toward the pop prior. The latter regularizes the model to preserve the base acoustic structure when category-specific modulation is absent. At inference, we set $r = 1$.

For flow generation, the noisy state, masked prior, and mask are concatenated:

$$z_\tau = \mathrm{Concat}\left(x_\tau, \widetilde{Mel}_{\mathrm{pop}}, M\right).$$

With endpoint $x_1 = Y_r$, the decoder predicts the conditional vector field $v_\theta(z_\tau, \tau, H_{\mathrm{frame}}, f_0, n_{\mathrm{type}})$. Details of the probability path are provided in Appendix C.

4.2 Technique-driven Phonological Modulation

TPM factorizes pronunciation control into a Standard Mandarin base-pinyin representation and a category-conditioned feature transformation. The content encoder receives only the tone-bearing Standard Mandarin pinyin sequence; opera-specific surface forms are not introduced as additional linguistic tokens. Let $H_{\mathrm{content}} \in \mathbb{R}^{N \times D}$ denote the character-level linguistic representation. For each phonological label l_i, TPM computes an embedding and projects it into feature-wise scale and shift vectors:

$$e_i = E_{\mathrm{tech}}(l_i), \qquad [\gamma_i; \beta_i] = P_{\mathrm{tech}}(e_i).$$

To coordinate TPM with stochastic routing, we define $a_i = r \cdot \mathbb{I}[l_i \neq 0]$ and apply FiLM-style modulation:

$$H_{\mathrm{mod},i} = H_{\mathrm{content},i} \odot (\mathbf{1} + a_i\gamma_i) + a_i\beta_i.$$

For Standard characters or identity-reconstruction examples, $a_i = 0$, so TPM reduces to the identity transformation. Consequently, for a character whose Standard Mandarin pinyin is *jian*, the base linguistic representation is derived from *jian* regardless of whether the character is assigned the Jian or Tuan category. The distinction is supplied only through the corresponding TPM label. Likewise, a Shang-kou-zi character with Standard Mandarin pinyin *bai* is represented by the base token *bai* together with its category label, rather than by an opera-surface token such as *bo*. We also zero-initialize the final projection of P_{tech}, ensuring that TPM starts from identity modulation.

Because each category transformation operates on the underlying content representation, its acoustic effect can depend on the represented syllable and context. This factorization avoids assigning a separate token to every base-pinyin–category combination and provides a mechanism for transferring category-conditioned transformations to combinations absent from the adaptation set. Whether this mechanism improves compositional generalization is evaluated empirically in Section 5.3.

TPM is applied before duration expansion:

$$H_{\mathrm{frame}} = \mathrm{LR}\left(H_{\mathrm{mod}}, d\right).$$

The resulting frame-level representation is aligned with the acoustic mask and conditions the inpainting decoder. Although TPM directly modulates only marked characters, supervision over the complete opera target enables global opera-domain acoustic adaptation.

4.3 Region-Aware Flow Matching

Given the route-dependent endpoint $x_1 = Y_r$, we sample $x_0 \sim \mathcal{N}(0, I)$ and $t \sim \mathcal{U}(0, 1)$. We construct the linear conditional flow-matching path as

$$x_t = (1 - t)\, x_0 + t x_1,$$

The corresponding target vector field is

$$u_t = \frac{\mathrm{d}x_t}{\mathrm{d}t} = x_1 - x_0.$$

The acoustic prior is provided as a conditioning input rather than as the initial state of the probability path. The inpainting decoder predicts

$$v_\theta = v_\theta\left(x_t, t, \widetilde{Mel}_{\text{pop}}, M, H_{\text{frame}}, f_0, n_{\text{type}}\right),$$

where $H_{\text{frame}} = \text{LR}\left(H_{\text{mod}}, d\right)$.

Because opera-specific regions are sparse, we assign a larger weight to the masked frames:

$$w_j = \lambda_{\text{keep}} + \left(\lambda_{\text{edit}} - \lambda_{\text{keep}}\right) M_j,$$

where $\lambda_{\text{edit}} = 5.0 \quad \text{and} \quad \lambda_{\text{keep}} = 1.0$. The region-aware flow-matching objective is

$$\mathcal{L}_{\text{RAFM}} = \mathbb{E}\left[\frac{\sum_{j=1}^{T} P_j w_j \left\|v_{\theta,j} - u_{t,j}\right\|_1}{\sum_{j=1}^{T} P_j w_j}\right],$$

where P_j is the valid-frame padding mask. This objective emphasizes the reconstruction of opera-specific regions while still preserving the surrounding acoustic structure.

5 Experiments

5.1 Setup

Opera Corpus and Data Split. We use an 8.35-hour, 44.1 kHz Peking Opera corpus containing 948 recordings from six professional Mei-school *Qingyi* singers. Expert-verified opera pronunciations are used only as evaluation references. During training and inference, character-level phonological labels are deterministically derived from the input lyrics using an expert-verified lexicon; no recording-derived oracle labels are used at inference. Corpus statistics and annotation procedures are provided in Appendix A and B.

Singer 3 is entirely excluded from adaptation, prompt construction, and ASR training. Its 44 recordings define the held-out target conditions and ground-truth references. From the remaining singers, we reserve 80 recordings exclusively as in-domain prompts and exclude them from adaptation; the remaining 824 recordings (7.58 hours) form the adaptation set. For each target condition, all systems receive the same lyrics, score, and frame-level melody through their native interfaces. The target waveform is used only to derive these controlled conditions and as an evaluation reference, never as an acoustic prompt or direct model input.

Prompt Conditions. We evaluate two settings. **In-domain opera prompts** use the 80 reserved recordings from singers represented in the adaptation set but with unseen prompt utterances; this evaluates timbre preservation for seen singers rather than unseen-singer generalization. **Cross-domain pop prompts** use 100 pop-singing segments from 50 singers excluded from both opera-domain adaptation and the foundation-model training data (25 Mandarin and 25 English singers, with two prompts each). This setting evaluates zero-shot transfer from unseen pop singers to Peking Opera.

For objective evaluation, every prompt is paired with all 44 held-out target conditions. Subjective evaluation uses the sampled pairs described in Appendix D.3 and summarized in Table 7. Throughout this paper, *zero-shot* refers to prompt-timbre transfer from singers excluded from opera-domain adaptation, not adaptation without opera recordings. SoulX-Singer is the frozen structural-prior model throughout.

Baselines. We compare ZeroOpera with StyleSinger, YingMusic-Singer, VevoSing, and frozen SoulX-Singer, and additionally construct SoulX-Singer-FFT and SoulX-Singer-LoRA from the same checkpoint and adaptation set. External systems retain their official inference pipelines and vocoders and are therefore compared at the system level. Within the SoulX-Singer group, ZeroOpera, FFT, LoRA, and the frozen baseline share the same foundation model, target F0, acoustic preprocessing, and Vocos waveform-generation pipeline. The adaptation variants are not matched in trainable parameters, architecture, training iterations, or optimization objective; therefore, this comparison evaluates the reported configurations rather than intrinsically ranking adaptation paradigms. Details are provided in Appendix C.

Table 1: Performance under in-domain opera prompts and cross-domain zero-shot pop prompts. Subjective scores are reported as MOS with 95% confidence intervals. Lower values are better for FFE and PER; higher values are better for AQ, PC, TS, and OF. Bold and underlined values indicate the best and second-best system results, respectively.

Model	FFE ↓	PER ↓	AQ ↑	PC ↑	TS ↑	OF ↑
In-Domain Opera Prompts						
GT	–	0.200	4.48 ± 0.08	4.47 ± 0.08	–	4.62 ± 0.07
StyleSinger	0.278	0.928	1.12 ± 0.06	1.01 ± 0.02	1.01 ± 0.02	1.00 ± 0.00
YingMusic	0.225	**0.211**	1.82 ± 0.10	2.46 ± 0.15	1.83 ± 0.13	1.36 ± 0.08
VevoSing	0.307	0.464	1.92 ± 0.12	1.88 ± 0.12	2.11 ± 0.15	1.39 ± 0.09
SoulX-Singer	<u>0.127</u>	0.289	<u>3.49 ± 0.11</u>	<u>3.68 ± 0.13</u>	<u>3.68 ± 0.12</u>	<u>3.69 ± 0.14</u>
ZeroOpera	**0.126**	<u>0.233</u>	$\mathbf{3.96 \pm 0.12}$	$\mathbf{3.94 \pm 0.13}$	$\mathbf{4.11 \pm 0.13}$	$\mathbf{4.07 \pm 0.14}$
Cross-Domain Zero-Shot Pop Prompts						
StyleSinger	0.500	0.718	1.19 ± 0.08	1.01 ± 0.02	1.05 ± 0.03	1.01 ± 0.01
YingMusic	0.162	<u>0.279</u>	1.95 ± 0.11	2.15 ± 0.14	1.61 ± 0.12	1.30 ± 0.08
VevoSing	0.401	0.604	2.09 ± 0.12	1.84 ± 0.13	2.61 ± 0.14	1.32 ± 0.08
SoulX-Singer	<u>0.156</u>	0.336	<u>3.23 ± 0.11</u>	<u>3.43 ± 0.12</u>	<u>3.17 ± 0.13</u>	<u>3.22 ± 0.18</u>
ZeroOpera	**0.149**	**0.263**	$\mathbf{3.57 \pm 0.14}$	$\mathbf{3.66 \pm 0.13}$	$\mathbf{3.47 \pm 0.15}$	$\mathbf{3.61 \pm 0.18}$

Evaluation Metrics. We measure melody fidelity using F0 Frame Error (FFE) and pronunciation using Phoneme Error Rate (PER). FFE is computed against the ground-truth singing, and PER is computed by our Peking Opera ASR against expert-verified pronunciation transcriptions. The ASR evaluation recordings are disjoint from its training data, and its held-out real-recording performance is reported in Appendix D.1. Because recognition remains challenging in this low-resource domain, PER is treated as a complementary objective metric rather than a replacement for perceptual evaluation.

Fifteen evaluators with professional Peking Opera experience rate Audio Quality (AQ), Pronunciation Clarity (PC), Timbre Similarity (TS), and Opera Flavor (OF) on five-point scales. We report MOS with 95% confidence intervals; the sampling, presentation, rating protocol, and confidence-interval calculation are detailed in Appendix D. We use listener-rated TS as the primary timbre metric because generic speaker embeddings can be confounded by the substantial pitch-register mismatch between pop prompts and Peking Opera targets; supporting analysis is provided in Appendix E.1.

5.2 Main Results

In-Domain Opera Prompts. As shown in Table 1, ZeroOpera achieves the best FFE and all four subjective scores among the evaluated systems, while reducing PER by 19.4% relative to the frozen SoulX-Singer. The simultaneous improvements in timbre similarity and Opera Flavor indicate that the proposed acoustic editing introduces opera-domain characteristics without compromising the prompt-conditioned structure. YingMusic obtains a lower aggregate PER but substantially weaker Opera Flavor. Since Standard characters constitute 68.61% of the corpus, aggregate PER does not isolate sparse opera-specific pronunciations; category-specific results are analyzed in Table 3. A representative case study comparing mel-spectrograms, ASR-recognized category-specific pronunciations, and signed F0 deviations is provided in Appendix E.3.

