

CLASVS: Continuous-Latent Autoregression for Melody-Preserving Lyric Editing in Singing Voice Synthesis

Yizhong Geng^{1,2*}, Tian-Hao Zhang^{2,3*}, Chunfeng Wang², Wenxin Fu¹, Yingming Gao¹,
Ruimin Wang², Zhou Pan², Kun Zhan^{2†}, Liang Li³, Ya Li^{1†}

¹Beijing University of Posts and Telecommunications, Beijing, China

²Li Auto Inc., Beijing, China

³Tsinghua University, Beijing, China

Abstract

Reference-conditioned melody-preserving lyric editing replaces words while retaining a performance’s timing, singer identity, and naturalness. Continuous-latent autoregression avoids finite codebooks and offers stepwise generation with learned stopping. Editing creates a conflict absent from ordinary reconstruction: training pairs reference cues with original lyrics, whereas inference asks revised lyrics to override source-lyric-correlated cues; one source-following patch can propagate through AR history. We introduce CLASVS. Its State–Control–Transition (SCT) routing keeps target-lyric and reference-melody controls persistent, returns semantic feedback on phonetic progress to the causal planner, and confines the previous latent patch to the local Transition. Progressive State–Control Grounding (PSCG) learns this contract through paired-edit-free, content-consistent Mandarin reconstruction. On two Mandarin benchmarks, CLASVS improves all four operations over discrete-AR Vevo2 and reduces macro-PER by 46.2%, while maintaining melody, singer similarity, and perceptual quality. Together, these results establish a strong continuous-AR operating point for score-annotation-free lyric edits and a basis for broader stepwise control. Audio demonstrations are available on our project page: <https://pipediperg.github.io/Li lyric-SVS/>.

Introduction

Melody-preserving lyric editing changes the words of a singing performance while retaining its melody, timing, singer identity, and naturalness. Score-conditioned singing voice synthesis provides precise control through aligned phonemes, notes, and durations, but such annotations are costly for arbitrary recordings (Liu et al. 2022; Zhang et al. 2023; Chunhui, Zeng, and He 2023; Kim, Cho, and Lee 2024; Wang et al. 2024; Zhang et al. 2024b; Wang et al. 2022; Zhang et al. 2022a, 2024a; Bai et al. 2026). Reference-conditioned lyric editing instead accepts a singing recording and revised lyrics directly (Zhang et al. 2022b, 2026; Hao et al. 2026; Park, Im, and Nam 2026). Removing score annotations, however, creates a selective-preservation problem: the model must retain the performance characteristics carried by the reference while rejecting its source lyrics.

Two architectural axes shape selective preservation. Discrete codecs compact acoustic streams for autoregressive (AR) modeling (Zeghidour et al. 2022; Défossez et al. 2023; Kumar et al. 2023; Wu et al. 2024; Zhang et al. 2026), but finite codebooks may lose subtle variation, whereas continuous or tokenizer free latent models reduce dependence on fixed codebooks and offer strong fidelity or scalable generation (Le et al. 2023; Shen et al. 2024; Mehta et al. 2024; Meng et al. 2025; Yang et al. 2025b). Generation order adds a complementary tradeoff: continuous nonautoregressive (NAR) systems such as YingMusic-Singer-Plus (YingMusic+) enable parallel generation, whereas AR provides stepwise state and learned stopping. Existing continuous AR models such as DiTAR and SemaVoice mainly target content consistent synthesis (Jia et al. 2025; Wang et al. 2026), leaving unresolved the supervision mismatch in reference-conditioned lyric editing under reconstruction only supervision.

This mismatch is especially severe in singing. Recent continuous or tokenizer free AR speech systems learn content consistent synthesis (Lian et al. 2026; Peng et al. 2025; Zhou et al. 2025), while unified speech editors rely on explicit instruction based editing supervision (Yan et al. 2025); neither setting directly provides singing counterfactuals, because resinging different lyrics also changes timing, phrasing, and acoustics. We therefore train on same-recording tuples whose reference, transcript, and acoustic target come from one performance, but this scalable reconstruction never exposes a conflict between reference and target lyrics or an edit template. At inference, revised lyrics create this unseen conflict, and one patch that follows the source lyrics can bias later AR history, requiring generalization without paired edits rather than memorized edit transformations.

CLASVS addresses this challenge with a State–Control–Transition (SCT) editing contract. Target lyrics and reference melody remain persistent *Control*; a frozen causal semantic encoder returns feedback on realized phonetic progress as recurrent *State* (Qian et al. 2022; Chen et al. 2022; Li et al. 2024); and the previous AudioVAE latent patch enters only the local *Transition*. This separates global lyric planning from local acoustic continuity. Latent patches never re-enter the causal planner, preserving continuity without biasing lyric planning. Progressive State–Control Grounding (PSCG) grounds this routing through paired-edit-free,

*Equal contribution.

†Corresponding authors.

content-consistent reconstruction. State Grounding pretrains then freezes the semantic encoder, Control Grounding perturbs melody cues and the previous latent patch, and Task Grounding transfers Mandarin coverage to curated singing. It needs no paired counterfactual edit recordings.

We evaluate two Mandarin benchmarks under a common automatic-input protocol using reference audio and target lyrics. Vevo2 is the primary comparator because it shares CLASVS’s sequential interface but predicts discrete acoustic tokens rather than continuous AudioVAE latent patches (Zhang et al. 2026). CLASVS improves all four operations over Vevo2, reducing macro-PER from .0699 to .0376 (46.2%); the profile transfers to the public benchmark. Ying-Music+ is the complementary continuous-NAR comparator (Hao et al. 2026): it remains stronger on substitution, global duration, and speed, whereas CLASVS leads deletion, insertion, macro-PER, voiced-onset MAE, naturalness, and lyric intelligibility. The two comparators expose complementary system trade-offs. Matched ablations provide within-stack SCT/PSCG evidence. Our contributions are:

- **SCT Editing Contract.** We formulate editing as a conflict among persistent target-lyric and reference-melody controls, phonetic progress, and the previous latent patch, and route them separately.
- **Paired-Edit-Free PSCG.** We learn this SCT routing through three-stage, content-consistent Mandarin reconstruction, without paired counterfactual edit recordings.
- **Continuous-AR Editing Results.** Under the common automatic-input protocol, CLASVS reduces macro-PER by 46.2% versus discrete-AR Vevo2 across two Mandarin benchmarks, improving all four operations while preserving melody, similarity, and perceptual quality.

