CLASVS: CONTINUOUS-LATENT AUTOREGRESSION FOR LENGTH-CHANGING LYRIC EDITS FROM A REFERENCE SINGING PERFORMANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-conditioned lyric editing changes words within a sung phrase while preserving its melody, singer identity, and naturalness. When edits change the syllable count, target lyrics must override source-correlated cues while retaining the reference phrasing. We introduce CLASVS, a continuous-latent autoregressive editor with learned stopping. Its State–Control–Transition (SCT) history-access contract keeps target lyrics and reference melody available throughout generation, returns causal semantic feedback from generated latent patches to the planner, and confines direct predecessor acoustics to local synthesis. Progressive State–Control Grounding (PSCG) learns these routes through semantic pretraining, control perturbations, and staged reconstruction without paired lyric-edit recordings or score annotations. Across two Mandarin benchmarks, CLASVS improves deletion and insertion over continuous-NAR YingMusic+ and all four edit operations over discrete-AR Vevo2. Blinded paired listening favors its naturalness and target-lyric intelligibility, complementing objective measurements of melody and singer preservation. Matched routing and training comparisons support the access design and grounding recipe. Together, these results connect reliable lyric-count editing with the quality of the finished singing, using performances and transcripts for training. Audio demos are available at `https://anonymous.4open.science/w/CLASVS-65B7/index.html`.

1 INTRODUCTION

Reference-conditioned lyric editing changes words while preserving melody, singer identity, and naturalness; timing may adapt to the revised text (Zhang et al., 2022; Park et al., 2026). Here, length-changing edits alter the target syllable count within a reference phrase. Score-conditioned synthesis requires costly phoneme, note, and duration alignment (Liu et al., 2022; Zhang et al., 2023; Wang et al., 2024). Reference-audio/target-lyric editors avoid supplied scores (Zhang et al., 2026a; Hao et al., 2026), but the reference carries both performance cues and source words.

This conflict is clearest for insertion and deletion (Zhang et al., 2022; Hao et al., 2026). Insertion must add syllables without losing the phrase; deletion must suppress old syllables without stopping early. Paired recordings differing only in lyrics are difficult to obtain because resinging changes phrasing and acoustics. We therefore train by reconstructing performances from their transcripts. At inference, target lyrics must override source-correlated cues, a leakage risk found in melody-guided editing (Hao et al., 2026). Autoregressive (AR) speech models can accumulate errors (Zhang et al., 2026b); a source-following patch may similarly bias later patches.

These failures make lyric progress and stopping central. YingMusic-Singer-Plus (YingMusic+), a continuous nonautoregressive (NAR) editor, offers parallel generation (Hao et al., 2026). Continuous-latent AR updates decisions from generated patches (Jia et al., 2025) and avoids quantized codec tokens (Meng et al., 2025); here it also learns when to stop. Yet detailed predecessor acoustics may reinforce source-lyric errors if fed directly to the planner. Which history should support global lyric planning, and which only local synthesis?

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.

Table 1: Generated-history access in related autoregressive audio models (Jia et al., 2025; Peng et al., 2025; Zhou et al., 2025; Rouard et al., 2025).

Model	Global recurrent input	Local path and role
DiTAR	Encoded acoustic patches	Prior patches → local diffusion (patch AR)
VibeVoice	Semantic/acoustic projections	Hybrid long-form speech feedback; no predecessor path
VoxCPM	Local-encoder embeddings	FSQ planning + residual acoustics; prior patch → diffusion
CALM†	Noised acoustic history	Clean recent latents → short context; noisy long-range history
CLASVS	Frozen semantic feedback	Previous latent patch → Flow-DiT; full acoustics remain local

Note. Access paths are compared across tasks, not performance. †Music configuration; its speech configuration differs. "Global" refers only to generated-history feedback.

CLASVS formalizes this choice as a State–Control–Transition (SCT) history-access contract. Target-lyric and reference-melody tokens remain persistent *Control* in the planner. A frozen causal semantic encoder reads generated AudioVAE latent patches and returns lyric-progress feedback as recurrent *State*. The full previous patch directly conditions only the next local Flow-DiT *Transition*; the planner receives semantic feedback instead of full predecessor acoustics. A reference singer vector conditions Flow-DiT separately, and a planner head learns when to stop.

That separation must still be learned without paired lyric-edit recordings. Progressive State–Control Grounding (PSCG) pretrains and freezes the semantic encoder in State Grounding; Control Grounding masks melody spans and sometimes drops the local predecessor throughout a three-stage Task Grounding curriculum, retaining lyrics. These perturbations encourage explicit controls over predecessor copying, while reconstruction requires no score annotations.

We evaluate lyric accuracy, performance preservation, and blinded paired listening. Across two Mandarin four-operation benchmarks, CLASVS improves deletion and insertion over continuous-NAR YingMusic+ and all operations over discrete-AR Vevo2. Listeners favor its naturalness and target-lyric intelligibility. Matched routing and equal-budget training controls connect SCT and PSCG to these editing outcomes, with State pretraining included in the budget. We make the following three contributions:

- We **formulate the SCT history-access contract**: persistent target and reference controls, recurrent semantic feedback, and a local previous latent patch.
- We **develop PSCG training** using semantic pretraining, control corruptions, and staged reconstruction without counterfactual edit pairs or score annotations.
- We **provide operation-resolved system evaluation** across two four-operation benchmarks, paired listening, matched routing, and equal-budget training controls.

2 Related Work

Comparable editing interfaces. EditSinger and MeloDISinger need alignment, duration, or edit-boundary cues; SoulX-Singer supplies a metadata pipeline (Zhang et al., 2022; Park et al., 2026; Qian et al., 2026). Vevo2 and YingMusic+ support reference-audio/target-lyric editing without supplied scores or boundaries (Zhang et al., 2026a; Hao et al., 2026). They form the four-operation system comparisons; Appendix Table 11 records the screening protocol.

Representation and generation order. Discrete codecs tokenize acoustic targets (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023; Wu et al., 2024), whereas continuous latents avoid a finite acoustic codebook (Le et al., 2023; Shen et al., 2024; Mehta et al., 2024; Meng et al., 2025). NAR offers parallelism; AR permits feedback and learned stopping. Neither choice specifies which generated history enters the planner.

Generated-history routing. Continuous-AR systems differ in acoustic and semantic feedback (Jia et al., 2025; Peng et al., 2025; Zhou et al., 2025; Rouard et al., 2025); DiTAR+ masks shallow acoustics (Zhang et al., 2026b), while instruction editors use edit supervision (Yan et al., 2025). SCT addresses reconstruction-trained lyric editing by sending semantic feedback globally and direct full-patch access locally. Table 1 compares routes, not quality.