Cross-Domain Pop Prompts. Under unseen pop-singer prompts, ZeroOpera achieves the best result for every reported metric. Relative to the frozen foundation model, it reduces both PER and FFE while improving Timbre Similarity by 0.30 MOS and Opera Flavor by 0.39 MOS. These joint gains suggest that opera-domain adaptation does not simply overwrite the identity cues supplied by the prompt.

5.3 Adaptation Analysis and Ablations

The lower panel of Table 2 evaluates TPM, the acoustic prior, phonology-aware inpainting, region-aware flow matching, and stochastic routing. Each variant modifies only the corresponding component while retaining the remaining ZeroOpera configuration.

Table 2: Comparison of adaptation strategies and component ablations under the in-domain opera-prompt setting. All adaptation variants use the same SoulX-Singer foundation model and Peking Opera adaptation set. Subjective scores for all nine configurations are obtained from one blinded nine-system listening test, with ZeroOpera (Full) included as the common reference. The upper and lower sections organize adaptation strategies and component ablations, respectively. Subjective scores are reported as MOS with 95% confidence intervals. Lower is better for FFE and PER; higher is better for AQ, PC, TS, and OF.

Model / Configuration	**FFE** ↓	**PER** ↓	**AQ** ↑	**PC** ↑	**TS** ↑	**OF** ↑
Adaptation Strategies						
SoulX-Singer (Frozen)	0.127	0.289	3.49 ± 0.11	3.68 ± 0.13	3.68 ± 0.12	3.69 ± 0.14
SoulX-Singer-FFT	0.145	0.265	3.25 ± 0.14	3.45 ± 0.13	1.25 ± 0.08	3.55 ± 0.15
SoulX-Singer-LoRA	0.138	0.385	3.65 ± 0.11	3.25 ± 0.14	3.75 ± 0.12	3.15 ± 0.16
Mechanism and Component Ablations						
TPM → Discrete Concat	0.132	0.412	3.85 ± 0.12	3.32 ± 0.15	4.05 ± 0.13	3.28 ± 0.14
w/o Region-Aware Weighting	0.158	0.285	3.90 ± 0.11	3.75 ± 0.12	4.08 ± 0.14	3.42 ± 0.15
w/o Stochastic Routing	0.135	0.298	3.52 ± 0.16	3.65 ± 0.14	3.95 ± 0.12	3.98 ± 0.13
w/o Acoustic Prior	0.254	0.362	2.85 ± 0.15	3.15 ± 0.14	1.18 ± 0.08	3.25 ± 0.16
w/o Inpainting (Full Prior Visible)	0.131	0.328	3.68 ± 0.12	3.42 ± 0.15	4.02 ± 0.13	3.38 ± 0.14
ZeroOpera (Full)	**0.126**	**0.233**	$\mathbf{3.96 \pm 0.12}$	$\mathbf{3.94 \pm 0.13}$	$\mathbf{4.11 \pm 0.13}$	$\mathbf{4.07 \pm 0.14}$

Adaptation Strategies. Table 2 compares configurations built on the same SoulX-Singer checkpoint and opera adaptation set. Full fine-tuning improves PER but substantially degrades timbre similarity, whereas the evaluated LoRA configuration better preserves timbre but provides weaker pronunciation and opera-style adaptation. ZeroOpera achieves the strongest balance across PER, TS, and OF. Since the configurations are not matched in trainable capacity or optimization budget, the comparison is specific to the evaluated settings rather than an intrinsic ranking of adaptation paradigms.

Technique-driven Phonological Modulation. Replacing TPM with discrete embedding concatenation increases the overall PER from 0.233 to 0.412. It also decreases PC from 3.94 to 3.32 and OF from 4.07 to 3.28. These results indicate that directly concatenating a category embedding is less effective than applying category-conditioned scale and shift transformations to the linguistic representation.

Appendix E.2 further visualizes and analyzes the effect of phonological-category conditioning. With all other inputs fixed, we compare the output obtained when all category labels are set to Standard with that obtained using the automatically assigned category labels. The resulting spectrogram difference shows that changing the category condition produces temporally localized acoustic modifications. This visualization demonstrates the model's sensitivity to the category condition, while the PER results provide the quantitative evidence for pronunciation accuracy.

To examine compositional generalization, we further divide the opera-specific characters in the held-out Singer 3 test set according to whether their base-pinyin–category combinations occur in the adaptation set. IV-Category contains combinations observed during adaptation, whereas OOV-Category contains previously unseen combinations. This partition is determined solely from the training and test annotations and is independent of model predictions.

As shown in Table 3, discrete concatenation obtains a PER of 0.355 on IV-Category and 0.861 on OOV-Category. ZeroOpera obtains PERs of 0.322 and 0.328, respectively, corresponding to a 61.9% relative reduction in OOV-Category PER compared with the evaluated discrete-concatenation baseline. The small IV–OOV gap is consistent with the intended factorization of base-pinyin content and phonological category.

This result provides preliminary evidence that TPM can transfer category-conditioned information to unseen base-pinyin–category combinations. However, the analysis is based on one held-out singer, with 36 IV-category tokens and 161 OOV-category tokens, and the automatic metric has a nonzero recognition floor on real recordings. The result should therefore not be interpreted as a general demonstration of compositional phonological generalization across singers, repertoires, or Peking Opera traditions.

Table 3: Category-specific PER on the held-out Singer 3 test set. IV-Cat. and OOV-Cat. denote base-pinyin–category combinations observed and unobserved in the acoustic adaptation set, respectively. The same partition is used for all systems and is independent of model predictions. The partition contains 36 IV-category tokens and 161 OOV-category tokens. Because PER is computed using an automatic recognizer with a nonzero error rate on real recordings, these values are interpreted comparatively. Lower is better.

Model	Standard	IV-Cat.	OOV-Cat.	Overall
GT Recording	0.145	0.318	0.322	0.200
YingMusic	**0.152**	0.338	0.342	**0.211**
SoulX-Singer	0.198	0.485	0.491	0.289
Discrete Concat	0.322	0.355	0.861	0.412
ZeroOpera	0.191	**0.322**	**0.328**	0.233

Acoustic Prior and Inpainting. For the *w/o Acoustic Prior* variant, we replace the masked acoustic prior with an all-zero tensor while retaining the musical conditions, linguistic conditions, TPM, routing strategy, decoder architecture, and training objective. Removing the acoustic prior substantially degrades all metrics, increasing FFE from 0.126 to 0.254 and PER from 0.233 to 0.362, while reducing TS from 4.11 to 1.18.

For the *w/o Inpainting* variant, we provide the complete acoustic prior to the decoder rather than hiding the category-specific regions. The original phonological mask is retained for TPM activation and region-aware weighting. This variant increases PER from 0.233 to 0.328 and decreases PC and OF from 3.94 and 4.07 to 3.42 and 3.38, respectively.

Region-Aware Flow Matching. We replace the region-aware weights with a uniform objective by setting $\lambda_{\text{edit}} = \lambda_{\text{keep}} = 1.0$. This modification increases PER from 0.233 to 0.285 and FFE from 0.126 to 0.158. PC and OF also decrease from 3.94 and 4.07 to 3.75 and 3.42, respectively. These results suggest that a uniform objective underemphasizes the relatively sparse opera-specific regions, whereas region-aware weighting encourages the model to learn the intended acoustic modifications without sacrificing the surrounding structure.

Stochastic Routing. Removing stochastic routing forces the model to use the opera-editing route throughout training and eliminates the identity-preserving reconstruction route. As a result, the overall PER increases from 0.233 to 0.298, while AQ and TS decrease from 3.96 and 4.11 to 3.52 and 3.95, respectively. OF changes more moderately from 4.07 to 3.98. These results indicate that opera-category conditioning alone can retain much of the perceived opera character, but produces less stable pronunciation and acoustic quality. The identity-preserving route therefore acts as a regularizer that reduces unnecessary modifications when category-specific editing is not required.

6 Conclusion

We presented ZeroOpera, a low-resource Peking Opera SVS framework that supports zero-shot prompt-timbre transfer by adapting the output of a frozen foundation model. ZeroOpera uses the foundation model's prompt-conditioned mel-spectrogram as a structural prior and performs opera-domain adaptation through phonology-aware acoustic inpainting and enhancement, while Technique-driven Phonological Modulation factorizes base linguistic content and phonological-category conditioning to provide compositional control over opera-specific pronunciation. Under the evaluated configurations, ZeroOpera improves pronunciation-related objective and subjective metrics while preserving the prompt-timbre capability of the frozen model. We further establish evaluation resources comprising an expert-verified phonological lexicon, an automatic annotation pipeline, and a Pinyin-syllable Peking Opera ASR model. The current evidence is limited to the Mei-school *Qingyi* tradition and does not establish generalization to other Peking Opera schools, role types, or Chinese opera traditions; additional limitations and future directions are discussed in Appendix F. Overall, our results provide an initial case study of adapting prompt-based generative audio models to a culturally specific, low-resource vocal domain.

AI USE STATEMENT

We used generative AI tools, including OpenAI ChatGPT, to assist with English-language editing, manuscript organization, and the improvement of clarity and presentation. The tools were also used to provide feedback on the phrasing and logical structure of technical explanations and experimental discussions. All AI-assisted suggestions were critically reviewed, revised, and approved by the authors.

The research problem, methodology, experimental design, implementation, data processing, evaluation protocol, and interpretation of the experimental results were determined and conducted by the authors. Generative AI tools were not used to fabricate data, generate experimental results, or replace expert annotations or subjective evaluations. Technical claims, numerical results, equations, and references were checked by the authors against the implementation, experimental records, and original sources. The authors take full responsibility for the accuracy, originality, and final content of the paper.

ETHICS STATEMENT

The Peking Opera recordings used in this study were collected and processed with authorization from the relevant performers or data providers. Performer identities are anonymized throughout the paper, and the recordings are used solely for research purposes within the scope of the applicable permissions. We will release only those data, annotations, models, and evaluation resources for which redistribution is permitted. Recordings without explicit redistribution permission will not be publicly released.

The subjective evaluation was conducted with professional Peking Opera practitioners who were informed of the research purpose and evaluation procedure before participation. Participation was voluntary, and no personally identifiable information is reported. The evaluation results are presented only in aggregated form.

We recognize that prompt-conditioned singing voice synthesis may be misused for unauthorized voice imitation or impersonation. ZeroOpera is intended for research, cultural preservation, and authorized creative applications, rather than for generating content that misrepresents a real performer. Any public release will therefore include a usage policy requiring appropriate authorization for voice prompts and generated content. We also encourage disclosure that audio is synthetically generated and recommend the use of provenance tracking, audio watermarking, or access controls in practical deployments. The system should not be regarded as a replacement for professional Peking Opera performers or as an authoritative representation of all Peking Opera schools and role types.