Method

Inputs, Time Bases, and Duration

Let $\mathbf{r}$ be reference audio and $\mathbf{x} = (x_1, \dots, x_{L_x})$ the target lyric-token sequence. A frozen melody tokenizer maps $\mathbf{r}$ to $\mathbf{m} = (m_1, \dots, m_{L_m})$ at approximately 6.25 Hz using a 512-entry codebook; it is trained from chroma-derived inputs so that the control emphasizes pitch and timing rather than lexical content. A frozen CAM++ encoder (Wang et al. 2023) extracts a 192-D singer vector $\mathbf{e}$. A frozen Ming-UniAudio AudioVAE (Yan et al. 2025) represents 16-kHz target audio as $\mathbf{z}^* \in \mathbb{R}^{T \times 64}$ at 50 Hz.

We group $P = 5$ frames into $\bar{\mathbf{z}}_k^* \in \mathbb{R}^{5 \times 64}$, so the acoustic recurrence runs at 10 Hz. A final short group is zero padded and retained as one patch in loss and decoding; no frame-level trimming is applied, giving at most 100-ms endpoint granularity, and this tail remains in every metric and listening item. Melody and acoustic tokens are *not* one-to-one aligned: the melody prefix remains in the planner cache while semantic feedback marks generated progress. The stop head learns output patch count while target/source syllable counts may differ under reference timing cues. Inference is bounded by $\min(K_{\text{ref}} + 8, 256)$ patches only as a rollout safeguard.

Training Tuples and the Content Conflict

Each tuple $(\mathbf{r}, \mathbf{x}, \mathbf{z}^*)$ comes from one recording: reference and target audio are the same performance, and $\mathbf{x}$ is its transcript. Thus training uses content-consistent reconstruction, never unavailable recordings of one melody sung with counterfactual lyrics. At evaluation, $\mathbf{x}$ changes while reference-derived controls remain tied to the reference audio. This paired-edit-free training turns lyric editing into a zero-paired-edit generalization test: success requires following revised lyrics despite residual source-lyric-correlated evidence, rather than fitting edit templates.

SCT/PSCG make that generalization learnable: target lyrics remain explicit, melody spans are perturbed, and an imperfect previous latent patch prevents unconditional predecessor copying. Empirically, output phones across all four operations are scored against both target and source under one max-length NED denominator. Their target-preference margin tests the realized shortcut directly rather than treating tokenizer purity as an assumption.

State–Control–Transition Routing

Figure 1 summarizes the SCT routing. Let $\mathbf{R}_k$ denote the recurrent context before patch k :

$$\mathbf{R}_k = (\mathbf{K}_k^{\text{plan}}, \mathbf{K}_k^{\text{sem}}, \bar{\mathbf{z}}_{k-1}),$$

$$p_\theta(\bar{\mathbf{z}}_{1:K} \mid \mathbf{x}, \mathbf{m}, \mathbf{e}) = \prod_{k=1}^K p_\theta(\bar{\mathbf{z}}_k \mid \mathbf{R}_k, \mathbf{x}, \mathbf{m}, \mathbf{e}). \quad (1)$$

The update $\mathcal{U}(\mathbf{R}_k, \bar{\mathbf{z}}_k)$ appends the new patch to the semantic encoder, pools its five states into semantic feedback, appends that feedback to the planner cache, and exposes the patch as the previous latent patch at $k + 1$. At $k = 1$, both recurrent paths are initialized with zeros.

Lyric and melody tokens prefill a Qwen3-0.6B causal planner (Yang et al. 2025a). After the first step, only semantic feedback $\mathbf{a}_{k-1}$ is appended:

$$(\mathbf{h}_k, \mathbf{K}_k^{\text{plan}}) = \mathcal{P}_\theta(\mathbf{a}_{k-1}, \mathbf{K}_{k-1}^{\text{plan}}). \quad (2)$$

The cache retains the full controls and all earlier feedback. A separate linear head predicts continue/stop from $\mathbf{h}_k$. Singer identity bypasses the planner and modulates each Flow-DiT block through adaptive layer normalization.

Control serialization and time bases. Lyrics and melody occupy distinct cached spans rather than a fused pseudo-score. Melody tokens retain reference-frame timestamps, whereas target lyric tokens retain sequence order without requiring note, duration, or edit-boundary labels. A query for patch k can attend to both complete control spans and semantic feedback through $k - 1$, but never to a future generated patch. AudioVAE frames use a 20-ms stride and five consecutive frames form one 100-ms transition. If the final transition is short, it is zero padded and retained as a complete patch through decoding. This contract keeps score-annotation-free controls globally available while making generated acoustic evidence strictly causal.

A frozen Whisper-style 32-layer, 1280-D causal audio encoder (Radford et al. 2023) provides the compact path; its