3 METHOD

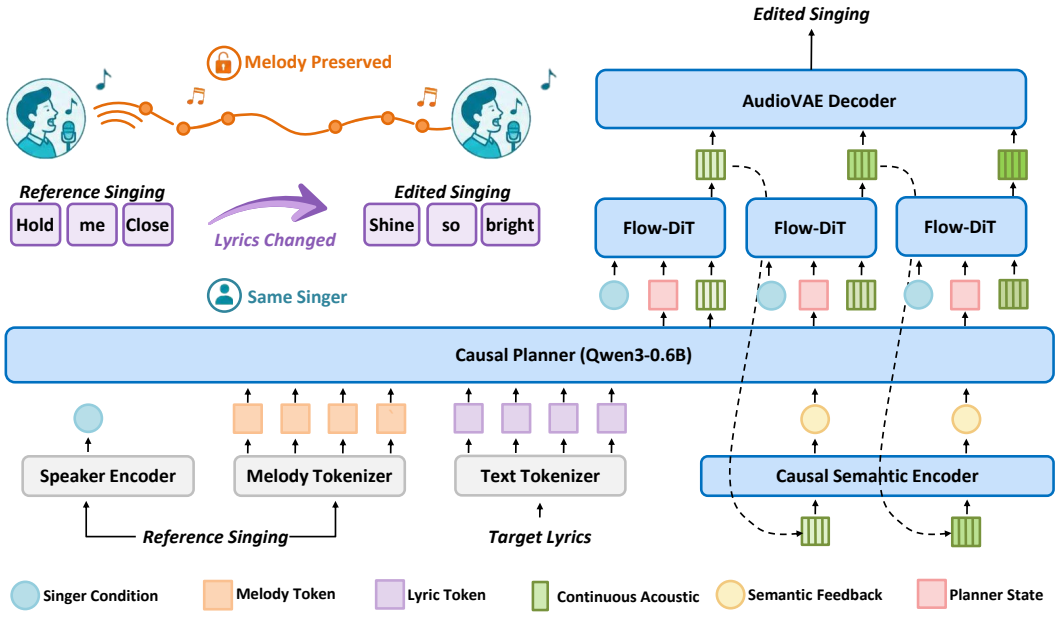

Figure 1: SCT history-access contract. Persistent lyric and melody controls remain in the planner, which receives semantic feedback. The previous latent patch and singer condition directly condition the shared Flow-DiT, which generates five 64-D latent frames per 100-ms step.

3.1 TASK AND TEMPORAL REPRESENTATION

Inputs and latent timeline. Let $\mathbf{r}$ be reference audio and $\mathbf{x}$ target lyric tokens. Frozen encoders map $\mathbf{r}$ to 6.25-Hz melody tokens $\mathbf{m}$ from chroma-derived inputs and a 192-D CAM++ singer vector $\mathbf{e}$ (Wang et al., 2023). A frozen Ming-UniAudio AudioVAE (Yan et al., 2025) represents 16-kHz audio as $\mathbf{z}^* \in \mathbb{R}^{T\times 64}$ at 50 Hz. Five frames form each patch $\bar{\mathbf{z}}_k^* \in \mathbb{R}^{5\times 64}$, giving 10-Hz recurrence and at most 100-ms endpoint granularity; a final short patch is zero-padded and retained.

Reconstruction versus editing. Every training tuple $(\mathbf{r}, \mathbf{x}, \mathbf{z}^*)$ is one performance and its transcript; evaluation changes $\mathbf{x}$ while reference controls remain tied to the source. The full melody prefix stays cached without phone alignment, semantic feedback marks generated progress, and a learned stop head selects patch count. The cap $\min(K_{\text{ref}} + 8, 256)$ bounds duration adaptation around the reference phrase. SCT addresses this reconstruction–editing gap.

3.2 SCT HISTORY-ACCESS CONTRACT

Persistent control and recurrent state. Target lyrics specify what must change, reference melody and identity specify what must be preserved, and generated history records what has been sung. The contract keeps controls persistent, returns semantic feedback globally, and confines direct predecessor acoustics to the local transition. Recurrent context includes the planner cache, semantic-encoder cache, and previous patch; State supplies the planner with a causal summary of realized singing.

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

$$\begin{aligned} \mathbf{R}_k &= (\mathbf{K}_k^{\text{plan}}, \mathbf{K}_k^{\text{sem}}, \bar{\mathbf{z}}_{k-1}), \\ p_\theta(\bar{\mathbf{z}}_{1:K} \mid \mathbf{x}, \mathbf{m}, \mathbf{e}) &= \prod_{k=1}^{K} p_\theta(\bar{\mathbf{z}}_k \mid \mathbf{R}_k, \mathbf{x}, \mathbf{m}, \mathbf{e}). \end{aligned} \tag{1}$$

The update $\mathcal{U}$ encodes the new patch, pools five semantic states into planner feedback, and exposes the same patch as the local predecessor at $k+1$; both recurrent paths start from 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:

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

The cache retains full controls and earlier feedback; a head predicts continue or stop from $\mathbf{h}_k$. Singer identity bypasses the planner and modulates Flow-DiT through adaptive layer normalization. The planner never observes future patches, and no note, duration, or edit-boundary labels are required.

Global semantic feedback and local transition. The semantic route uses a custom, randomly initialized Whisper-style encoder (Radford et al., 2023), not pretrained Whisper weights. It has 32 layers, width 1280, 20 attention heads, and FFN width 5120. A linear $64 \to 1280$ adapter accepts AudioVAE frames without downsampling; sinusoidal positions, a frame-causal mask, and incremental KV caching preserve causal access. After State Grounding, the adapter and encoder are frozen; a trainable projection maps states to 1024 dimensions before five-frame pooling:

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

In parallel, the complete 5×64 predecessor enters an 8-block conditional Flow-DiT (Lipman et al., 2023; Peebles & Xie, 2023) with one flow-time/planner token, five predecessor tokens, and five noisy current-patch tokens. Semantic feedback persists globally, whereas the full latent patch conditions only the next transition. Incremental caches avoid prefix re-encoding, although planner memory remains linear in rollout length.

Unified and access-matched routes. To test this access pattern, we first define an original unified-context route (U-old):

$$\mathbf{u}_{k-1} = \mathbf{a}_{k-1} + W \,\text{vec}(\bar{\mathbf{z}}_{k-1}) + \mathbf{b}, \qquad W \in \mathbb{R}^{1024 \times 320}. \tag{4}$$

U-old retains one planner token and all 11 Flow-DiT tokens but leaves its global acoustic path unmasked during local dropout. U-sync-add synchronizes the mask; U-sync-gate gates fusion. A gated pair then changes only planner access (fixed null versus real predecessor), holding projection, normalization, gate, tokens, parameters, local path, dropout, and 576 GPU-hours per seed fixed (Appendix Table 10).