The current dataset focuses on the Mei-school *Qingyi* tradition and therefore does not represent the full diversity of Peking Opera or other Chinese opera traditions. We report this limitation explicitly and avoid generalizing our findings beyond the cultural and acoustic scope supported by the data. The authors take responsibility for complying with the applicable consent, licensing, privacy, and research-integrity requirements.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details required to reproduce ZeroOpera throughout the paper and appendices. The model architecture, Technique-driven Phonological Modulation, stochastic routing strategy, and region-aware flow-matching objective are described in Section 4. Appendix A documents the Peking Opera phonological categories, the expert-verified phonological-category lexicon, and the automatic annotation procedure. Appendix B reports the corpus statistics, singer-disjoint adaptation and test splits, prompt construction, and evaluation pairing protocol. Model configurations, acoustic representations, training hyperparameters, optimization settings, baseline configurations, and the complete training algorithm are provided in Appendix C. Metric definitions, ASR validation, subjective evaluation procedures, and audio-processing settings are documented in Appendix D. Additional analyses and audio examples are included in the supplementary materials. We intend to release the implementation, configuration files, automatic annotation resources, and evaluation scripts upon publication. Resources involving performer recordings will be released only where redistribution is permitted.

REFERENCES

Peng Bai, Yue Zhou, Meizhen Zheng, Wujin Sun, and Xiaodong Shi. Improving chinese pop song and hokkien gezi opera singing voice synthesis by enhancing local modeling. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 3302–3312, 2023.

Ruiqi Li, Yu Zhang, Yongqi Wang, Zhiqing Hong, Rongjie Huang, and Zhou Zhao. Robust singing voice transcription serves synthesis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9751–9766, 2024.

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

Changhao Pan, Dongyu Yao, Yu Zhang, Wenxiang Guo, Jingyu Lu, Zhiyuan Zhu, and Zhou Zhao. Synthetic singers: A review of deep-learning-based singing voice synthesis approaches. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 396–416, 2025.

Jiale Qian, Hao Meng, Tian Zheng, Pengcheng Zhu, Haopeng Lin, Yuhang Dai, Hanke Xie, Wenxiao Cao, Ruixuan Shang, Jun Wu, et al. Soulx-singer: Towards high-quality zero-shot singing voice synthesis. *arXiv preprint arXiv:2602.07803*, 2026.

qiuqiao. SOFA. `https://github.com/qiuqiao/SOFA`, 2023.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Haojie Wei, Xueke Cao, Tangpeng Dan, and Yueguo Chen. Rmvpe: A robust model for vocal pitch estimation in polyphonic music. *arXiv preprint arXiv:2306.15412*, 2023.

Yusong Wu, Shengchen Li, Chengzhu Yu, Heng Lu, Chao Weng, Liqiang Zhang, and Dong Yu. Peking opera synthesis via duration informed attention network. *arXiv preprint arXiv:2008.03029*, 2020.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, et al. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *International Conference on Learning Representations*, volume 2025, pp. 58436–58459, 2025.

Yu Zhang, Rongjie Huang, Ruiqi Li, JinZheng He, Yan Xia, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. Stylesinger: Style transfer for out-of-domain singing voice synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19597–19605, 2024a.

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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1960–1975, 2024b.

Junjie Zheng, Chunbo Hao, Guobin Ma, Xiaoyu Zhang, Gongyu Chen, Chaofan Ding, Zihao Chen, and Lei Xie. Yingmusic-singer: Zero-shot singing voice synthesis and editing with annotation-free melody guidance. *arXiv preprint arXiv:2512.04779*, 2025.

Meizhen Zheng, Peng Bai, Xiaodong Shi, Xun Zhou, and Yiting Yan. Ft-gan: Fine-grained tune modeling for chinese opera synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19697–19705, 2024.

Xun Zhou, Wujin Sun, and Xiaodong Shi. A high-quality melody-aware peking opera synthesizer using data augmentation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1092–1097. IEEE, 2023.

Yue Zhou, Peng Bai, and Xiaodong Shi. Univoice: a unified framework for text-to-speech, singing voice synthesis, and opera singing synthesis. *Applied Intelligence*, 56(4):104, 2026.

A PHONOLOGY AND ANNOTATION

Traditional Peking Opera pronunciation draws on the *Zhongzhou* rhyme tradition and historical Beijing speech, and therefore differs from modern Standard Mandarin in several systematic ways. In this work, we focus on three pronunciation categories that are particularly relevant to the Mei-school *Qingyi* repertoire: *Jian*, *Tuan*, and *Shang-kou-zi*. Together with a Standard category, they form the four-way label set used by Technique-driven Phonological Modulation (TPM): $l_i \in \{0, 1, 2, 3\}$.

These labels describe categorical pronunciation patterns associated with lyric characters. They do not constitute a complete phonetic transcription of Peking Opera singing. Fine-grained acoustic characteristics, such as articulatory emphasis, resonance, vowel prolongation, and micro-prosodic variation, are learned from the acoustic training targets rather than explicitly encoded by the labels.

A.1 PEKING OPERA PHONOLOGICAL CATEGORIES

Jian and Tuan. The *Jian–Tuan* distinction reflects a historical contrast that has largely merged in modern Standard Mandarin. In relevant syllabic contexts, Standard Mandarin uses the palatal initials *j*, *q*, and *x*, whereas traditional opera pronunciation may distinguish two character classes:

- **Jian (Label 1):** Characters in this category are realized with dental or alveolar affricative/fricative initials, represented here as *z*, *c*, or *s*. For example, "arrow" has the Standard Mandarin pinyin *jian* but is represented as *zian* in the corresponding opera pronunciation.
- **Tuan (Label 2):** Characters in this category retain palatal initials represented as *j*, *q*, or *x*. Their surface pinyin may therefore coincide with the Standard Mandarin form. For example, "sword" remains *jian*. We preserve the Tuan category because its historical and lexical status distinguishes it from Jian characters and prevents the Jian transformation from being transferred to inappropriate characters.

Shang-kou-zi. *Shang-kou-zi* (Label 3) denotes a lexically specified set of characters whose traditional opera pronunciations involve vowel or final patterns that differ from their Standard Mandarin forms. For example, "white" is represented as *bai* in Standard Mandarin and *bo* in the opera pronunciation, while "brother" changes from *ge* to *guo*. These character-specific variants contribute substantially to pronunciation clarity and perceived operatic style.

Standard. The Standard category (Label 0) contains characters that are not assigned any of the three category-specific pronunciation patterns above. Label 0 indicates the absence of an explicit Jian, Tuan, or Shang-kou-zi transformation; it does not imply that the complete acoustic realization is identical to ordinary Mandarin speech or pop singing. Standard-category characters may still exhibit opera-specific resonance, articulation, duration, vibrato, and prosodic realization. ZeroOpera learns these continuous acoustic characteristics through full-spectrogram supervision and acoustic enhancement, rather than representing them with additional discrete labels.

Table 4 provides representative examples of the four categories. The opera-pronunciation column is included only for linguistic illustration and evaluation design. For both SoulX-Singer and ZeroOpera, the linguistic input remains the Standard Pinyin column. For example, the Jian character "arrow" is represented by the base token *jian* with label 1 rather than by the token *zian*; the Shang-kou-zi character "white" is represented by *bai* with label 3 rather than by *bo*. Thus, ZeroOpera is trained using Standard Mandarin pinyin together with category labels, rather than manually aligned opera-specific phoneme sequences or opera-surface pinyin tokens.

A.2 EXPERT-VERIFIED CATEGORY LEXICON

We construct an expert-verified opera-pronunciation lexicon that maps lyric characters to their opera-specific pronunciation forms and phonological categories. The lexicon contains opera-surface pronunciation forms for linguistic documentation, forced alignment, and evaluation-reference construction. However, these opera-surface forms are not passed to the SoulX-Singer prior model or the ZeroOpera content encoder. At model input, every lyric character is represented by its tone-bearing Standard Mandarin pinyin token, while the lexicon contributes only its four-way phonological-category label. Its category-label component contains 1,755 unique character–category entries: 487

Table 4: Representative Peking Opera pronunciation categories and their corresponding TPM labels. The examples illustrate the category definitions and are not intended as a complete phonological inventory.

Meaning	Standard Pinyin	Opera Form	Category	Label
Sing	chang	chang	Standard	0
Sky	tian	tian	Standard	0
Arrow	jian	**zian**	Jian	1
Autumn	qiu	**ciu**	Jian	1
Small	xiao	**siao**	Jian	1
Sword	jian	jian	Tuan	2
Bridge	qiao	qiao	Tuan	2
White	bai	**bo**	Shang-kou-zi	3
Brother	ge	**guo**	Shang-kou-zi	3
Far	yuan	**wan**	Shang-kou-zi	3

Table 5: Statistics of the expert-verified Peking Opera phonological-category lexicon. Counts denote unique character–category entries rather than their occurrences in the acoustic corpus.

Category	Character Entries
Jian	487
Tuan	830
Shang-kou-zi	438
Total	1,755

Jian entries, 830 Tuan entries, and 438 Shang-kou-zi entries, as summarized in Table 5. Characters without a matching category entry are assigned the Standard label.

Given an input lyric sequence, we first obtain the tone-bearing Standard Mandarin pinyin sequence using a Standard Mandarin G2P conversion. In parallel, lexicon lookup produces one phonological-category label for each lyric character. The resulting model input is therefore the pair of aligned sequences (p, l), where p contains only Standard Mandarin pinyin tokens and l contains the four-way phonological labels. The same deterministic procedure is used during adaptation and inference. Consequently, ZeroOpera does not require category labels derived from the target recording, manually aligned opera-specific phoneme sequences, or test-time access to the reference audio.

The category lexicon and the expert-verified phoneme transcriptions serve different purposes. The lexicon is used to generate the character-level TPM labels during model adaptation and inference. By contrast, the complete opera-pronunciation transcriptions are used only as references for pronunciation evaluation and are not provided to ZeroOpera as Pinyin-syllable training supervision.

A.3 Automatic Annotation and Musical-Condition Extraction

Figure 3 illustrates the offline preprocessing pipeline used to construct the annotations and musical conditions for the Peking Opera recordings. This pipeline is applied only to the acoustic adaptation corpus and evaluation references; it does not imply that a target vocal recording is required during synthesis inference.

Given a vocal recording, RMVPE (Wei et al., 2023) extracts its continuous frame-level F0 trajectory. In parallel, the corresponding lyrics are processed in two aligned representations. A Standard Mandarin G2P conversion produces the tone-bearing pinyin tokens used as linguistic input to SoulX-Singer and ZeroOpera. Separately, the expert-verified opera lexicon provides character-level phonological-category labels. For corpus preprocessing only, the lexicon's opera-surface pronunciation forms are additionally used to construct linguistically appropriate forced-alignment and evaluation references. These opera-surface forms are never used as input tokens to the synthesis models. Characters without a matching opera-specific entry are assigned the Standard category.