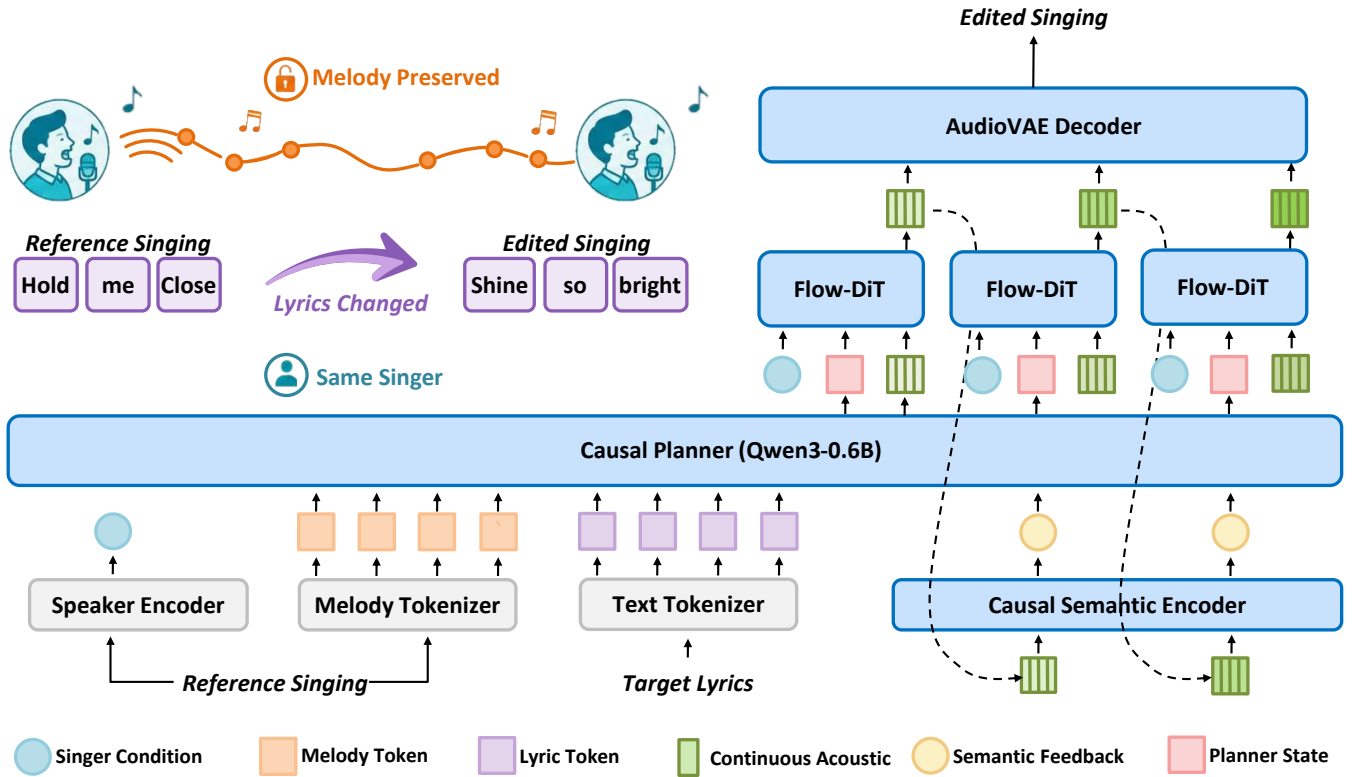


Figure 1: CLASVS generates 64-D continuous AudioVAE latent patches every 100 ms. Target lyrics and reference-melody tokens persist in the causal planner’s cache. Each patch updates semantic feedback and becomes the previous latent patch for the next step; the concatenated sequence is decoded once. Repeated Flow-DiT blocks share parameters.

states are projected to 1024-D before five-frame pooling:

$$\begin{aligned}
 (\mathbf{U}_{k-1}, \mathbf{K}_k^{\text{sem}}) &= \mathcal{E}_{\text{sem}}(\bar{\mathbf{z}}_{k-1}, \mathbf{K}_{k-1}^{\text{sem}}), \\
 \mathbf{a}_{k-1} &= P^{-1} \sum_{p=1}^P \text{Proj}(\mathbf{U}_{k-1,p}). \quad (3)
 \end{aligned}$$

In parallel, the previous latent patch enters an 8-block Flow-DiT that combines conditional flow matching with a transformer denoiser (Lipman et al. 2023; Peebles and Xie 2023). Each transition receives one flow-time/planner token, five predecessor tokens, and five noisy current-patch tokens. The two paths are complementary views of the same realized patch: semantic feedback supports long-range progress, whereas the previous latent patch preserves local continuity.

The paths share each generated patch but differ in reach: semantic feedback persists in the planner cache for all later steps, whereas the previous latent patch conditions only the next Flow-DiT transition. This asymmetric lifetime separates global planning memory from local acoustic continuity.

The semantic route compresses each five-frame patch into one 1024-D feedback vector, while the local route retains the complete 5×64 patch for only the following transition. This design lets the planner accumulate content progress without an unrestricted acoustic history, preserves causal local continuity in Flow-DiT, and uses incremental caches to avoid re-encoding the generated prefix.

Progressive State–Control Grounding

PSCG grounds the SCT editing contract (Figure 2). *State Grounding* pretrains the causal semantic encoder with $\mathcal{L}_{\text{state}} = \mathcal{L}_{\text{ASR}} + 5\mathcal{L}_{\text{MSMel}}$ before freezing it, so semantic feedback tracks realized content rather than copying acoustics. *Control Grounding* applies history dropout ($p_h = .10$) to the teacher-forced previous latent patch and masks reference-melody spans ($p_m = .15$), forcing recovery from imperfect local evidence while retaining both controls. *Task Grounding* schedules 72K broad-singing, 24K speech-assisted, and 24K curated-singing updates, supplying coverage before editing fidelity (Geng et al. 2026).

Each stage ablation disables only its defining operation while preserving width, 120K updates, optimizer, Mandarin pool, manifest, and checkpoint rule. Without State Grounding, the semantic encoder is randomly initialized and trained jointly for all 120K updates; no untrained encoder is frozen. Removed inputs are zeroed. The matched unified-context AR projects semantic feedback and the previous latent patch to the planner width, combines them through one shared recurrent route visible to both causal planning and local generation, and otherwise retains the same controls, AudioVAE, planner, Flow-DiT, data, parameters, and updates. Pretrained-frozen/pretrained-joint/random-joint/random-frozen PER is .0376/.0391/.0511/.0918. History-dropout, melody-mask, joint-corruption, and curriculum-order removals yield PER