3.3 PSCG Training

PSCG training grounds semantic feedback, lyric/melody controls, and singing generation (Figure 2). State Grounding pretrains feedback; Control Grounding spans the three-stage Task Grounding curriculum used for singing generation.

State grounding. The random semantic encoder is trained for 100K updates on 80%/20% broad singing/speech using character CTC and multi-scale log-mel reconstruction:

$$\mathcal{L}_{\text{state}} = \mathcal{L}_{\text{CTC}} + 5\mathcal{L}_{\text{MSMel}}. \tag{5}$$

CTC supplies transcript supervision without frame alignment; Mel reconstruction retains an acoustic signal. The heads are discarded and the encoder frozen for stable downstream feedback. The resulting representation exposes lyric progress to the planner while retaining the acoustic cues characterized by the probes below.

Control grounding. Training zeros the local teacher-forced predecessor with $p_h = .10$ and masks melody spans with $p_m = .15$ while retaining lyrics. These corruptions force use of remaining controls without supplying a counterfactual waveform. History dropout is local; semantic history stays intact and still reaches the planner during training.

Task grounding. Downstream training uses 72K broad-singing updates, 24K singing/speech updates (1:1), and 24K curated-singing updates (Geng et al., 2026). Shuffled and reordered controls use the same exposure multiset to isolate the effect of stage order.

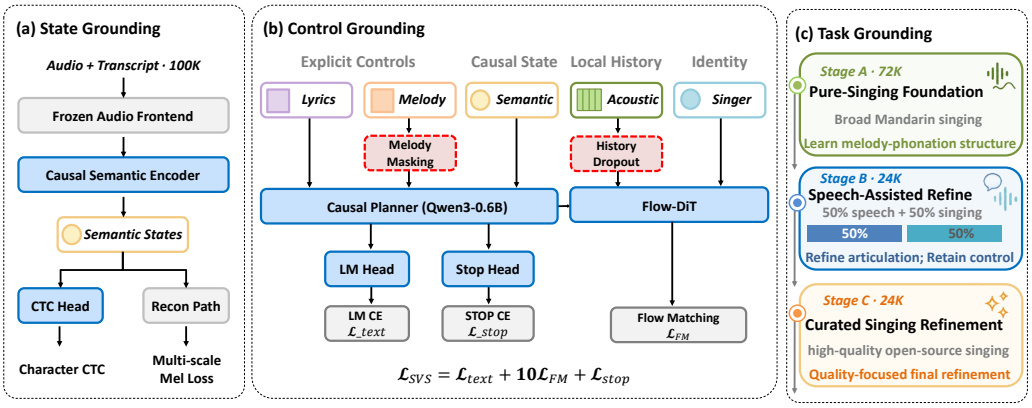

Figure 2: Progressive State–Control Grounding (PSCG). State Grounding pretrains the causal encoder with character CTC and Mel reconstruction; Control Grounding zeros only the local previous-latent-patch input and masks melody while retaining lyrics; Task Grounding schedules broad singing, mixed singing/speech, and curated singing. K denotes thousands of updates.

Training objective and setup. Both feedback paths train on teacher-forced latent histories with the local-history dropout defined above; caches reset between samples and gradients span each sample. We train the feedback projection, planner, Flow-DiT, and heads while keeping the semantic encoder frozen. Singer conditioning enters block-specific AdaLN mappings, with a learned null embedding for CFG.

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

$$\mathcal{L} = \mathcal{L}_{\text{text}} + 10\mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{stop}}. \tag{6}$$

$\mathcal{L}_{\text{text}}$ is lyric-prefix cross-entropy for the shared planner; $\mathcal{L}_{\text{FM}}$ supervises acoustic realization. Balanced $\mathcal{L}_{\text{stop}}$ labels the final patch as stop. Joint lyric–melody–singer dropout ($p = .10$) enables CFG (Ho & Salimans, 2022); frontends stay frozen.

Training uses about 8,000 hours of permission-documented Mandarin singing, 2,000 hours of Emilia speech (He et al., 2024), and 300 hours of curated open singing. Three 120K-update seeds use BF16, ZeRO-2, AdamW, global batch 128, and eight H200s (median 71.8 hours, excluding State pretraining). A frozen 1.350B AudioVAE accompanies .765B trainable parameters.

3.4 Autoregressive Inference

Rollout and stopping. After lyric/melody prefill, Flow-DiT combines the planner state, singer condition, and predecessor patch to generate $\bar{\mathbf{z}}_k$ with 24 Euler steps and guidance 2.0. Each patch is retained before testing stop $\geq .5$ or the cap, then feeds both history paths. The concatenated sequence is decoded once.

Semantic feedback tracks realized lyrics while reference melody describes timing; phones need not align with melody tokens. Evaluation therefore checks target following and duration separately, since matching duration alone can hide omissions or source reversion.

Testable consequences of history access. The two history routes imply different failure modes. Without semantic feedback, the planner retains target lyrics and reference melody but loses an explicit record of realized content; insertions and deletions should then be more vulnerable to omissions, repetitions, or premature stopping. Without the local predecessor, Flow-DiT retains the plan but loses the nearest acoustic boundary, potentially reducing pitch and patch continuity. Direct planner access to that predecessor may instead reinforce source-lyric cues. We therefore compare matched planner-access variants, remove each route separately, and perturb generated-history inputs during rollout. The tests measure target following and source reversion alongside pitch continuity and stopping, rather than treating duration agreement as evidence of correct lyrics.

4 EXPERIMENTS

4.1 PROTOCOL

Baseline comparability. At the September 20, 2026 freeze, headline baselines needed an official runnable checkpoint, four-operation support, and one output from the official pipeline. They received the same unaligned inputs: reference singing, target lyrics, and a source transcript if needed, without supplied score, duration, alignment, or edit boundaries. Internal preprocessing was allowed. Only Vevo2 (discrete AR) and YingMusic+ (continuous NAR) qualified (Appendix Table 11).

Benchmarks. CLA-LyricEdit-320 contains 320 edits from 160 excerpts, 80 songs, and 40 singers, balanced across partial/full substitution (PSub/FSub), deletion, and insertion. Splits are song-disjoint; two annotators verify targets and operations ($\kappa = .93$). We also use the published Mandarin four-operation LyricEditBench portion: 600 items per operation, 2,400 total (Hao et al., 2026). Macro-PER averages within each operation before weighting the four equally, retaining deletion and insertion as distinct requirements in the aggregate score and subsequent analysis.