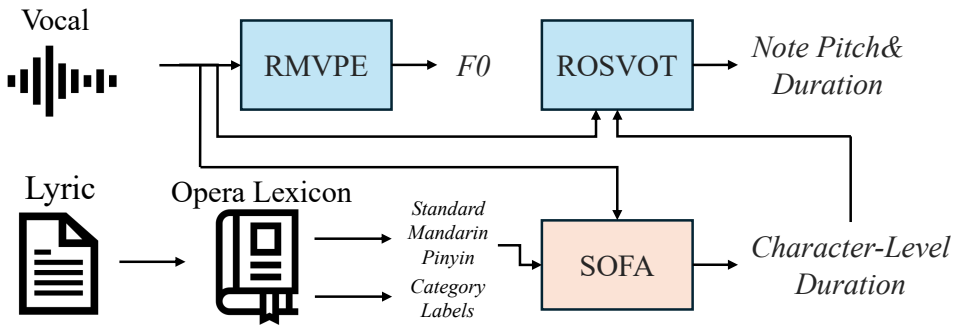


Figure 3: Offline preprocessing and annotation pipeline for the Peking Opera corpus. RMVPE extracts frame-level F0 from each vocal recording. A Standard Mandarin G2P conversion derives the tone-bearing pinyin tokens used as synthesis-model input. In parallel, the expert-verified opera lexicon derives character-level phonological-category labels and, only for corpus annotation, opera-surface pronunciation forms used by forced alignment. SOFA aligns the pinyin sequence with the vocal recording to obtain character-level durations. ROSVOT then uses the recording and the SOFA-derived character boundaries to transcribe note pitches and durations. This recording-dependent pipeline is used for corpus preprocessing and controlled evaluation; synthesis inference does not require a target vocal recording.

SOFA(qiuqiao, 2023) performs forced alignment between the vocal recording and the opera-pronunciation pinyin sequence used only for corpus annotation, in order to estimate character-level boundaries and durations. The vocal recording and SOFA-derived character boundaries are subsequently provided to ROSVOT(Li et al., 2024), which transcribes the note-level pitches and durations. The resulting frame-level F0 trajectory, note pitches and durations, note attributes, and character durations are temporally aligned with the lyric sequence. For acoustic adaptation, the stored linguistic input tokens are the tone-bearing Standard Mandarin pinyin tokens, together with the character-level phonological-category labels. The opera-surface pinyin sequence used for forced alignment is retained only as annotation metadata and is not fed to either synthesis model.

The phonological-category labels are obtained deterministically from the input lyrics and the same lexicon lookup is available at inference time. By contrast, the RMVPE-, SOFA-, and ROSVOT-derived conditions shown in Figure 3 are extracted from the recordings for corpus preprocessing and controlled evaluation. During practical text-to-singing inference, the target lyrics, score or melody condition, and acoustic prompt are provided by the user; no target recording or recording-derived alignment is required.

B Dataset and Protocol

This section describes the corpus composition, singer-disjoint data split, prompt construction, evaluation pairing protocol, and compositional coverage of the held-out test set. The phonological categories and automatic annotation procedure are described separately in Appendix A.

B.1 Corpus Composition

Our Peking Opera corpus contains 948 recordings from six professional singers, with a total duration of 8 hours, 20 minutes, and 44 seconds (approximately 8.35 hours). All recordings belong to the Mei-school *Qingyi* tradition, a role type within the broader Dan category. To protect performer privacy, we anonymize the performers as Singer 1–Singer 6.

Character-level phonological-category labels are generated using the deterministic expert-verified lexicon described in Appendix A.2. Table 6 reports the audio duration and phonological-category token distribution for each singer. The complete corpus contains 1,303 *Jian* tokens, 1,160 *Tuan* tokens, 2,705 *Shang-kou-zi* tokens, and 11,295 Standard tokens. These categories account for 7.91%,

Table 6: Audio duration and phonological-category distribution for each singer. Counts denote lyric-character occurrences, and percentages in parentheses are calculated relative to the total number of lyric characters for the corresponding singer.

Singer	Duration	Jian	Tuan	Shang-kou-zi	Standard
Singer 1	00:30:04	90 (9.53%)	61 (6.46%)	159 (16.84%)	634 (67.16%)
Singer 2	02:00:10	341 (7.66%)	317 (7.12%)	758 (17.03%)	3,035 (68.19%)
Singer 3	00:25:37	48 (8.39%)	41 (7.17%)	108 (18.88%)	375 (65.56%)
Singer 4	01:49:35	182 (6.91%)	167 (6.34%)	433 (16.43%)	1,853 (70.32%)
Singer 5	02:47:05	491 (8.25%)	458 (7.69%)	911 (15.30%)	4,094 (68.76%)
Singer 6	00:48:13	151 (7.92%)	116 (6.08%)	336 (17.62%)	1,304 (68.38%)
Overall	**08:20:44**	**1,303 (7.91%)**	**1,160 (7.05%)**	**2,705 (16.43%)**	**11,295 (68.61%)**

7.05%, 16.43%, and 68.61% of all lyric-character occurrences, respectively. The category proportions are broadly similar across singers, despite differences in recording duration.

B.2 Singer-Disjoint Data Split

We designate Singer 3 as the held-out target singer. All 44 recordings from Singer 3, totaling 25 minutes and 37 seconds, are excluded from opera-domain adaptation, acoustic-prompt construction, and ASR training. These recordings define the 44 target conditions used for synthesis evaluation and provide the corresponding ground-truth references.

The target recordings are used to derive the target lyrics, musical scores, and frame-level melody conditions required by the evaluated systems. They are also used as ground-truth references for objective and subjective evaluation. However, the target waveforms are never supplied as acoustic prompts or direct acoustic inputs to any synthesis system.

From Singers 1, 2, 4, 5, and 6, we reserve 80 recordings, totaling 20 minutes and 23 seconds, exclusively as in-domain opera prompts. These prompt recordings are excluded from the opera adaptation set. After excluding the 44 Singer 3 target recordings and the 80 reserved opera-prompt recordings, the acoustic adaptation set contains 824 recordings with a total duration of 7 hours, 34 minutes, and 44 seconds, or approximately 7.58 hours.

Singer identities represented by the in-domain opera prompts also occur in the adaptation set, but the prompt utterances themselves do not. Consequently, this condition evaluates timbre preservation for seen singers on unseen utterances; it does not constitute an unseen-singer zero-shot condition.

B.3 Prompt Construction and Evaluation Pairing

We evaluate two complementary acoustic-prompt conditions: in-domain opera prompts and cross-domain pop prompts.

In-domain opera prompts. The in-domain prompt pool consists of the 80 reserved Peking Opera recordings from Singers 1, 2, 4, 5, and 6. The singers are represented in the opera adaptation set, but the specific prompt recordings are strictly excluded from adaptation. This condition is therefore used to evaluate in-domain timbre preservation under a seen-singer, unseen-utterance setting.

Cross-domain pop prompts. The cross-domain prompt pool consists of 100 pop-singing segments from 50 singers, with two segments per singer and a total duration of 13 minutes and 34 seconds. These singers are excluded from opera-domain adaptation. This condition evaluates zero-shot transfer of timbre from unseen pop singers to Peking Opera synthesis.

For the complete objective-evaluation pool, every prompt is paired with all 44 held-out target conditions. The 80 opera prompts therefore produce 3,520 synthesized samples, while the 100 pop prompts produce 4,400 synthesized samples. The complete pairing protocol yields 7,920 synthesized samples, as summarized in Table 7.

Table 7: Summary of the complete evaluation-pairing protocol. "Singer in adaptation" indicates whether the prompt singer identity is represented in the Peking Opera adaptation set. "Prompt in adaptation" indicates whether the specific prompt segment is included in that set. Every prompt is paired with all 44 held-out target conditions.

Setting	Prompt Singers	Prompts	Targets	Synthesized Samples	Singer in Adaptation?	Prompt in Adaptation?
Opera prompts	5	80	44	3,520	Yes	No
Pop prompts	50	100	44	4,400	No	No
Overall	55	180	44	7,920	Mixed	No

Table 8: Compositional coverage of category-specific base-pinyin–category combinations in the held-out Singer 3 test set. IV and OOV denote combinations observed and unobserved in the acoustic adaptation set, respectively. Types count unique combinations, while tokens count their occurrences in the test set. Standard- category characters are excluded from this diagnostic partition.

Category	IV Types	IV Tokens	OOV Types	OOV Tokens
Jian	3	8	23	40
Tuan	3	6	20	35
Shang-kou-zi	8	22	58	86
Total	14	36	101	161

The subjective evaluation uses a sampled subset of these prompt–target pairs rather than all 7,920 synthesized samples. Its sampling procedure, number of ratings, and listener protocol are described in Appendix D.3.

B.4 Compositional Coverage in the Test Set

TPM represents each phonological category separately from the underlying base-pinyin representation. This factorization is designed to support generalization to base-pinyin–category combinations that are absent from the acoustic adaptation set. We therefore partition the opera-specific character occurrences in the held-out Singer 3 test set according to whether their corresponding base-pinyin–category combinations occur in the adaptation set.

Here, base pinyin always denotes the tone-bearing Standard Mandarin pinyin token used as model input, rather than an opera-surface pronunciation form. A combination is classified as in-vocabulary (IV) when the same Standard Mandarin base-pinyin token and phonological category occur together in the acoustic adaptation set. It is classified as out-of-vocabulary (OOV) when that particular base-pinyin–category pairing is absent from the adaptation set. This definition concerns combinations rather than individual pinyin syllables, phonological categories, or lexicon entries. An OOV combination may therefore contain a Standard Mandarin base-pinyin token and a category that have each been observed separately during adaptation but have not been observed together.

This partition includes only characters assigned to the *Jian*, *Tuan*, or *Shang-kou-zi* categories. Standard-category characters are excluded because the purpose of the partition is to diagnose generalization of the category-conditioned phonological transformations.

Table 8 reports both the number of unique base-pinyin–category combinations (types) and their character occurrences (tokens) in the held-out test set. The partition contains 14 IV types with 36 tokens and 101 OOV types with 161 tokens. In total, it covers all 197 opera-specific character occurrences in the 44 held-out Singer 3 recordings.

The predominance of OOV combinations makes this partition a challenging diagnostic of compositional phonological generalization. However, it is derived from a single held-out singer and contains substantially fewer IV than OOV tokens. We therefore interpret the resulting comparison as preliminary evidence rather than as a comprehensive benchmark of generalization across singers, repertoires, role types, or Peking Opera traditions.

Table 9: Control configurations and parameter counts of the evaluated systems. All systems receive the same target lyrics and acoustic prompt. Melody-controllable systems use the same target F0 trajectory, converted into the input format required by each system's official inference interface. StyleSinger only supports score-level control and therefore uses the corresponding target score. For systems without Peking Opera adaptation, the parameter count denotes the original model size; for the adaptation variants, it denotes the number of parameters updated during adaptation. The SoulX-Singer backbone remains frozen in ZeroOpera.

Model	Native Control	Control Used	Opera Adaptation	Parameters
StyleSinger	Score	Target score	No	42M
YingMusic-Singer	Melody	Target F0	No	300M
VevoSing	Melody	Target F0	No	1.13B
SoulX-Singer	Score / Melody	Target F0	No	449.3M
SoulX-Singer-FFT	Score / Melody	Target F0	Yes	449.3M
SoulX-Singer-LoRA	Score / Melody	Target F0	Yes	$\sim$8.29M
ZeroOpera (Ours)	Score / Melody	Target F0	Yes	143.17M

C Implementation Details

C.1 Model and Training Configurations

Table 9 summarizes the control configurations, Peking Opera adaptation settings, and parameter counts of the evaluated systems. All melody-controllable systems use the same target F0 trajectory, converted into the representation required by each system's official inference interface. StyleSinger does not support melody control and is therefore evaluated using the corresponding target score.