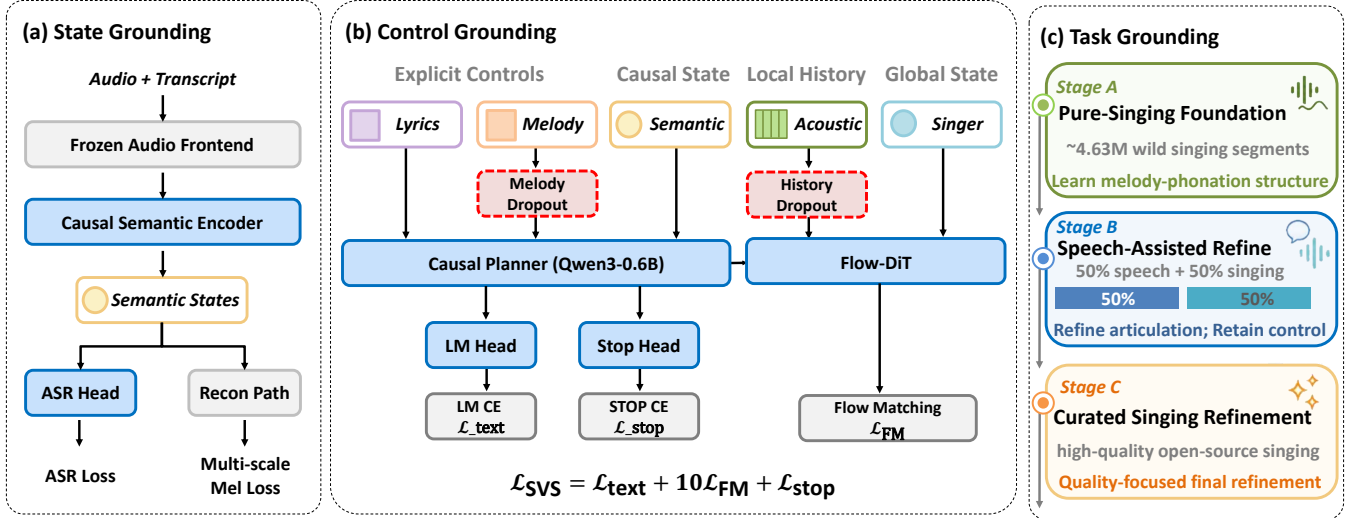


Figure 2: Progressive State-Control Grounding (PSCG). State Grounding pretrains the semantic encoder for phonetic progress; Control Grounding applies history dropout to the previous latent patch and masks melody while retaining lyrics; Task Grounding moves from broad to curated Mandarin singing.

.0441/.0419/.0468/.0415; full tables are in the supplement.

For target patch $\bar{\mathbf{z}}_k^*$, noise $\epsilon_k \sim \mathcal{N}(0, I)$, and $\tau \sim \mathcal{U}(0, 1)$, conditional flow matching uses $\mathbf{y}_{k,\tau} = (1 - \tau)\epsilon_k + \tau\bar{\mathbf{z}}_k^*$ and target velocity $\mathbf{v}_k^* = \bar{\mathbf{z}}_k^* - \epsilon_k$. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{text}} + 10\mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{stop}}. \quad (4)$$

$\mathcal{L}_{\text{text}}$ is causally shifted cross-entropy on nonpadding target-lyric positions during prefix teacher forcing; melody tokens, separators, and recurrent semantic feedback are excluded. It is a planner-side target-content regularizer, not direct acoustic or counterfactual supervision; its effect on generated audio is tested by ablation. $\mathcal{L}_{\text{stop}}$ is balanced cross-entropy with the final valid patch labeled stop. Joint lyric-melody-singer dropout ($p = .10$) trains the unconditional guidance branch (Ho and Salimans 2022), which supports classifier-free guidance during sampling. We optimize the planner, adapters, Flow-DiT, and heads while freezing all frontends. Each patch starts from Gaussian noise and uses 24 Euler steps with guidance 2.0. We emit one candidate, stop at posterior .5 or the rollout budget, concatenate the patches, and decode once; the companion archive includes executable reference code and module cards.

Experiments

Data, Comparators, and Statistics

Training audio comprises about 8,000 hours of permission-documented Mandarin singing, 2,000 hours of Mandarin Emilia speech (He et al. 2024), and 300 hours of curated open-source singing; all data are Mandarin. Study-specific evaluation recordings came from consenting singers and contain no private information; listeners consented, and responses were de-identified. Three 120K-update seeds use BF16, ZeRO-2, AdamW, and global batch 128. The checkpoint has 0.765B trainable/2.115B total parameters, includ-

ing the frozen 1.350B AudioVAE; versioned feature extractors are excluded. Cold-start/steady-tail RTF and memory cover the operational pipeline; external counts follow publisher scope.

Vevo2 is the primary discrete-AR comparator: both systems derive melody automatically and decode sequentially under the common automatic-input protocol, but Vevo2 predicts discrete acoustic tokens and CLASVS predicts continuous AudioVAE latent patches. YingMusic+ is the complementary continuous-NAR comparator with parallel inference. These comparisons establish operating profiles; causal claims rely on matched SCT/PSCG tests in Table 5.

CLA-LyricEdit-320 contains 320 edits from 160 excerpts spanning 80 songs and 40 singers, balanced across partial-line substitution (PSub), full-line substitution (FSub), deletion, and insertion. PSub replaces a contiguous 2–6-syllable span while retaining context, FSub replaces every lexical syllable in a line, and deletion and insertion remove or add 2–6 syllables. Two annotators verify targets and operations ($\kappa = .93$); all splits are song-disjoint.

FireRedASR-AED-L transcribes each output; pypinyin yields phones for PER and normalized edit distance (NED). Source-lyric reversion rate (SrcRev) counts outputs closer to source than target under NED. Melody metrics are common-voiced RMVPE log-F0 Pearson correlation (FPC) (Wei et al. 2023), voiced-onset MAE, and reference-duration deviation (RefDur; DUR-A/R1); singer similarity (SIM) is WavLM-base-plus-SV cosine (Chen et al. 2022). A pre-generation stratified draw yields 80 blinded audio outputs spanning five conditions, evaluated by twenty native Mandarin listeners for naturalness (N-MOS), melody (M-MOS), and lyric intelligibility (L-MOS). We report operation-balanced macro-PER, 10,000 song-clustered draws, listener/item random effects (Kirkland et al. 2023; Camp et al. 2023), and Holm correction; evaluator and listener details appear in the supplement.