Metrics and listening. FireRedASR-AED-L and pypinyin give phoneme error rate (PER) and NED; SrcRev counts source-closer outputs. Preservation uses RMVPE log-F0 correlation (FPC) (Wei et al., 2023), voiced-onset error, reference-duration deviation, and WavLM speaker cosine (Chen et al., 2022). Listening pairs 80 edits (20 per operation) across three systems (240 audios); 20 Mandarin listeners give eight ratings per audio (1,920 ratings; 5,760 N-/M-/L-MOS scores). Objective intervals use 10,000 song-clustered draws (Appendix D); paired MOS analyses account for repeated listeners and shared content, with Holm correction (Kirkland et al., 2023; Camp et al., 2023) (Appendix F).

Table 2: Editing accuracy, performance preservation, and perceptual quality.

A. Editing accuracy	PER ↓				
System	PSub	FSub	Del	Ins	Macro [95% CI]
CLA-LyricEdit-320: 320 edits, 80 songs					
Vevo2	.0617	.0563	.1033	.0581	.0699 [.0617, .0781]
YingMusic+	**.0171**	**.0245**	.0847	.0382	.0411 [.0356, .0469]
CLASVS	.0505	.0442	**.0260**	**.0298**	**.0376** [.0326, .0429]
Public LyricEditBench: four Mandarin operations					
Vevo2	.0664	.0607	.1123	.0649	.0761 [.0672, .0850]
YingMusic+	**.0220**	**.0230**	.0930	.0440	.0455 [.0384, .0506]
CLASVS	.0508	.0435	**.0398**	**.0357**	**.0425** [.0368, .0485]

B. Performance preservation	primary benchmark			
System	FPC ↑	SIM ↑	RefDur (%) ↓	Onset (ms) ↓
Vevo2	.7440	.892	1.5995	69.58
YingMusic+	.9340	.906	**.1465**	18.80
CLASVS	**.9410**	**.914**	.9906	**7.68**

C. Perceptual quality	80 paired edits, 240 audios		
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]

Notes. Bold marks the best point estimate. Brackets in A/C give 95% macro-PER / individual-system MOS intervals; paired contrasts are reported separately in Figure 3 and Appendix Table 14. Primary PER intervals use song-clustered resampling. Listening uses 20 listeners, shared edits, and eight ratings per audio; N-/M-/L-MOS assess naturalness, melody, and target-lyric intelligibility.

4.2 EDITING ACCURACY AND PERFORMANCE PRESERVATION

Successful lyric editing must follow revised text while retaining the performance carried by the reference. We evaluate both demands on the primary benchmark and the public split, separating substitutions from length-changing edits. Operation-level results show where the editor advances through changed lyric lengths and where it makes replacements within the source timing.

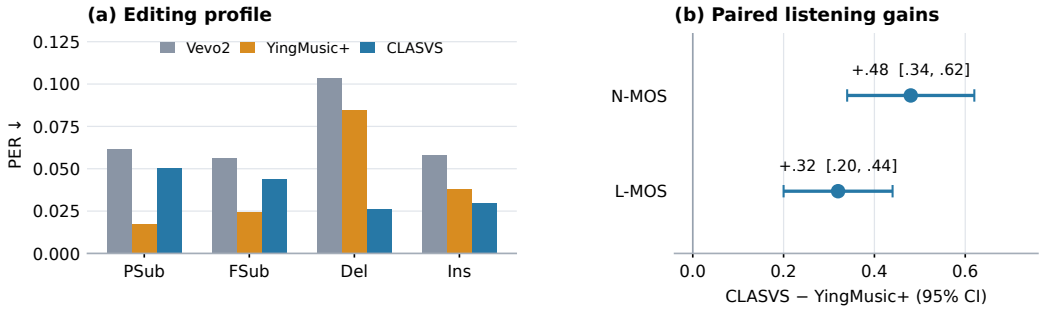

Figure 3: Operation-resolved system evaluation. Left: PER by operation on the primary benchmark. Right: paired CLASVS–YingMusic+ listener contrasts for naturalness, clarity, and artifacts (N-MOS) and target-lyric intelligibility (L-MOS), with 95% intervals. Lower PER and positive MOS differences favor CLASVS; operation bars are point estimates.

Heard quality on matched edits. In system-blinded listening on 80 paired edits, CLASVS scores 4.13 versus YingMusic+'s 3.65 N-MOS for naturalness, clarity, and artifacts: $+.48$ $[+.34, +.62]$ (Holm $p < .001$). Target-lyric L-MOS is 4.42 versus 4.10: $+.32$ $[+.20, +.44]$ ($p < .001$). Listeners see the target lyrics and judge their intelligibility and correctness, complementing automatic PER with a direct judgment of the finished singing. Figure 3 shows listener contrasts beside operation-level phone errors. Audio demos are available for direct listening to representative edits.

Editing accuracy and preservation. On deletion and insertion, CLASVS lowers primary mean PER from YingMusic+'s .0615 to .0279; both operations also improve on public LyricEditBench. Across four equally weighted operations, primary macro-PER is .0376 versus .0411, a paired difference of $-.0035$ [95% CI $-.0056, -.0013$]. Public-set macro scores are .0425 and .0455. Against Vevo2, CLASVS improves all operations, reducing primary macro-PER from .0699 to .0376. Preservation is measured at complementary temporal scales: FPC is .9410 versus YingMusic+'s .9340, and voiced-onset error is 7.68 versus 18.80 ms. On PSub, unedited-region log-spectral distance is .097 versus .112, SIM is .923 versus .912, and edit-boundary onset error is 8.0 versus 20.2 ms. YingMusic+ has closer global endpoint agreement (.1465% versus .9906% RefDur), while these local measures favor CLASVS. Paired M-MOS is 4.28 versus 4.27 ($p = .79$). Table 2 and Appendix D give the full profile and statistical definitions.

Operation-dependent gains. Primary Del/Ins mean PER falls from YingMusic+'s .0615 to .0279 (54.6%), whereas substitution mean is .0474 versus .0208. This profile identifies complementary strengths across operations. Four-way macro summarizes balanced editing capabilities; paired listening evaluates finished singing. Target margins and stopping audits assess target-lyric realization.

Inference cost. At the reported quality settings, single-H200 batch-one warm end-to-end RTF is 1.45, with 5.8 GiB incremental and 9 GiB resident peak memory (Appendix G).

4.3 Target Following and History Routing

Target-following audit. NED divides phone edit distance by the larger sequence length, and target-preference margin subtracts output–target NED from output–source NED. This directional comparison complements target-only PER by measuring whether the output is closer to revised or source lyrics. Every operation has a positive margin (.2248–.7209; overall .3685), supporting target following under source-correlated reference evidence. In the no-edit control, source/target NEDs are .018/.018, PER is .019, and margin and strict SrcRev are zero by definition (Appendix E).