Because the released systems expose different native control interfaces, comparisons with external baselines are interpreted as system-level comparisons under their strongest applicable control modes. Comparisons among the SoulX-Singer-based variants directly control for the foundation model, target F0, and waveform-generation pipeline.

StyleSinger. We use the official implementation and released checkpoint of StyleSinger (Zhang et al., 2024a), whose model architecture contains approximately 42M trainable parameters. StyleSinger employs separate phoneme and note encoders with a hidden dimension of 256. The phoneme encoder consists of four Feed-Forward Transformer blocks, while the note encoder represents note pitch, note type, and note duration. Reference-style information is modeled using a pretrained wav2vec 2.0 encoder together with a Residual Style Adaptor (RSA). The RSA contains a five-layer convolutional encoder and a residual quantization module with a codebook size of 128 and a quantization depth of 4. StyleSinger further uses a 100-step pitch diffusion predictor for F0 and voiced/unvoiced prediction and a four-step diffusion decoder for mel-spectrogram generation. We follow the official acoustic configuration with a sampling rate of 48 kHz, an analysis window size of 1,024, a hop size of 256, and 80 mel-frequency bins. Waveforms are generated using the official BigVGAN vocoder. For evaluation, the input lyrics and musical scores are converted to the format required by the official inference pipeline, and each evaluation prompt is provided as the reference singing sample for style extraction. We do not perform Peking Opera adaptation or test-singer fine-tuning for this baseline.

YingMusic-Singer. We use the official implementation and released checkpoint of YingMusic-Singer (Zheng et al., 2025), without additional fine-tuning on our Peking Opera corpus. Its model architecture contains approximately 300M parameters. YingMusic-Singer is an annotation-free, melody-guided zero-shot SVS system whose generative backbone is a Diffusion Transformer initialized from F5-TTS. The backbone comprises 12 decoder layers with a hidden dimension of 1,024 and 16 attention heads. The model is conditioned on lyrics, an acoustic prompt for zero-shot timbre transfer, and frame-level melody representations extracted by its online melody encoder. In our evaluation, the melody-reference input is constructed from the target musical conditions rather than taken from the original ground-truth Peking Opera recording. The online melody extractor is trained under the guidance of a frozen SOME teacher model, while Centered Kernel Alignment encourages correspondence between the extracted melody representations and the intermediate features of the flow model. The released model is further optimized using Flow-GRPO with content-accuracy and

melodic-similarity rewards. Following the official inference configuration, we generate audio at 24 kHz using an Euler ODE solver with 32 sampling steps and a classifier-free guidance scale of 2. We retain the official acoustic-generation and waveform-reconstruction pipeline for system-level evaluation.

VevoSing. We use the released VevoSing implementation, which adapts the VEVO framework (Zhang et al., 2025) to singing voice synthesis. Its model architecture contains approximately 1.13B trainable parameters. The autoregressive Transformer consists of 12 layers with a hidden size of 1,920, an intermediate size of 7,680, and 16 attention heads. The model uses Whisper representations together with 24-dimensional chromagram features and employs style and content–style codebooks containing 512 and 16,384 entries, respectively. Audio is represented as 128-bin mel-spectrograms at 24 kHz, with a hop size of 480 and an FFT/window size of 1,920. We use the released VevoSing checkpoint and its corresponding inference configuration without additional adaptation to our Peking Opera corpus.

SoulX-Singer. We use the official SoulX-Singer implementation and checkpoint (Qian et al., 2026). Its model architecture contains 449.3M trainable parameters. The content encoder consists of four layers with 512-dimensional text, pitch, note-type, and F0 embeddings. The flow-matching decoder contains 22 Transformer layers with a hidden size of 1,024 and 16 attention heads. The model generates 128-bin mel-spectrograms at 24 kHz using a hop size of 480 and an FFT/window size of 1,920. Following the official inference configuration, we use 32 sampling steps and a classifier-free guidance scale of 3. For all experiments, we adopt the frame-level melody control mode. The same target melody condition is used for the frozen SoulX-Singer baseline, SoulX-Singer-FFT, SoulX-Singer-LoRA, and ZeroOpera. The released SoulX-Singer checkpoint is evaluated without Peking Opera adaptation and is kept frozen when used to generate the acoustic prior for ZeroOpera.

SoulX-Singer-FFT. We initialize SoulX-Singer-FFT from the pretrained SoulX-Singer checkpoint and fine-tune all 449.3M model parameters on the Peking Opera adaptation set. The model is trained for 30,000 iterations with a learning rate of 0.00005 and a batch size of 4 per GPU, yielding a global batch size of 32 across eight NVIDIA H200 GPUs. We use the same audio preprocessing, adaptation set, and waveform-generation pipeline as for ZeroOpera. During inference, we use 32 flow-matching steps and a classifier-free guidance scale of 3.

SoulX-Singer-LoRA. We initialize SoulX-Singer-LoRA from the pretrained SoulX-Singer checkpoint and insert LoRA adapters into its flow-matching decoder while keeping all original model parameters frozen. Specifically, rank-16 adapters with a scaling factor of 32 and a dropout rate of 0.05 are inserted into the query, key, value, and output projections, as well as the gate, up, and down projections of the feed-forward layers. No bias parameters are trained. This configuration introduces approximately 8.29M trainable parameters. We train the model on the same Peking Opera adaptation set for 7,500 iterations with a learning rate of 0.002 and a batch size of 4 per GPU, yielding a global batch size of 32 across eight NVIDIA H200 GPUs. The audio preprocessing and waveform-generation pipeline are the same as those used for SoulX-Singer-FFT. During inference, we use 32 flow-matching steps and a classifier-free guidance scale of 3.

ZeroOpera. ZeroOpera uses the output of the frozen pretrained SoulX-Singer model as the acoustic prior for phonology-aware inpainting. The acoustic priors are generated offline and cached before training. Its trainable flow-matching decoder consists of six Transformer layers with a hidden size of 1,024 and eight attention heads. Together with TPM and the associated conditioning modules, ZeroOpera contains 143.17M trainable parameters, while all parameters of the SoulX-Singer backbone remain frozen. The model predicts 128-bin mel-spectrograms at 24 kHz using a hop size of 480 and an FFT/window size of 1,920. We train ZeroOpera for 300,000 iterations with a learning rate of 0.0002, 1,500 warm-up iterations, and gradient clipping with a maximum norm of 1. The batch size is 4 per GPU, yielding a global batch size of 32 across eight NVIDIA H200 GPUs. The complete training process takes approximately six hours. During inference, we use 32 flow-matching steps and a classifier-free guidance scale of 3.

Scope of the Adaptation Comparison. The full-fine-tuning, LoRA, and ZeroOpera configurations share the same foundation checkpoint, adaptation data, target melody conditions, acoustic preprocessing, and waveform-generation pipeline. They are not matched in trainable parameter count, training iterations, wall-clock cost, or optimization objective. The reported LoRA result corresponds specifically to rank 16 and the insertion locations described above; we do not claim that it represents

the optimal performance achievable by all LoRA configurations. A broader evaluation over LoRA ranks, adapter locations, parameter-matched architectures, and compute-matched training schedules is left for future work.

C.2 ZeroOpera Training Procedure

Before training ZeroOpera, we generate the acoustic priors offline. For each training target segment, we randomly select a temporally non-overlapping segment from the same song recording as the acoustic prompt. The frozen SoulX-Singer receives this prompt together with the target musical conditions and generates

$$Mel_{\text{pop}} = \mathcal{F}_{\text{base}}\left(C, Mel_{\text{prompt}}\right).$$

The resulting prompt-conditioned acoustic prior is cached together with the corresponding training target. During ZeroOpera training, the cached prior is loaded directly, so SoulX-Singer is not executed at each optimization step.

Algorithm 1 Training Procedure of ZeroOpera

Require: Trainable inpainting decoder θ, TPM parameters ϕ, and cached training set $\mathcal{D}_{\text{cache}}$ containing offline-generated acoustic priors.
Require: Region weights $\lambda_{\text{edit}} = 5.0$ and $\lambda_{\text{keep}} = 1.0$.
for $s = 1, \ldots, 300{,}000$ **do**
  Sample a training example and its cached acoustic prior: $(p, f_0, n_{\text{type}}, Mel_{\text{pop}}, Mel_{\text{opera}}, l, d) \sim \mathcal{D}_{\text{cache}}$.
  Construct the frame-level phonological mask: $M \leftarrow \text{Expand}\left(\{\mathbb{I}[l_i \neq 0]\}_{i=1}^{N}, d\right)$.
  Construct the masked acoustic prior: $\widetilde{Mel}_{\text{pop}} \leftarrow Mel_{\text{pop}} \odot (\mathbf{1} - M)$.
  Sample the routing variable: $r \sim \text{Bernoulli}(0.5)$.
  if $r = 1$ **then**
    Set the flow endpoint: $Y_r \leftarrow Mel_{\text{opera}}$.
  else
    Set the flow endpoint: $Y_r \leftarrow Mel_{\text{pop}}$.
  end if
  Obtain the base linguistic representation H_{content} from the content encoder using the tone-bearing Standard Mandarin pinyin sequence p; do not use opera-surface pinyin as model input.
  Compute the TPM scale and shift: $(\gamma, \beta) \leftarrow \text{TPM}_{\phi}(l)$.
  Compute the category- and route-dependent gate: $a_i \leftarrow r \cdot \mathbb{I}[l_i \neq 0]$.
  Modulate the linguistic representation: $H_{\text{mod},i} \leftarrow H_{\text{content},i} \odot (\mathbf{1} + a_i\gamma_i) + a_i\beta_i$.
  Expand the modulated representation: $H_{\text{frame}} \leftarrow \text{LR}(H_{\text{mod}}, d)$.
  Sample $x_0 \sim \mathcal{N}(0, I)$ and $\tau \sim \mathcal{U}(0, 1)$, and construct the intermediate flow state: $x_\tau \leftarrow [1 - \tau]x_0 + \tau Y_r$.
  Construct the target vector field: $v_{\text{gt}} \leftarrow Y_r - x_0$.
  Form the decoder input: $z_\tau \leftarrow \text{Concat}\left(x_\tau, \widetilde{Mel}_{\text{pop}}, M\right)$.
  Predict the conditional vector field: $v_\theta \leftarrow v_\theta\left(z_\tau, \tau, H_{\text{frame}}, f_0, n_{\text{type}}\right)$.
  Assign frame-level region weights: $w_j \leftarrow \lambda_{\text{keep}} + (\lambda_{\text{edit}} - \lambda_{\text{keep}})\, M_j$.
  Compute the region-aware flow-matching loss $\mathcal{L}_{\text{RAFM}}$.
  Update only θ and ϕ.
end for

Algorithm 1 summarizes the ZeroOpera optimization procedure after offline acoustic-prior construction. For each target segment, a temporally non-overlapping segment from the same song recording is randomly selected as the SoulX-Singer acoustic prompt. The frozen SoulX-Singer uses this prompt and the target musical conditions to generate the corresponding Mel_{pop}. The generated prior is cached before ZeroOpera training and loaded directly during optimization.