System	Representation / generation	Count / basis	Required control	Quality inference	Status
Vevo2	Discrete / AR	1.97B / publisher	Ref. audio + target lyrics	Official recipe	Released checkpoint
YingMusic-Singer-Plus	Continuous / NAR	0.73B / publisher	Ref. audio + target lyrics	Official recipe	Released checkpoint
CLASVS	Continuous / AR	0.765B / 2.115B	Ref. audio + target lyrics	24 steps; CFG 2.0	Ours

Table 1: Released comparators and CLASVS share the automatic-input protocol and emit one candidate. CLASVS count is trainable/total generative-stack size, excluding separately versioned frozen feature extractors; external counts retain publisher scope. Quality settings appear here and runtime settings in Table 4.

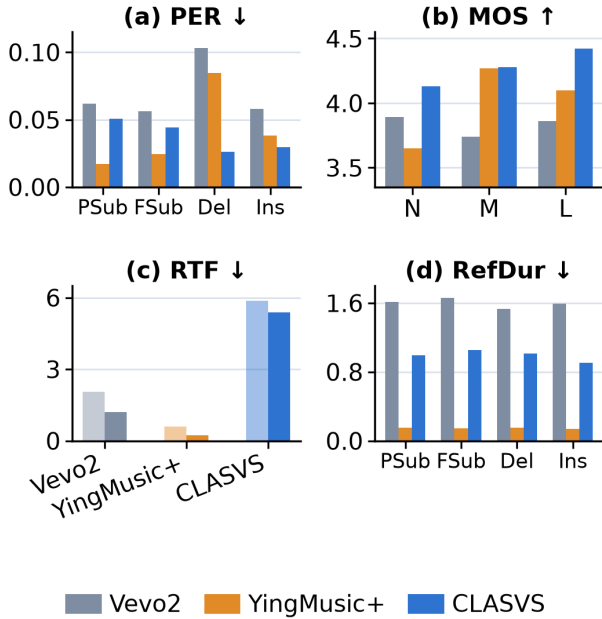


Figure 3: Common automatic-input comparison across all three systems: operation PER, listener ratings, cold-start/steady-tail RTF (lighter/solid), and RefDur.

SCT Control Audit

Table 2 exposes the lexical-change baseline under a common denominator. Output–target NED is low in every operation (.0252–.0490), whereas output–source NED closely tracks the required target–source change overall (.4050 versus .3965). All four target–preference margins are positive (.2248–.7209; overall .3685). Because chroma-derived melody tokens need not be lexically invariant, this behavioral audit asks whether outputs follow the target despite any residual reference cue. Low target NED and positive margins reject simple source copying while remaining sensitive to additional errors.

Public-Benchmark Replication

Table 3 uses the frozen released LyricEditBench split and evaluator (Hao et al. 2026), restricted to its four Mandarin-only operations. CLASVS improves all four operations over Vevo2 and lowers macro-PER from .0761 to .0425; against YingMusic+, it improves deletion and insertion while YingMusic+ remains stronger on PSub and FSub (.0425 versus

.0443 overall). This reproduces the discrete-AR gain and operation-specific profile within the Mandarin-only scope.

Editing Accuracy, Quality, and Identity

Table 4 and Figure 3 show a broad gain over the discrete-AR comparator: relative to Vevo2, CLASVS reduces macro-PER by 46.2%, improves every operation, FPC, SIM, RefDur, voiced-onset MAE, all three MOS scales, and peak memory, at the cost of slower RTF. Against the continuous-NAR YingMusic+ comparator, CLASVS has lower macro-PER by $-.0035$ [$-.0056, -.0013$] and better deletion/insertion accuracy (.0279 versus .0615 mean). CLASVS N-MOS rises from 3.65 to 4.13 ($+.48$ [$+.34, +.62$], adjusted $p < .001$), reflecting a marked listener preference in naturalness and acoustic clarity. YingMusic+ remains stronger on substitution, global RefDur, and speed. The comparison therefore supports a complete-system continuous-AR operating point rather than uniform or one-factor architectural dominance.

PSCG and Inference Diagnostics

PSCG development sweeps favor moderate latent-patch corruption: increasing p_h from .10 to .20 lowers PSub PER from .0505 to .0437 and SrcRev from .081 to .055, but worsens macro-PER to .0397; .10 is the locked trade-off. Removing State Grounding changes PER/FPC to .0511/.9168, supporting semantic feedback as an explicitly grounded phonetic-progress State.

Table 5 uses three seeds. Full versus matched unified improves PER by $-.0071$ [$-.0085, -.0057$], FPC by $+.0131$ [$+.0098, +.0164$], and SrcRev by $-.037$ [$-.048, -.026$]. The route removals are complementary rather than interchangeable: removing semantic feedback most strongly harms lyric accuracy and target following, whereas removing the previous latent patch most strongly harms melody preservation. Grounding-stage removals show the same division of labor at training time. Removing $\mathcal{L}_{\text{text}}$ raises PER to .0458 and SrcRev to .137, further connecting target organization in the causal planner to generated audio.

Configuration discipline. All matched rows retain the same frontends, data, width, 0.765B trainable budget, 120K updates, manifest, and checkpoint rule; three-seed PER SD across routing/stage rows is .0009–.0016. The explicit 2×2 State-Grounding check separates initialization from freezing; its one-time semantic pretraining is an additional cost outside the downstream schedule, so this check is not compute matched. Separate removals isolate history dropout, melody masking, and curriculum order. Released checkpoints remain

Evaluation	PSub	FSub	Del	Ins	Overall	Interpretation
Output–target NED↓	.0490	.0430	.0252	.0289	.0365	Desired content
Output–source NED↑	.3107	.7639	.2500	.2954	.4050	Rejected reference words
Target–source NED	.3020	.7600	.2380	.2860	.3965	Required lexical change
Target–preference margin↑	.2617	.7209	.2248	.2665	.3685	Output–source minus output–target

Table 2: Counterfactual content audit. NED is phone edit distance divided by the larger sequence length; target-preference margin is output–source minus output–target. No-edit output–target/source is .018/.019.

System	PSub	FSub	Del	Ins	Macro-PER↓
Vevo2	.0664	.0607	.1123	.0649	.0761
YingMusic-Singer-Plus	.0214	.0186	.0946	.0426	.0443
CLASVS	.0508	.0435	.0398	.0357	.0425

Table 3: Four Mandarin-only LyricEditBench operations under the common automatic-input protocol and public evaluator. Cross-lingual categories are excluded; the discrete-AR gain and operation-specific profile transfer.