Routing ablations. In the gated matched pair, changing only planner access to the predecessor lowers PER from .0408 to .0390 (4.4% relative), with all three seeds favoring local-only access and a paired difference of +.0018 [95% CI +.0004,+.0032]. Deletion and insertion account for the clearest effects: local-only access lowers their PER by 8.4% and 9.1% relative to global access.

Table 3: Controlled routing and training comparisons under the locked continuous-AR protocol.

A. History access shared downstream recipe

Configuration	Change from Full	PER ↓	FPC ↑	SrcRev ↓
Full SCT + PSCG	Reference configuration	**.0376**	**.9410**	**.081**
Unified (U-old)	Planner acoustics; local mask	.0450	.9300	.120
U-sync-add	Shared acoustic dropout mask	.0418	.9350	.103
Matched local-only	Gated; fixed-null planner acoustics	.0390	.9395	.088
Matched global-access	Gated; real planner acoustics	.0408	.9370	.096
No semantic feedback	Remove semantic route	.0564	.9205	.154
No local acoustic patch	Remove local predecessor	.0492	.8896	.086

B. Training budget including State pretraining

Configuration	Updates	Trainable	Total budget	PER ↓
Random-joint	120K	1.395B	1.20	.0511
Random-joint	130K	1.395B	1.30	.0477
Random-joint	140K	1.395B	1.40	.0444
Full SCT + PSCG	120K	.765B	1.40	**.0376**

C. Training controls 120K downstream updates

Configuration	Change from Full	PER ↓	SrcRev ↓
No history dropout	Set local $p_h = 0$ only	.0440	.130
No Control Grounding	Remove both corruptions	.0468	.121
No Task Grounding	Shuffle the same exposures	.0439	.102

Notes. Panel A reports macro-PER. U-old adds 328,704 parameters; synchronized masking and gating are detailed in Appendix Table 9. The matched pair changes only planner-cache acoustic input, with all listed components and 576 GPU-hours per seed held equal (Table 10). Panel B normalizes GPU-hours to Full's downstream run, not FLOPs. No-history-dropout averages seeds 0/1/2.

FPC/SrcRev point estimates also favor local-only (.9395/.088 versus .9370/.096). Separately, U-old's .0450 falls to .0418 with synchronized dropout and .0408 with gating; removing semantic feedback or the local acoustic patch yields .0564/.0492 PER (Appendix Tables 9–10).

Probe profile. Song-disjoint probes describe the deployed interfaces: semantic/acoustic reached-position MAE is .082/.191 (time baseline .120), singer accuracy 8.6/91.2% (chance 2.5%), and log-F0 R^2 .09/.84 (Figure 4). Time plus semantics reaches .080 and a semantic MLP .078; Appendix E specifies the features, labels, and controls.

Intervention response. Let $M_{\text{lyric}} = \text{NED}(y, x_{\text{src}}) - \text{NED}(y, x_{\text{tgt}})$ and $M_{\text{melody}} = \text{FPC}(y, r) - \text{FPC}(y, r_{\text{donor}})$. Semantic-history shuffling reduces lyric margin .3685→.2469 (aggregate retention .67); predecessor shuffling reduces melody margin .200→.170 (.85). Only generated history is perturbed and caches are rebuilt. These rollout interventions show that semantic and acoustic histories each contribute to the corresponding measured outcome.

4.4 Grounding and Training Budget

GPU-hour and training controls. Routing tests ask where generated history enters; the training comparison includes the extra State Grounding phase. Panel B separates update count from logged GPU-hours normalized to Full's downstream run. Random-joint trains 1.395B parameters; Full trains .765B plus .40 pretraining units. At the same reported 1.40-unit budget, PSCG reaches .0376 PER against .0444 for 140K-step random-joint, a 15.3% relative reduction. Panel C shows higher PER when local history dropout is removed, both corruptions are removed, or training exposures are shuffled; melody-mask removal yields .0419, and task-stage reordering .0415 (Appendix A).

PSCG training and edit generalization. Reconstruction pairs a performance with its original transcript, while evaluation pairs that performance with revised lyrics. State Grounding yields linearly accessible progress, and the reported corruption and stage-order controls show lower PER for PSCG training. Positive target margins, no-edit calibration, and routing ablations jointly show target following across the four edit types.

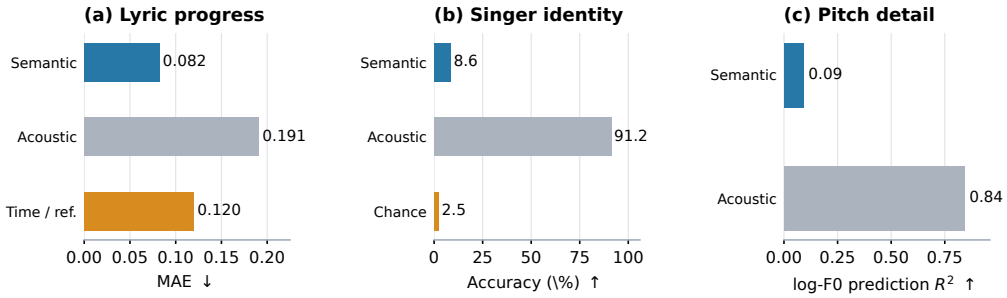

Figure 4: Diagnostic information profiles for the two history routes. Linear probes compare semantic feedback with AudioVAE latents for target-position progress, singer identity, and frame-level log-F0. Orange bars mark the normalized-time and chance baselines.

4.5 Stopping and Transfer Robustness

Stopping behavior. Budget-hit rates are 0/0/0/1.25% for PSub/FSub/Del/Ins; premature/delayed stops reach at most 1.25/2.50%. The current patch is retained before testing the learned stop rule, and forced cap stops are recorded separately (Appendix E). These measurements distinguish budget-limited termination from learned endpoint errors. No deletion reaches the cap. Target-phone errors and source-versus-target margins provide complementary measures of the realized lyric content.

Singer-held-out transfer. Across 24 post-freeze singers and 192 edits, CLASVS/YingMusic+ gives PER .0472 [.0416,.0533]/.0521 [.0461,.0586]. These singers were absent from training and development, and the audit was collected after checkpoint freeze. Deletion/insertion PER is .0380/.0418 versus .0960/.0574 for YingMusic+, extending the main operation pattern to new voices. FPC/SIM is .934/.905 versus .928/.898; Appendix D reports the complete operation and preservation profile.

5 Discussion and Limitations

Lyric-count accuracy and the quality of the finished singing give complementary evidence of editing utility. Deletion and insertion gains over YingMusic+ recur on the primary benchmark, public LyricEditBench, and post-freeze singers. Blinded listeners favor CLASVS in naturalness and target-lyric intelligibility. Figure 3 and the audio demos connect operation-level accuracy with directly perceptible differences in the completed renditions.