For each training example, we construct a frame-level phonological mask from the automatically assigned character-level category labels and their corresponding durations. Regions associated with *Jian*, *Tuan*, and *Shang-kou-zi* are removed from the acoustic prior, whereas the remaining regions are retained as visible context for the inpainting decoder.

During training, stochastic routing selects between two reconstruction targets with equal probability. The pop-to-opera route uses the ground-truth Peking Opera mel-spectrogram as the flow endpoint and activates TPM to inject category-conditioned phonological modulation. The identity-reconstruction route uses the unmasked acoustic prior as the flow endpoint and disables TPM through the routing gate. This route regularizes the decoder to recover masked content without introducing unnecessary opera-specific modifications. In both routes, the decoder predicts the complete mel-spectrogram rather than only the masked regions.

The same region-aware weighting is applied to both routing paths. Frames in opera-specific phonological regions receive a weight of 5.0, whereas the remaining valid frames receive a weight of 1.0. Consequently, the pop-to-opera route emphasizes reconstruction of category-specific pronunciations, while the identity-reconstruction route emphasizes faithful recovery of the masked foundation-model output. Supervision remains active over the complete non-padding acoustic sequence, allowing the decoder to enhance both the edited regions and the surrounding acoustic context.

Only the inpainting decoder and TPM-related parameters are optimized. SoulX-Singer is used only during offline acoustic-prior construction and is not included in the ZeroOpera training computation graph. At inference time, stochastic routing is disabled by setting $r = 1$. ZeroOpera therefore always uses the pop-to-opera route, with the automatically assigned phonological-category labels activating TPM and the corresponding regions of the acoustic prior being masked for reconstruction.

D Evaluation Details

D.1 Peking Opera ASR Evaluation

To support objective pronunciation evaluation, we construct a pronunciation-oriented Peking Opera ASR model by fine-tuning the official Qwen3-ASR-0.6B checkpoint (Shi et al., 2026). Qwen3-ASR-0.6B consists of a Qwen3-0.6B language model, an audio–text projector, and a 180M-parameter AuT encoder with a hidden size of 896. The audio encoder processes 128-dimensional filter-bank features and applies eight-fold temporal downsampling.

The original Qwen3-ASR model generates written transcriptions. To adapt it for Peking Opera pronunciation recognition, we extend its vocabulary with complete Pinyin syllable tokens and convert the training transcriptions into Pinyin-syllable sequences as the recognition targets. We perform full-parameter supervised fine-tuning using the official Qwen3-ASR fine-tuning framework. Following its supervised fine-tuning recipe, we optimize the autoregressive cross-entropy loss over the target Pinyin syllable tokens. The model is fine-tuned for 5,000 optimization steps with a learning rate of 2×10^{-5}. During inference, we use greedy autoregressive decoding without an external language model.

The ASR training data follow the same split as the ZeroOpera adaptation data. They comprise recordings from Singers 1, 2, 4, 5, and 6 after excluding the 80 recordings reserved as in-domain acoustic prompts. All recordings from Singer 3 are excluded from ASR training. In particular, none of the 44 real Singer 3 recordings used for synthesis evaluation is used for ASR training, validation, checkpoint selection, or hyperparameter tuning. The 80 in-domain prompt recordings are likewise excluded from all ASR training data.

Although the recognizer directly generates complete Pinyin syllables, pronunciation errors are evaluated at the phoneme level. We apply the same deterministic Pinyin-to-phoneme conversion to both the recognized and expert-verified reference Pinyin sequences. We then compute the edit distance between the resulting phoneme sequences. Phoneme Error Rate is defined as $\mathrm{PER} = (S+D+I)/N$, where S, D, and I denote the numbers of phoneme substitutions, deletions, and insertions, respectively, and N is the number of phonemes in the reference sequence.

For category-specific evaluation, phoneme errors are attributed according to the phonological category of the corresponding reference character after sequence alignment. Substitution and deletion errors are assigned to the category of the aligned reference character. Insertions do not have a unique reference category and are therefore included only in the overall PER. As a result, the overall PER is not necessarily the frequency-weighted average of the category-specific PER values.

Table 10: Validation of the Peking Opera ASR on the 44 real recordings from held-out Singer 3. None of these recordings is used for ASR training, validation, checkpoint selection, or hyperparameter tuning. PER is computed after converting the recognized and reference Pinyin-syllable sequences into phoneme sequences. Lower is better.

Recognizer	Overall	Standard	Jian	Tuan	Shang-kou
General Mandarin ASR	0.685	0.412	0.945	0.456	0.912
Peking Opera ASR	**0.200**	**0.145**	**0.265**	**0.218**	**0.282**

We evaluate the recognizer on the 44 real recordings from held-out Singer 3 using expert-verified Peking Opera pronunciation transcriptions as references. Table 10 compares the fine-tuned Peking Opera ASR with a general Mandarin ASR model on the same recordings. The Peking Opera ASR achieves an overall PER of 0.200, compared with 0.685 for the general Mandarin recognizer. It also yields lower PER for the Standard, *Jian*, *Tuan*, and *Shang-kou-zi* categories. These results indicate that domain-specific fine-tuning makes the recognizer better suited to Peking Opera pronunciation evaluation.

Because the Peking Opera ASR still obtains a nonzero PER on real recordings, we interpret synthesis PER as a comparative ASR-based pronunciation metric rather than an absolute transcription error rate or a replacement for expert listening evaluation.

D.2 F0 Frame Error Evaluation

We evaluate target-pitch accuracy using F0 Frame Error (FFE), which jointly accounts for gross pitch errors and voicing-decision errors. For both the ground-truth recording and the synthesized audio, we extract frame-level F0 trajectories using the YIN pitch estimator. All evaluated recordings are processed at 22.05 kHz. We use an analysis window of 1,024 samples and a hop size of 256 samples, corresponding to approximately 46.4 ms and 11.6 ms, respectively. The F0 search range is restricted to 100–500 Hz. The same pitch-extraction configuration is applied to the ground-truth and synthesized recordings.

The extracted reference and synthesized F0 trajectories are matched on their common temporal support before frame-wise comparison. Let f_t^{ref} and f_t^{syn} denote the reference and synthesized F0 values, respectively, at frame t. Unvoiced frames are represented by zero F0. Their corresponding voicing indicators are defined as

$$v_t^{\mathrm{ref}} = \mathbb{I}\left[f_t^{\mathrm{ref}} > 0\right]$$

and

$$v_t^{\mathrm{syn}} = \mathbb{I}\left[f_t^{\mathrm{syn}} > 0\right].$$

A voicing-decision error occurs when the reference and synthesized frames have different voiced/unvoiced decisions:

$$\mathbb{I}_{\mathrm{VDE}}(t) = \mathbb{I}\left[v_t^{\mathrm{ref}} \neq v_t^{\mathrm{syn}}\right].$$

For frames that are voiced in both recordings, a gross pitch error occurs when the absolute relative F0 deviation exceeds 20%:

$$\mathbb{I}_{\mathrm{GPE}}(t) = \mathbb{I}\left[v_t^{\mathrm{ref}} = 1,\, v_t^{\mathrm{syn}} = 1,\, \frac{\left|f_t^{\mathrm{syn}} - f_t^{\mathrm{ref}}\right|}{f_t^{\mathrm{ref}}} > 0.2\right].$$

FFE is computed as the proportion of aligned frames that contain either a gross pitch error or a voicing-decision error:

$$\mathrm{FFE} = \frac{\sum_{t=1}^{T} \mathbb{I}_{\mathrm{GPE}}(t) + \sum_{t=1}^{T} \mathbb{I}_{\mathrm{VDE}}(t)}{T},$$

where T is the total number of temporally matched frames. Since gross pitch errors are defined only on jointly voiced frames, while voicing-decision errors occur when the two voicing decisions differ, the two error categories are mutually exclusive. A lower FFE indicates better agreement with both the target F0 contour and the target voicing pattern.

We compute FFE separately between each synthesized utterance and its corresponding ground-truth recording, and report the mean FFE over all utterances in the evaluation set. No global pitch normalization, transposition, or register compensation is applied before the calculation, since FFE is intended to measure fidelity to the absolute target melody.

D.3 Subjective Evaluation

We conduct blind subjective evaluations with 15 evaluators who have professional experience in Peking Opera. The same rating definitions, reference audio, anonymization procedure, and randomized presentation protocol are used for the main system comparison in Table 1 and the adaptation and component-ablation comparison in Table 2. These comparisons are administered as two separate listening tests.

For each selected prompt–target pair, all systems evaluated within the corresponding listening test are anonymized as "Model 1," "Model 2," and so on. The correspondence between anonymous identifiers and actual systems is hidden from evaluators, and the presentation order is randomized independently for every evaluation item. The main comparison contains five synthesis systems. The adaptation and component-ablation comparison contains nine configurations: the three adaptation strategies, the five component ablations, and ZeroOpera (Full). Figure 4 illustrates the interface and rating instructions used for the main comparison. The adaptation and ablation test follows the same rating protocol and reference-audio design, but presents nine anonymized system outputs per prompt–target pair.

Each listening test uses 20 sampled prompt–target pairs per prompt condition, and every synthesized sample is rated by all 15 evaluators. Thus, each system receives $20 \times 15 = 300$ ratings per prompt condition and evaluation dimension. The five-system main comparison yields 100 synthesized samples per prompt condition, whereas the nine-system adaptation and ablation comparison yields 180 synthesized samples under the in-domain opera-prompt condition. For the main comparison, which includes both in-domain opera and cross-domain pop prompt conditions, this corresponds to 200 synthesized samples and 3,000 sample-level evaluations. Table 11 summarizes both listening-test designs.

For each evaluation item, evaluators can listen to the acoustic prompt, the corresponding ground-truth recording, and the anonymized synthesized outputs included in the relevant listening test. The acoustic prompt serves as the reference for Timbre Similarity because the intended singer identity of each synthesized sample is determined by the prompt. The ground-truth recording provides a reference for the target lyrics, melody, and Peking Opera realization, but its singer identity is not treated as the target timbre. The acoustic prompt serves as the reference for Timbre Similarity because the intended singer identity of each synthesized sample is determined by the prompt. The ground-truth recording provides a reference for the target lyrics, melody, and Peking Opera realization, but its singer identity is not treated as the target timbre.

The ground-truth recordings are also rated for Audio Quality, Pronunciation Clarity, and Opera Flavor using the same five-point criteria. We do not evaluate Timbre Similarity for the ground-truth recordings because they are original recordings of the target passages rather than prompt-conditioned synthesis, and their singer identities generally differ from those of the acoustic prompts. The ground-truth scores therefore provide a perceptual reference for the quality, intelligibility, and operatic characteristics achievable by real recordings, rather than a reference for prompt-timbre similarity.