System	PSub	FSub	Del	Ins	Macro-PER [95% CI]↓	FPC↑	SIM↑	RefDur./Onset↓	C/S RTF↓	ΔGiB↓
Vevo2	.0617	.0563	.1033	.0581	.0699 [.0617,.0781]	.7440	.892	1.6009%/69.58ms	2.060/1.197	8.62
YingMusic-Singer-Plus	.0171	.0245	.0847	.0382	.0411 [.0356,.0469]	.9340	.906	.1464% /18.80ms	.594/.226	6.74
CLASVS	.0505	.0442	.0260	.0298	.0376 [.0326,.0429]	.9410	.914	.9923%/7.68ms	5.858/5.380	5.82

Table 4: One candidate per item/run on 320 Mandarin edits. RefDur is DUR-A/R1 absolute output/reference duration deviation; Onset is voiced-onset MAE. CLASVS intervals resample songs and three training runs; released comparators condition on one checkpoint. C/S is cold-start/steady-tail RTF; ΔGiB is incremental peak memory in a separate runtime benchmark.

Configuration	PER↓	FPC↑	SrcRev↓
Full SCT + PSCG	.0376	.9410	.081
No semantic-feedback route	.0564	.9205	.154
No latent-patch route	.0492	.8896	.086
Unified-context AR	.0447	.9279	.118
– State Grounding	.0511	.9168	.143
– Control Grounding	.0468	.9242	.121
– Task Grounding	.0439	.9298	.102

Table 5: SCT and PSCG checks under one locked continuous-AR protocol. Unified-context AR is matched in frontends, planner, generator, data, width, parameters, and updates; lower SrcRev is better.

system comparators, not causal ablations; runtime settings are separately labeled.

Inference-time sensitivity checks. Figure 4 creates held-out counterfactual conflicts. Target-lyric swaps fix the reference-melody control, singer vector, flow noise, and roll-out budget; matched reference-melody swaps fix all other inputs. Order-breaking and replacement tests agree: disrupting semantic feedback lowers the target-lyric control margin, disrupting latent patches lowers the reference-melody control margin, and clustered intervals exclude zero. A progress probe favors semantic feedback.

Analysis and Discussion

System-level evidence. Vevo2 and YingMusic+ are the discrete-AR and continuous-NAR comparators under the

System	N-MOS	M-MOS	L-MOS
Vevo2	3.89 [3.78,4.00]	3.74 [3.63,3.85]	3.86 [3.75,3.97]
YingMusic+	3.65 [3.55,3.74]	4.27 [4.19,4.35]	4.10 [4.01,4.19]
CLASVS	4.13 [4.04,4.22]	4.28 [4.20,4.36]	4.42 [4.34,4.50]
Δ vs. YingMusic+	+ .48	+ .01	+ .32
Adjusted <i>p</i>	< .001	.79	< .001

Table 6: Ordinal listener–item analysis over 80 blinded audio outputs; deltas use YingMusic+, and the M-MOS difference is not detectable.

common automatic-input protocol. They establish complete-system profiles; matched unified-context, route, and stage interventions support SCT/PSCG attribution. The public Mandarin benchmark reproduces the operation-specific profile.

Why the two comparators are complementary. Vevo2 shares CLASVS’s sequential, automatically extracted melody-token interface but predicts discrete acoustic tokens instead of continuous AudioVAE latent patches; CLASVS improves every reported editing and preservation measure. YingMusic+ generates continuous features in parallel and leads substitution and RTF; CLASVS leads deletion, insertion, voiced-onset MAE, N-MOS, L-MOS, and macro-PER. Together with the matched unified-context ablation, these complementary profiles across the evaluated systems support—without factorial attribution—high-fidelity step-wise control with learned stopping.

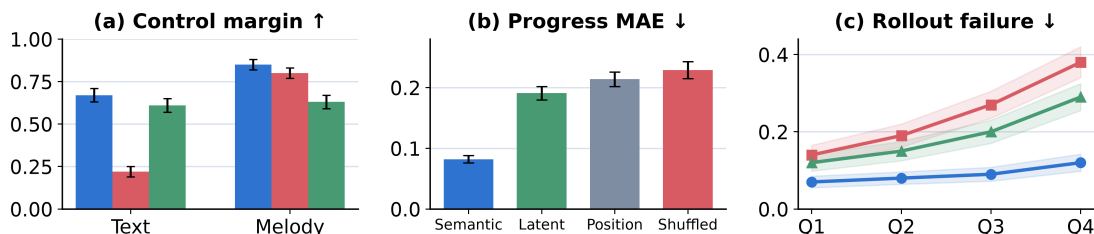


Figure 4: SCT diagnostics (song-clustered 95% intervals): paired control swaps, a progress probe, and failure by output-length quartile. Blue/red/green denote full/semantic-feedback-shuffled/latent-patch-shuffled; gray is position-only.

Zero-paired-edit generalization. Resinging revised words changes the performance, so paired counterfactual edit recordings are not naturally observable. Training never reveals the conflict, edit-span alignment, or operation template; CLASVS applies its SCT editing contract only at test time. The .3685 target-preference margin and deletion/insertion gains evidence behavior beyond reconstruction; progress-probe MAE (.082 versus .191), route shuffles, and history dropout independently support the contract across held-out cases.

Robustness across edit demands. CLASVS improves every Vevo2 operation, especially length-changing edits. Against YingMusic+ it leads deletion/insertion but trails substitution; the same profile recurs on the public benchmark under matched evaluation. Output–target/source NED rules out retained-word shortcuts; RefDur and voiced-onset MAE track reference pacing. Thus lexical, temporal, and replicated evidence complement macro-PER.

State size and runtime. Here *compact* is temporal: one 1024-D semantic-feedback vector enters the cache per patch; the previous latent patch stays one patch wide. Cache growth remains linear during long rollouts and incremental memory is 5.82 GiB; long inputs require future compression.

Metric and deployment trade-offs. CLASVS improves all Vevo2 measures under matched inputs, at slower RTF. Against YingMusic+ it leads macro-PER, deletion/insertion, naturalness, and voiced-onset MAE but trails substitution, global RefDur, and RTF, so it favors offline use.