SCT makes generated-history access explicit. The matched gated pair isolates direct planner access to the previous acoustic patch and favors local-only access across three seeds, chiefly on deletion and insertion. Route removals measure the value of retaining each path; probes and perturbations characterize the deployed interfaces. Equal-budget training controls support PSCG within the continuous-AR stack, completing the evidence for the design.

The evaluation covers Mandarin edits within reference phrases. Substitution and insertion/deletion exhibit distinct accuracy profiles, while global endpoint agreement and local continuity measure different preservation properties. Additional languages and larger phrase-length changes are natural next settings for testing this editing approach.

6 Conclusion

CLASVS combines SCT and PSCG to learn lyric editing from singing performances and transcripts. Persistent controls guide the planner, causal semantic feedback summarizes generated history, and predecessor acoustics support local synthesis. This design learns phrase-level lyric-count edits without counterfactual edit pairs or score annotations. On two Mandarin benchmarks, it improves deletion and insertion over continuous-NAR YingMusic+ and all operations over discrete-AR Vevo2. Blinded listeners favor its naturalness and target-lyric intelligibility. Matched access comparisons, route removals, and equal-budget training controls support the design choices. The system unites target-lyric following with the quality of the finished singing.

AI USE STATEMENT

Generative AI tools assisted with format migration, manuscript and figure revision, language editing, literature comparison, consistency checks, simulated peer review, and suggestions for experimental and statistical clarification. Experimental results were supplied by the authors. AI tools did not train the reported models, collect listener ratings, or independently regenerate experimental scores. The authors retain responsibility for the manuscript, its experimental evidence, and its citations.

ETHICS STATEMENT

Study-specific evaluation recordings came from consenting singers, listeners consented to participate, and responses were de-identified. Recognizable-voice editing can nevertheless enable impersonation or unauthorized alteration (Zang et al., 2024). Source-level permissions, opt-out and removal procedures, identity-authorized checkpoint access, and localized provenance or watermarking should therefore govern data use, release, and responses to misuse (Roman et al., 2024).

REPRODUCIBILITY STATEMENT

The main text and appendix specify the SCT routes, model objectives, training schedule, inference settings, comparator interfaces, metrics, statistical procedures, listening design, and diagnostic experiments. The appendix also describes run identifiers, randomization, and item-level accounting. Upon acceptance, we will release the training and evaluation code, trained checkpoints, permission-cleared training and benchmark audio with manifests, per-item outputs and scores, and de-identified listening ratings.

REFERENCES

Joshua Camp, Tom Kenter, Lev Finkelstein, and Rob Clark. MOS vs. AB: Evaluating text-to-speech systems reliably using clustered standard errors. In *Proceedings of Interspeech 2023*, pp. 1090–1094, 2023. doi: 10.21437/Interspeech.2023-2014. URL https://www.isca-archive.org/interspeech_2023/camp23_interspeech.html.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://doi.org/10.1109/JSTSP.2022.3188113.

Wang Chunhui, Chang Zeng, and Xing He. XiaoiceSing 2: A high-fidelity singing voice synthesizer based on generative adversarial network. In *Proceedings of Interspeech 2023*, pp. 5401–5405, 2023. doi: 10.21437/Interspeech.2023-119. URL https://www.isca-archive.org/interspeech_2023/chunhui23_interspeech.html.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=ivCd8z8zR2.

Yizhong Geng, Yanliang Li, Jinghan Yang, Tianhan Jiang, Boxun An, Ya Li, and Xiaoyu Shen. Bridging the stability-expressivity gap: Synthetic data scaling and preference alignment for low-resource spoken language models. *arXiv preprint arXiv:2605.27383*, 2026.

Chunbo Hao, Junjie Zheng, Guobin Ma, Yuepeng Jiang, Huakang Chen, et al. YingMusic-Singer: Controllable singing voice synthesis with flexible lyric manipulation and annotation-free melody guidance, 2026. URL https://arxiv.org/abs/2603.24589. Forthcoming in Interspeech 2026.

Haorui He et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 885–890. IEEE, 2024. doi: 10.1109/SLT61566.2024.10832365. URL https://doi.org/10.1109/SLT61566.2024.10832365.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Dongya Jia et al. 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*, pp. 27255–27270. PMLR, 2025. URL https://proceedings.mlr.press/v267/jia25a.html.

Taewoo Kim, Choongsang Cho, and Young Han Lee. Period singer: Integrating periodic and aperiodic variational autoencoders for natural-sounding end-to-end singing voice synthesis. In *Proceedings of Interspeech 2024*, pp. 1875–1879, 2024. doi: 10.21437/Interspeech.2024-1720. URL https://www.isca-archive.org/interspeech_2024/kim24p_interspeech.html.

Ambika Kirkland, Shivam Mehta, Harm Lameris, Gustav Eje Henter, Eva Szekely, and Joakim Gustafson. Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. In *Proceedings of the 12th ISCA Speech Synthesis Workshop*, pp. 41–47, 2023. doi: 10.21437/SSW.2023-7. URL https://www.isca-archive.org/ssw_2023/kirkland23_ssw.html.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/58d0e78cf042af5876e12661087bea12-Abstract-Conference.html.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, et al. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2d8911db9ecedf866015091b28946e15-Abstract-Conference.html.

Shun Lei, Yixuan Zhou, Boshi Tang, Max W. Y. Lam, Feng Liu, Hangyu Liu, Jingcheng Wu, Shiyin Kang, Zhiyong Wu, and Helen Meng. SongCreator: Lyrics-based universal song generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/92a7a03e1c716970848a4a86cc8243ee-Abstract-Conference.html.

Ruiqi Li, Rongjie Huang, Yongqi Wang, Zhiqing Hong, and Zhou Zhao. Self-supervised singing voice pre-training towards speech-to-singing conversion. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9819–9831. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.585. URL https://aclanthology.org/2024.findings-acl.585/.

Shi Lian, Changtao Li, Bohan Li, Hankun Wang, Da Zheng, et al. dots.tts technical report, 2026. URL https://arxiv.org/abs/2606.07080.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. DiffSinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the 36th Conference on Artificial Intelligence*, volume 36, pp. 11020–11028, 2022. doi: 10.1609/aaai.v36i10.21350. URL https://ojs.aaai.org/index.php/AAAI/article/view/21350.

Zhijun Liu, Shuai Wang, Sho Inoue, Qibing Bai, and Haizhou Li. Autoregressive diffusion transformer for text-to-speech synthesis, 2024. URL https://arxiv.org/abs/2406.05551.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 11341–11345, 2024. doi: 10.1109/ICASSP48485.2024.10448291. URL https://doi.org/10.1109/ICASSP48485.2024.10448291.

Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, et al. Autoregressive speech synthesis without vector quantization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1287–1300, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.65. URL https://aclanthology.org/2025.acl-long.65/.

Yoonjeong Park, Jaekwon Im, and Juhan Nam. MeloDISinger: Melody-aware & duration-preserving singing voice editing with audio infilling, 2026. URL https://arxiv.org/abs/2606.30580. Forthcoming in Interspeech 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, et al. VibeVoice technical report, 2025. URL https://arxiv.org/abs/2508.19205.

Jiale Qian, Hao Meng, Tian Zheng, Pengcheng Zhu, Haopeng Lin, et al. SoulX-Singer: Towards high-quality zero-shot singing voice synthesis, 2026. URL https://arxiv.org/abs/2602.07803. Technical report.

Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, et al. 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*, pp. 18003–18017. PMLR, 2022. URL https://proceedings.mlr.press/v162/qian22b.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 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*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. 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*, pp. 43180–43196. PMLR, 2024. URL https://proceedings.mlr.press/v235/san-roman24a.html.

Simon Rouard, Manu Orsini, Axel Roebel, Neil Zeghidour, and Alexandre Défossez. Continuous audio language models, 2025. URL https://arxiv.org/abs/2509.06926.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, et al. NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Rc7dAwVL3v.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. CAM++: A fast and efficient network for speaker verification using context-aware masking. In *Proceedings of Interspeech 2023*, pp. 5301–5305, 2023. doi: 10.21437/Interspeech.2023-1513. URL https://www.isca-archive.org/interspeech_2023/wang23ha_interspeech.html.

Huimeng Wang, Hui Lu, Jiajun Deng, Haoning Xu, Youjun Chen, et al. SemaVoice: Semantic-aware continuous autoregressive speech synthesis, 2026. URL https://arxiv.org/abs/2605.16964.

Yongqi Wang, Ruofan Hu, Rongjie Huang, Zhiqing Hong, Ruiqi Li, et al. 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*, pp. 4780–4794. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.268. URL https://aclanthology.org/2024.naacl-long.268/.

Haojie Wei, Xueke Cao, Tangpeng Dan, and Yueguo Chen. RMVPE: A robust model for vocal pitch estimation in polyphonic music. In *Proceedings of Interspeech 2023*, pp. 5421–5425, 2023. doi: 10.21437/Interspeech.2023-528. URL https://www.isca-archive.org/interspeech_2023/wei23b_interspeech.html.

Yuning Wu, Chunlei Zhang, Jiatong Shi, Yuxun Tang, Shan Yang, and Qin Jin. TokSing: Singing voice synthesis based on discrete tokens. In *Proceedings of Interspeech 2024*, pp. 2549–2553, 2024. doi: 10.21437/Interspeech.2024-2360. URL https://www.isca-archive.org/interspeech_2024/wu24q_interspeech.html.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, et al. Ming-UniAudio: Speech LLM for joint understanding, generation and editing with unified representation, 2025. URL https://arxiv.org/abs/2511.05516.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Chenyu Yang, Shuai Wang, Hangting Chen, Jianwei Yu, Wei Tan, Rongzhi Gu, Yaoxun Xu, Yizhi Zhou, Haina Zhu, and Haizhou Li. SongEditor: Adapting zero-shot song generation language model as a multi-task editor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25597–25605, 2025b. URL https://ojs.aaai.org/index.php/AAAI/article/view/34750.

Yongyi Zang, Jiatong Shi, You Zhang, Ryuichi Yamamoto, Jionghao Han, et al. CtrSVDD: A benchmark dataset and baseline analysis for controlled singing voice deepfake detection. In *Proceedings of Interspeech 2024*, pp. 4783–4787, 2024. doi: 10.21437/Interspeech.2024-2242. URL https://www.isca-archive.org/interspeech_2024/zang24_interspeech.html.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. doi: 10.1109/TASLP.2021.3129994. URL https://doi.org/10.1109/TASLP.2021.3129994.

Lichao Zhang, Zhou Zhao, Yi Ren, and Liqun Deng. 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*, pp. 4503–4509, 2022. doi: 10.24963/ijcai.2022/625. URL https://www.ijcai.org/proceedings/2022/625.

Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, et al. Vevo2: A unified and controllable framework for speech and singing voice generation. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2262–2275, 2026a. doi: 10.1109/TASLPRO.2026.3671970. URL https://doi.org/10.1109/TASLPRO.2026.3671970.

Yongmao Zhang, Heyang Xue, Hanzhao Li, Lei Xie, Tingwei Guo, et al. VISinger2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer. In *Proceedings of Interspeech 2023*, pp. 4444–4448, 2023. doi: 10.21437/Interspeech.2023-391. URL https://www.isca-archive.org/interspeech_2023/zhang23e_interspeech.html.

Yu Zhang et al. 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, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.117. URL https://aclanthology.org/2024.emnlp-main.117/.

Ziyu Zhang, Tianlun Zuo, Hanzhao Li, Haoyu Zhang, and Lei Xie. DiTAR+: Dual optimization for robust autoregressive diffusion speech synthesis, 2026b. URL https://arxiv.org/abs/2609.13909.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, Zhiyong Wu, and Zhiyuan Liu. VoxCPM: Tokenizer-free TTS for context-aware speech generation and true-to-life voice cloning, 2025. URL https://arxiv.org/abs/2509.24650.

A Additional PSCG and Inference Diagnostics

Comparison scope. Routing variants share downstream width and training settings, but joint semantic training increases the trainable parameter count from .765B to 1.395B. We therefore separate initialization/freezing diagnostics from the additional GPU-hour-matched comparison.

Table 4: Semantic initialization and separated training controls. Dashes indicate unreported values.

Configuration	PER ↓	FPC ↑	SrcRev ↓
Pretrained / frozen (Full)	**.0376**	**.9410**	**.081**
Pretrained / joint	.0391	.9368	.089
Random / joint, 120K	.0511	.9168	.143
Random / frozen	.0918	.8724	.221
No history dropout (seeds 0/1/2)	.0440	—	.130
No melody masking	.0419	.9312	.094
Neither history dropout nor melody masking	.0468	.9242	.121
No Task Grounding (global exposure shuffle)	.0439	.9298	.102
Reordered curriculum stages	.0415	.9331	.095

Training controls. State Grounding uses an extra pretraining phase. Joint variants unfreeze the whole semantic encoder without freezing other modules to compensate. The initialization/freezing cells use 120K downstream updates and are not equal-total-compute comparisons. No Task Grounding globally shuffles the same exposure multiset, preserving broad singing/speech/curated singing at 70%/10%/20%. Reordered stages instead use curated 24K, broad 72K, then mixed 24K, and separately test the effect of curriculum order.