All synthesized samples are rated independently on a five-point scale along the following four dimensions:

- **Audio Quality (AQ):** the overall perceptual quality of the synthesized audio, including clarity, stability, and the absence of audible artifacts such as electronic noise, distortion, background noise, or unstable loudness.
- **Pronunciation Clarity (PC):** the intelligibility and clarity of the sung lyrics, including whether individual syllables can be recognized without severe slurring, omission, or unclear consonant–vowel articulation. Evaluators are instructed to assess intelligibility rather than the correctness of opera-specific pronunciation rules in this dimension.

Table 11: Summary of the subjective listening-test protocol per prompt condition. The main comparison is used for Table 1; the adaptation and ablation comparison is used for Table 2. Every synthesized sample and ground-truth recording is rated by all 15 evaluators.

Protocol Item	**Main Comparison**	**Adaptation/Ablation**
Prompt–target pairs	20	20
Synthesized systems	5	9
Synthesized samples	100	180
Evaluators	15	15
Ratings per synthesized sample	15	15
Ratings per system and dimension	300	300
Rating scale	1–5	1–5
Rated dimensions for synthesis	AQ, PC, TS, OF	AQ, PC, TS, OF
Rated dimensions for ground truth	AQ, PC, OF	AQ, PC, OF

- **Timbre Similarity (TS):** the perceived similarity between the singer identity of the synthesized sample and that of the corresponding acoustic prompt. Evaluators are instructed to focus on vocal identity and underlying timbre while disregarding differences in lyrics, melody, pitch range, and singing style as much as possible.
- **Opera Flavor (OF):** the degree to which the synthesized sample exhibits appropriate Peking Opera characteristics, including genre-specific articulation, resonance, vibrato, ornamentation, phrasing, and the realization of opera-specific pronunciations such as *Jian–Tuan* and *Shang-kou-zi*.

A score of 1 indicates very poor performance or very low similarity, whereas a score of 5 indicates excellent performance or very high similarity. The four dimensions are rated independently for each synthesized sample.

For each system, prompt condition, and evaluation dimension, we compute the mean opinion score over the 300 ratings obtained from 20 prompt–target pairs and 15 evaluators. Let $\bar{x}$, s, and N denote the sample mean, sample standard deviation, and number of valid ratings, respectively. The reported confidence interval is computed as

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{N}}.$$

Here, $N = 20 \times 15 = 300$ for each system, prompt condition, and evaluation dimension. The same calculation is applied to AQ, PC, and OF for the ground-truth recordings, while TS is not defined for the ground truth.

E Additional Analyses

E.1 Pitch Sensitivity of WavLM-Based Similarity

Speaker representations extracted by pretrained models are commonly used to evaluate speaker or timbre similarity in speech and singing synthesis. However, Peking Opera often involves a substantial pitch-register mismatch between the acoustic prompt and the target singing condition. This issue is particularly pronounced in the cross-domain setting, where a relatively low-register pop-singing prompt may be used to synthesize a high-register Peking Opera passage. Under such conditions, an embedding-based similarity metric may respond not only to singer identity, but also to differences in pitch register and vocal production.

We analyze this issue at two complementary levels. First, we conduct a system-level comparison over the cross-domain pop-prompt evaluation set. For SoulX-Singer and ZeroOpera, we compute WavLM-based speaker-embedding cosine similarity (SIM) under two pitch conditions. In the *Target-F0* condition, the synthesized output follows the F0 contour specified by the target musical condition. In the *Prompt-register* condition, the same output is pitch-shifted toward the register of its corresponding acoustic prompt. We report both SIM and F0 Frame Error (FFE), since shifting an output toward the prompt register may increase its embedding similarity while simultaneously moving its melody away from the intended target F0.

As shown in Table 12, pitch shifting produces the same trade-off for both SoulX-Singer and ZeroOpera. For SoulX-Singer, shifting the outputs toward their corresponding prompt registers increases the aggregate SIM from 0.639 to 0.856, while FFE deteriorates from 0.156 to 0.582. Similarly, the SIM of ZeroOpera increases from 0.655 to 0.879, whereas its FFE increases from 0.149 to 0.580. Thus, outputs that are considerably less faithful to the intended target melody can receive substantially higher embedding-similarity scores after their pitch registers are moved closer to those of the prompts.

The native output of VevoSing provides complementary system-level evidence. In the evaluated setting, VevoSing does not explicitly adjust its output to the target pitch register. Its synthesized F0 is generally lower than the target F0 and closer to the register of the acoustic prompt. VevoSing consequently obtains a relatively high SIM of 0.820 despite an FFE of 0.401. This observation is consistent with the controlled pitch-shifting results: proximity to the prompt register may increase WavLM-based similarity even when the synthesized output does not accurately follow the target melody.

Second, we perform a within-prompt analysis to isolate the effect of pitch register from variation across acoustic prompts. We select one fixed pop-singing prompt and collect 72 synthesized outputs conditioned on that prompt. Starting from their original target-F0 versions, we shift all outputs downward by the same amount, from 0 to 20 semitones in one-semitone increments. At each shift value, we compute SIM between every transformed output and the same fixed prompt and then average the scores over the 72 outputs.

As shown in Figure 5, the mean SIM generally increases with the magnitude of the downward pitch shift, rising from 0.0973 without pitch shifting to 0.5074 at a downward shift of 20 semitones. Although minor fluctuations occur at small shift values, the overall trend is pronounced and approximately monotonic over the larger shift range. Because the acoustic prompt and the set of source outputs are fixed throughout this analysis, the systematic score variation provides direct evidence that WavLM-based SIM is sensitive to pitch register.

The absolute SIM values in Figure 5 are not expected to match those in Table 12. The table reports aggregate results over the full cross-domain pop-prompt evaluation set, which contains multiple prompts and prompt–target pairs, whereas the figure reports a controlled within-prompt analysis for one fixed prompt and its associated set of 72 synthesized outputs. The two experiments therefore serve different purposes: the table demonstrates the practical metric trade-off at the system level, while the figure isolates the sensitivity of SIM to pitch shifting under a fixed prompt.

These findings reveal an important limitation of embedding-based similarity in cross-register Peking Opera synthesis. A system that accurately follows the target F0 may receive a relatively low SIM because its output register differs from that of the prompt. Conversely, an output that deviates from the target melody but remains closer to the prompt register may receive a higher SIM. We therefore treat WavLM-based SIM as a supplementary diagnostic metric rather than a standalone measure of timbre preservation. Listener-rated Timbre Similarity is used as the primary measure of perceived singer identity in our subjective evaluation.

E.2 Effect of Phonological-Category Conditioning

We conduct a controlled intervention to visualize the effect of phonological-category conditioning. Given the same lyrics, acoustic prompt, musical conditions, and sampling configuration, we compare two outputs from ZeroOpera. In the first output, all phonological-category labels are replaced with the Standard label, thereby disabling category-specific TPM modulation. In the second output, ZeroOpera uses the corresponding lexicon-derived phonological-category labels. The pinyin annotations and category boundaries in this visualization are obtained from the reference lyrics and annotations; they are not predictions from the Peking Opera ASR model.

Figure 6 shows the two generated spectrograms and their signed acoustic residual. Visible spectral differences appear around the characters assigned to the *Shang-kou-zi* category, including changes in harmonic-energy distributions and spectral structure. Differences are also visible in the surrounding context, which is expected because ZeroOpera predicts and enhances the complete spectrogram rather than independently replacing only the marked frames. This controlled comparison demonstrates that the phonological-category condition affects the generated acoustic realization. It does

Table 12: System-level analysis of pitch register and WavLM-based speaker-embedding similarity over the cross-domain pop-prompt evaluation set. Target-F0 denotes synthesis following the intended target melody, whereas Prompt-register denotes the corresponding outputs pitch-shifted toward the registers of their respective acoustic prompts. Native output denotes the original VevoSing output without pitch-shifting post-processing. SIM is the cosine similarity between WavLM embeddings extracted from each acoustic prompt and its corresponding synthesized output. The aggregate results show that proximity to the prompt register can increase SIM while degrading target-pitch accuracy.

Model	**Pitch Condition**	**FFE** $\downarrow$	**SIM** $\uparrow$
SoulX-Singer	Target-F0	0.156	0.639
SoulX-Singer	Prompt-register	0.582	0.856
ZeroOpera	Target-F0	0.149	0.655
ZeroOpera	Prompt-register	0.580	0.879
VevoSing	Native output	0.401	0.820

not, by itself, establish pronunciation correctness; the corresponding quantitative evidence is provided by the category-specific PER results in Table 3.

E.3 Qualitative Pronunciation and F0 Comparison

Figure 7 presents a representative comparison of the ground-truth recording and outputs from ZeroOpera, SoulX-Singer, StyleSinger, VevoSing, and YingMusic-Singer. Panels (a)–(f) show the spectrograms with the extracted F0 contours overlaid in cyan. The pinyin annotations displayed in these panels are decoded by our Pinyin-syllable Peking Opera ASR model rather than manually assigned from the spectrograms.

To determine the correctness marks, we compare the ASR-recognized pronunciation of each highlighted syllable with its expert-verified reference pronunciation. A check mark indicates that the recognized pronunciation agrees with the reference, whereas a cross mark indicates a mismatch. Category-specific syllables selected for comparison are highlighted in red, while the remaining syllable annotations are shown in gray. Thus, the marks visualize the same ASR-based pronunciation evidence used in the objective PER evaluation; they should not be interpreted as judgments made solely through visual inspection of the spectrograms.

In this example, ZeroOpera correctly realizes more of the highlighted category-specific pronunciations than the evaluated baselines while retaining an F0 trajectory close to that of the ground-truth recording. SoulX-Singer also closely follows the target melodic contour, but several of its highlighted pronunciations are recognized as inconsistent with the expert-verified reference. The signed F0 deviations in panel (g) further compare ZeroOpera and SoulX-Singer with the ground-truth contour. This example provides a qualitative illustration of the pronunciation and melodic behavior reflected in the corpus-level PER and FFE results; the main conclusions are based on the aggregate results rather than this individual case.

F Limitations and Future Work

ZeroOpera supports zero-shot prompt-timbre transfer after low-resource adaptation to Peking Opera, but it has several limitations. First, our automated G2P pipeline relies on a deterministic expert-curated lexicon covering Jian, Tuan, and Shang-kou-zi categories. It may not fully resolve context-dependent pronunciations or polyphonic characters. Future work will explore LLM integration to capture deeper semantic contexts for technique prediction. Second, because our structural prior is built upon a pop music foundation model, synthesizing extreme operatic vocal techniques—such as intense false vocal cord vibrations or highly complex rapid vibratos—can occasionally produce subtle acoustic artifacts. Finally, the iterative ODE solving process inherent to the Continuous Normalizing Flow (CNF) architecture results in slower inference speeds compared to single-step generative models.

To address these constraints, our future work will focus on three corresponding directions. We plan to integrate Large Language Models (LLMs) to automatically predict phonological technique tags directly from raw text, enabling a fully end-to-end pipeline. To enhance expressiveness and robustness in extreme acoustic regions, we aim to scale up the inpainting stream and introduce domain-specific acoustic discriminators. Furthermore, we will explore advanced acceleration techniques, such as flow distillation and consistency trajectory models, to significantly reduce inference latency and facilitate real-time interactive applications.