What the stopping audit adds. The stop head determines realized patch count across all operations; $\min(K_{\text{ref}} + 8, 256)$ is only a guard. Budget-hit rates are 0/0/0/1.25% for PSub/FSub/Del/Ins; premature/delayed-stop rates never exceed 1.25/2.50%. Here “length change” refers to target/source syllable count, not a required change in global audio duration: insertion and deletion may redistribute local rate while retaining the reference timeline. Together with output–source NED, these rates show that deletion accuracy is not obtained by truncation or retained words.

Transfer audits. Across 24 post-freeze singers (192 edits), CLASVS/YingMusic+ gives PER .0472/.0521, FPC .934/.928, SIM .905/.898, and voiced-onset MAE 9.4/21.6 ms. The public benchmark retains YingMusic+’s

substitution lead and CLASVS’s deletion/insertion advantage, confirming the operation-specific profile.

Perceptual interpretation. CLASVS improves N-MOS by +.48 [+ .34, +.62] and L-MOS by +.32 [+ .20, +.44] over YingMusic+ under the same listening protocol; M-MOS differs by an undetectable +.01. N-MOS explicitly includes clarity and artifacts, so 4.13 versus 3.65 supports clearer output; L-MOS independently confirms target intelligibility. These perceptual gains complement deletion/insertion accuracy without treating the M-MOS null as equivalence.

Limitations, Reproducibility, and Ethics

Evaluation uses one recognizer and reference-timing metrics on Mandarin 2–6-syllable edits; cross-lingual and long-edit validation remain open. Upon publication, we will release training, inference, and evaluation code/configurations, per-item scores, CLA-LyricEdit-320 annotations and permitted audio, plus identity-authorized inference checkpoints; restricted training audio cannot be redistributed.

Recognizable-voice editing can enable impersonation (Zang et al. 2024). Source-level permissions, opt-out and removal procedures, identity-authorized access, and localized provenance therefore govern data use, checkpoint release, and responses to misuse reports (Roman et al. 2024).

Conclusion

CLASVS establishes continuous-latent autoregression for melody-preserving Mandarin lyric editing without manual score annotations. Its SCT editing contract separates target-lyric and reference-melody controls, semantic feedback on phonetic progress, and the previous AudioVAE latent patch; PSCG grounds this routing through content-consistent reconstruction. It follows revised lyrics without paired counterfactual edit recordings. Under the common automatic-input protocol, CLASVS improves every Vevo2 operation and attains the lowest macro-PER, especially on deletion and insertion. It preserves melody and singer identity while improving naturalness and lyric intelligibility over the continuous-NAR comparator, though parallel generation is faster and stronger on substitution. Interventions show semantic feedback guides progress while the previous latent patch preserves local continuity. These results provide a continuous-AR foundation for broader controllable singing performance.

References

- Bai, B.; Geng, Y.; Wang, F.; Wang, C.; Guo, P.; Gao, Y.; and Li, Y. 2026. HQ-SVC: Towards High-Quality Zero-Shot Singing Voice Conversion in Low-Resource Scenarios. In *Proceedings of the 40th Conference on Artificial Intelligence*, volume 40(36), 30013–30021.
- Camp, J.; Kenter, T.; Finkelstein, L.; and Clark, R. 2023. MOS vs. AB: Evaluating Text-to-Speech Systems Reliably Using Clustered Standard Errors. In *Proceedings of Interspeech 2023*, 1090–1094.
- Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; et al. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.
- Chunhui, W.; Zeng, C.; and He, X. 2023. XiaoiceSing 2: A High-Fidelity Singing Voice Synthesizer Based on Generative Adversarial Network. In *Proceedings of Interspeech 2023*, 5401–5405.
- Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2023. High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research*.
- Geng, Y.; Li, Y.; Yang, J.; Jiang, T.; An, B.; Li, Y.; and Shen, X. 2026. Bridging the Stability-Expressivity Gap: Synthetic Data Scaling and Preference Alignment for Low-Resource Spoken Language Models. *arXiv preprint arXiv:2605.27383*.
- Hao, C.; Zheng, J.; Ma, G.; Jiang, Y.; Chen, H.; et al. 2026. YingMusic-Singer: Controllable Singing Voice Synthesis with Flexible Lyric Manipulation and Annotation-free Melody Guidance. Forthcoming in Interspeech 2026, arXiv:2603.24589.
- He, H.; et al. 2024. Emilia: An Extensive, Multilingual, and Diverse Speech Dataset for Large-Scale Speech Generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 885–890. IEEE.
- Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598.
- Jia, D.; et al. 2025. DiTAR: Diffusion Transformer Autoregressive Modeling for Speech Generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 27255–27270. PMLR.
- Kim, T.; Cho, C.; and Lee, Y. H. 2024. Period Singer: Integrating Periodic and Aperiodic Variational Autoencoders for Natural-Sounding End-to-End Singing Voice Synthesis. In *Proceedings of Interspeech 2024*, 1875–1879.
- Kirkland, A.; Mehta, S.; Lameris, H.; Henter, G. E.; Székely, E.; and Gustafson, J. 2023. Stuck in the MOS Pit: A Critical Analysis of MOS Test Methodology in TTS Evaluation. In *Proceedings of the 12th ISCA Speech Synthesis Workshop*, 41–47.
- Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2023. High-Fidelity Audio Compression with Improved RVQGAN. In *Advances in Neural Information Processing Systems*, volume 36.
- Le, M.; Vyas, A.; Shi, B.; Karrer, B.; Sari, L.; et al. 2023. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. In *Advances in Neural Information Processing Systems*, volume 36.
- Li, R.; Huang, R.; Wang, Y.; Hong, Z.; and Zhao, Z. 2024. Self-Supervised Singing Voice Pre-Training towards Speech-to-Singing Conversion. In *Findings of the Association for Computational Linguistics: ACL 2024*, 9819–9831. Association for Computational Linguistics.
- Lian, S.; Li, C.; Li, B.; Wang, H.; Zheng, D.; et al. 2026. dots.tts Technical Report. arXiv:2606.07080.
- Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*.
- Liu, J.; Li, C.; Ren, Y.; Chen, F.; and Zhao, Z. 2022. DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism. In *Proceedings of the 36th Conference on Artificial Intelligence*, volume 36, 11020–11028.
- Mehta, S.; Tu, R.; Beskow, J.; Székely, É.; and Henter, G. E. 2024. Matcha-TTS: A Fast TTS Architecture with Conditional Flow Matching. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 11341–11345.
- Meng, L.; Zhou, L.; Liu, S.; Chen, S.; Han, B.; et al. 2025. Autoregressive Speech Synthesis without Vector Quantization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1287–1300. Vienna, Austria: Association for Computational Linguistics.
- Park, Y.; Im, J.; and Nam, J. 2026. MeloDISinger: Melody-Aware & Duration-Preserving Singing Voice Editing with Audio Infilling. Forthcoming in Interspeech 2026, arXiv:2606.30580.
- Peebles, W.; and Xie, S. 2023. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.
- Peng, Z.; Yu, J.; Wang, W.; Chang, Y.; Sun, Y.; et al. 2025. VibeVoice Technical Report. arXiv:2508.19205.
- Qian, K.; Zhang, Y.; Gao, H.; Ni, J.; Lai, C.-I.; et al. 2022. ContentVec: An Improved Self-Supervised Speech Representation by Disentangling Speakers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 18003–18017. PMLR.
- Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; Mcleavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 28492–28518. PMLR.
- Roman, R. S.; Fernandez, P.; Elsahar, H.; Défossez, A.; Furon, T.; and Tran, T. 2024. Proactive Detection of Voice Cloning with Localized Watermarking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 43180–43196. PMLR.