Figure 4: Interface used for the blind subjective listening evaluation. We sample 20 prompt–target pairs for each of the in-domain opera-prompt and cross-domain pop-prompt settings, and every audio sample is rated by the same 15 evaluators. The upper panel presents the Chinese instructions and five-point rating criteria for Audio Quality (AQ), Pronunciation Clarity (PC), Timbre Similarity (TS), and Opera Flavor (OF). The figure illustrates the main system-comparison interface, where evaluators can listen to the acoustic prompt, the corresponding ground-truth recording, and five anonymized system outputs for each prompt–target pair. The adaptation and component-ablation comparison in Table 2 uses the same rating criteria, reference-audio design, anonymization procedure, and randomized presentation protocol, but presents nine anonymized system outputs per prompt–target pair. The acoustic prompt is used as the reference for timbre similarity, while the ground-truth recording provides a reference for the target lyrics, melody, and Peking Opera realization. The ground-truth recording is also rated for AQ, PC, and OF, but not for TS. System identities are hidden, and the presentation order of all evaluated systems is randomized for each evaluation item.

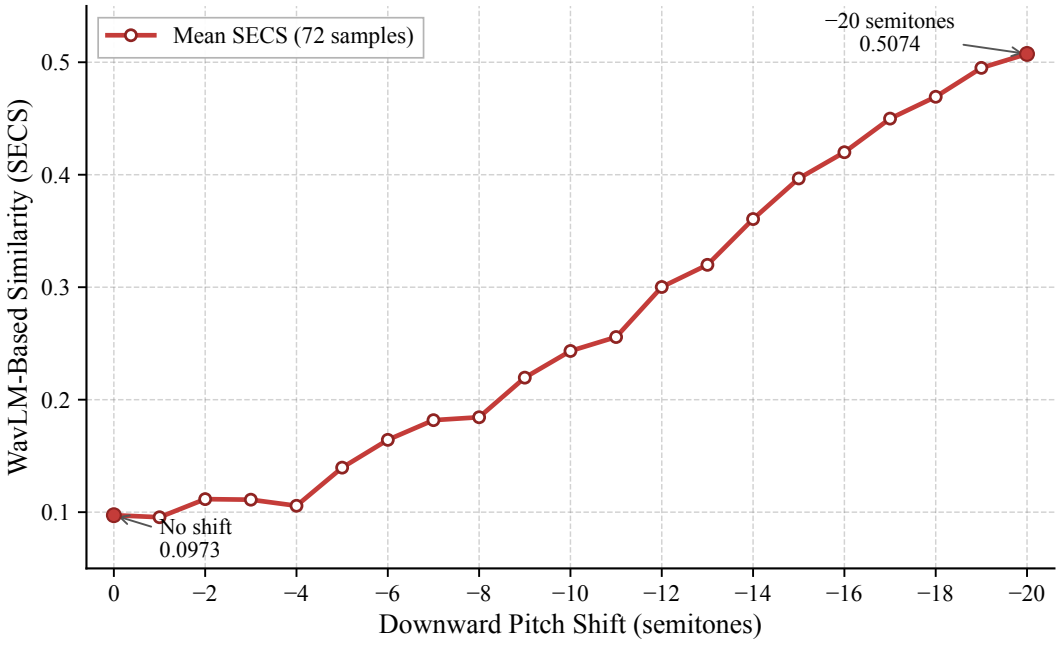


Figure 5: Within-prompt sensitivity of WavLM-based speaker-embedding cosine similarity (SIM) to progressive downward pitch shifting. We fix one pop-singing prompt and use 72 synthesized outputs conditioned on that prompt. The same downward shift, ranging from 0 to 20 semitones, is applied to all outputs at each step. Each point reports the mean SIM between the 72 transformed outputs and the fixed acoustic prompt. Mean SIM generally increases from 0.0973 without pitch shifting to 0.5074 at a 20-semitone downward shift. Unlike Table 12, which reports aggregate results over the full cross-domain evaluation set, this figure isolates pitch sensitivity under a single fixed prompt.

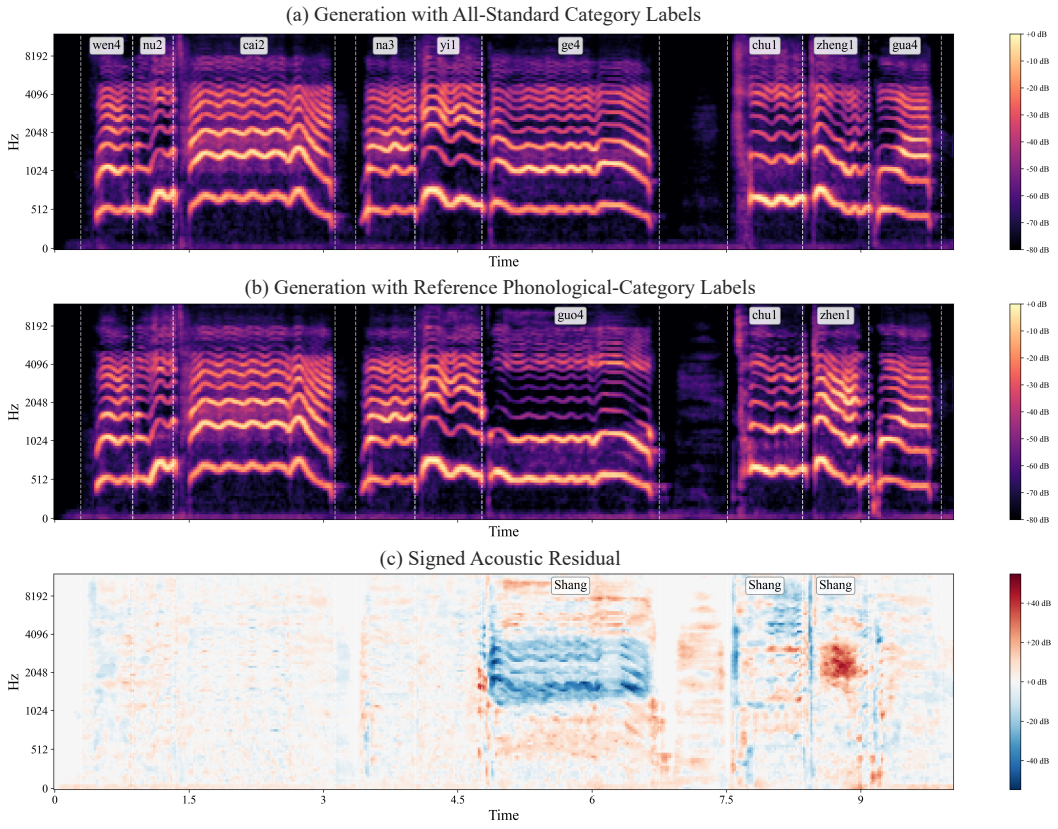


Figure 6: Controlled visualization of phonological-category conditioning with all other synthesis inputs fixed. (a) Spectrogram generated after replacing all phonological-category labels with the Standard label, thereby disabling category-specific TPM modulation. (b) Spectrogram generated using the corresponding lexicon-derived phonological-category labels. (c) Signed acoustic residual computed between (b) and (a), where warm and cool colors indicate positive and negative spectral-energy differences, respectively. The *Shang-kou-zi* regions are marked using reference lyric annotations. The pinyin labels and category boundaries in this figure are derived from the reference annotations rather than ASR predictions.

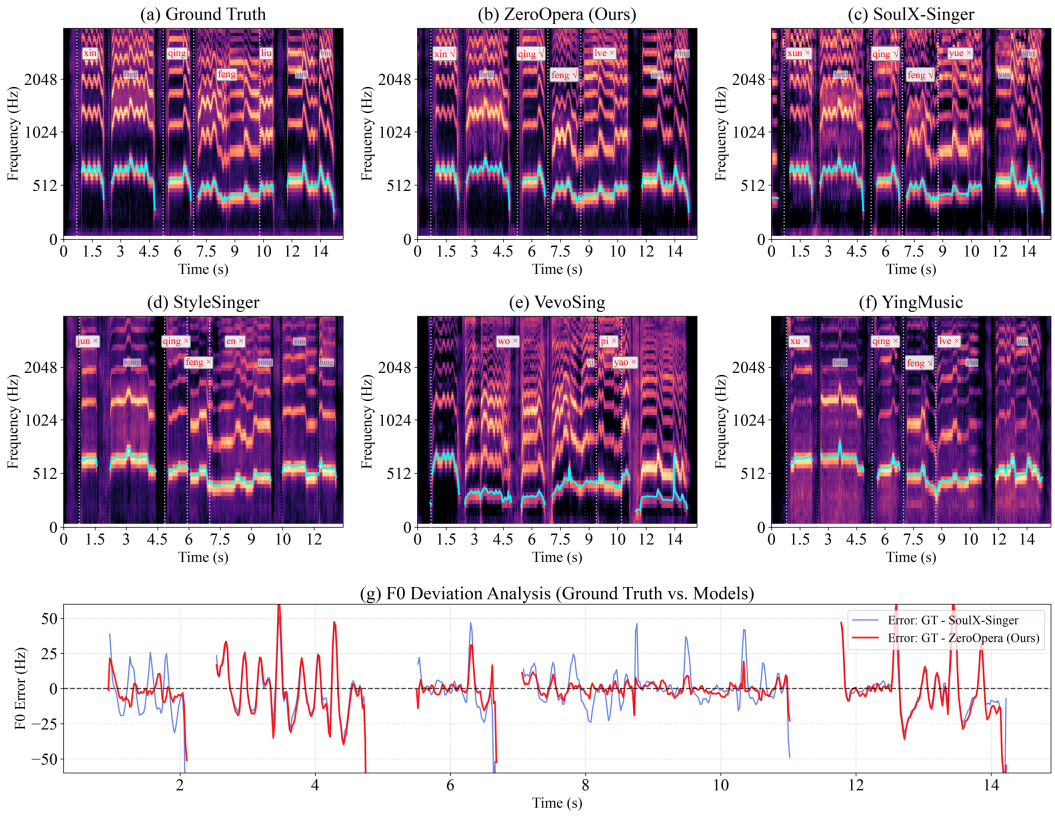


Figure 7: Qualitative comparison of spectrograms, F0 trajectories, and ASR-recognized pronunciations for a representative example. Panels (a)–(f) show the ground-truth recording and outputs from ZeroOpera, SoulX-Singer, StyleSinger, VevoSing, and YingMusic- Singer, with extracted F0 contours overlaid in cyan. The displayed pinyin annotations are generated by our Pinyin-syllable Peking Opera ASR model. Selected category-specific syllables are highlighted in red, while the remaining syllable annotations are shown in gray. A check mark indicates that the ASR-recognized pronunciation agrees with the corresponding expert-verified reference pronunciation, whereas a cross mark indicates a mismatch. Panel (g) reports the signed F0 deviations of SoulX-Singer and ZeroOpera relative to the ground-truth recording.