- Shen, K.; Ju, Z.; Tan, X.; Liu, E.; Leng, Y.; et al. 2024. NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers. In *International Conference on Learning Representations*.
- Wang, H.; Lu, H.; Deng, J.; Xu, H.; Chen, Y.; et al. 2026. SemaVoice: Semantic-Aware Continuous Autoregressive Speech Synthesis. arXiv:2605.16964.
- Wang, H.; Zheng, S.; Chen, Y.; Cheng, L.; and Chen, Q. 2023. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. In *Proceedings of Interspeech 2023*, 5301–5305.
- Wang, Y.; Hu, R.; Huang, R.; Hong, Z.; Li, R.; et al. 2024. Prompt-Singer: Controllable Singing-Voice-Synthesis with Natural Language Prompt. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4780–4794. Association for Computational Linguistics.
- Wang, Y.; Wang, X.; Zhu, P.; Wu, J.; Li, H.; Xue, H.; Zhang, Y.; Xie, L.; and Bi, M. 2022. OpenCpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis. In *Proceedings of Interspeech 2022*, 4242–4246.
- Wei, H.; Cao, X.; Dan, T.; and Chen, Y. 2023. RMVPE: A Robust Model for Vocal Pitch Estimation in Polyphonic Music. In *Proceedings of Interspeech 2023*, 5421–5425.
- Wu, Y.; Zhang, C.; Shi, J.; Tang, Y.; Yang, S.; and Jin, Q. 2024. TokSing: Singing Voice Synthesis Based on Discrete Tokens. In *Proceedings of Interspeech 2024*, 2549–2553.
- Yan, C.; Jin, C.; Huang, D.; Yu, H.; Peng, H.; et al. 2025. Ming-UniAudio: Speech LLM for Joint Understanding, Generation and Editing with Unified Representation. arXiv:2511.05516.
- Yang, A.; Li, A.; Yang, B.; et al. 2025a. Qwen3 Technical Report. arXiv:2505.09388.
- Yang, S.-W.; Kim, B.; Huang, K.-P.; Tang, Q.; Phan, H.; et al. 2025b. Generative Audio Language Modeling with Continuous-Valued Tokens and Masked Next-Token Prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 70793–70812. PMLR.
- Zang, Y.; Shi, J.; Zhang, Y.; Yamamoto, R.; Han, J.; et al. 2024. CtrSVDD: A Benchmark Dataset and Baseline Analysis for Controlled Singing Voice Deepfake Detection. In *Proceedings of Interspeech 2024*, 4783–4787.
- Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2022. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507.
- Zhang, L.; Li, R.; Wang, S.; Deng, L.; Liu, J.; Ren, Y.; He, J.; Huang, R.; Zhu, J.; Chen, X.; and Zhao, Z. 2022a. M4Singer: A Multi-Style, Multi-Singer and Musical Score Provided Mandarin Singing Corpus. In *Advances in Neural Information Processing Systems*, volume 35.
- Zhang, L.; Zhao, Z.; Ren, Y.; and Deng, L. 2022b. EditSinger: Zero-Shot Text-Based Singing Voice Editing System with Diverse Prosody Modeling. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 4503–4509.
- Zhang, X.; Zhang, J.; Wang, Y.; Wang, C.; Chen, Y.; et al. 2026. Vevo2: A Unified and Controllable Framework for Speech and Singing Voice Generation. *IEEE Transactions on Audio, Speech and Language Processing*, 34: 2262–2275.
- Zhang, Y.; Pan, C.; Guo, W.; Li, R.; Zhu, Z.; et al. 2024a. GTSinger: A Global Multi-Technique Singing Corpus with Realistic Music Scores for All Singing Tasks. In *Advances in Neural Information Processing Systems*, volume 37.
- Zhang, Y.; Xue, H.; Li, H.; Xie, L.; Guo, T.; et al. 2023. VISinger2: High-Fidelity End-to-End Singing Voice Synthesis Enhanced by Digital Signal Processing Synthesizer. In *Proceedings of Interspeech 2023*, 4444–4448.
- Zhang, Y.; et al. 2024b. 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*, 1960–1975. Miami, Florida, USA: Association for Computational Linguistics.
- Zhou, Y.; Zeng, G.; Liu, X.; Li, X.; Yu, R.; Wang, Z.; Ye, R.; Sun, W.; Gui, J.; Li, K.; Wu, Z.; and Liu, Z. 2025. VoxCPM: Tokenizer-Free TTS for Context-Aware Speech Generation and True-to-Life Voice Cloning. arXiv:2509.24650.