Table 5: Measured GPU-hours relative to Full's downstream training run.

Configuration	Updates	GPU-hour ratio	PER ↓
Random-joint	120K	1.20	.0511
Random-joint	130K	1.30	.0477
Random-joint	140K	1.40	.0444
Full + State pretraining	120K	1.40	**.0376**

GPU-hour accounting. Eight H200s were used for each run. Mean recorded GPU-hours are 574.4 for Full downstream, 229.8 for State pretraining, and 689.3/746.7/804.2 for random-joint at 120K/130K/140K. Timing includes initialization, loading, training, validation, and checkpoint saving, but excludes queueing, downloads, and offline features. Ratios divide by 574.4; Full plus State and 140K random-joint both round to 1.40. At that rounded ratio, Full reaches .0376 PER versus .0444 for random-joint.

B Illustrative Insertion Example

Case overview. Figure 5 shows one fixed LyricEditBench insertion. YingMusic+ realizes "qing qing" as "qian qian" (PER .107; FPC .830), whereas CLASVS produces the intended phones and more closely follows the reference F0 (PER .000; FPC .942). The legacy in-panel labels "YingMusic-Singer-Plus" and "CLA-SVS" denote YingMusic+ and CLASVS, respectively.

Reading the panels. The upper panel aligns reference, target, and recognized phones; the F0 and spectrogram panels compare pitch and acoustic continuity on a shared time axis. Alignment and shading are post-generation aids, not input masks or timestamps; recognized labels describe outputs.

Case interpretation. The example illustrates how the requested insertion appears in the recognized phones and how the output follows the reference pitch and local acoustic transitions. Its zero PER describes phone accuracy; the F0 and spectrogram panels provide complementary views of the same rendition. Aggregate evaluation across edits is reported in the benchmark and listening tables.

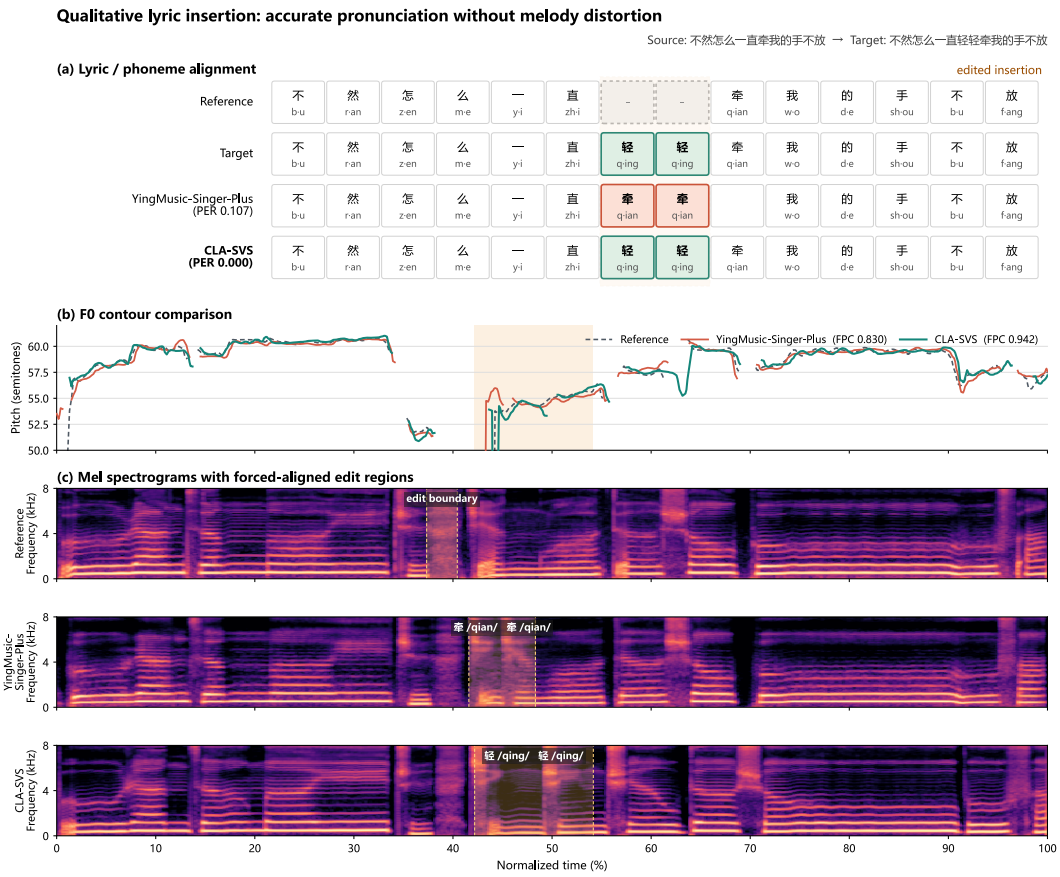

Figure 5: Qualitative lyric insertion from LyricEditBench. Shading marks the forced-aligned edit region; the target is the lyric/phone plan because no target waveform exists.

C Additional Context and Implementation Details

C.1 Extended Related Work

Score supervision and reference-conditioned editing. Score-conditioned SVS obtains precise control from aligned phonemes, notes, and durations (Liu et al., 2022; Zhang et al., 2023; Chunhui et al., 2023; Kim et al., 2024; Wang et al., 2024; Zhang et al., 2024), but their annotation is costly for arbitrary recordings. Editing systems reduce this burden through localized or duration-preserving infilling (Zhang et al., 2022; Park et al., 2026), or direct conditioning on reference singing and revised lyrics (Zhang et al., 2026a; Hao et al., 2026; Qian et al., 2026). These systems establish annotation-light editing, yet expose a selective-preservation problem: a reference carries melody, singer, and expression to retain and source lyrics to reject.

Discrete and continuous acoustic representations. Discrete neural codecs provide compact streams for acoustic language modeling (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023; Wu et al., 2024), but their finite codebooks can remove subtle variation. Continuous diffusion, flow, and tokenizer-free models reduce reliance on fixed acoustic codebooks while preserving rich acoustics (Le et al., 2023; Shen et al., 2024; Mehta et al., 2024; Liu et al., 2024; Rouard et al., 2025; Lian et al., 2026; Peng et al., 2025; Yan et al., 2025; Zhou et al., 2025); MELLE further reports robustness and quality gains over codec-LM counterparts (Meng et al., 2025). These results motivate continuous latent targets, but do not by themselves establish a causal advantage for lyric editing. The main paper therefore treats Vevo2 and YingMusic+ as released-system comparators and relies on matched within-stack variants for architectural attribution.

Continuous-latent autoregression and editing-specific state. DiTAR aggregates each realized acoustic patch into an inter-patch language model, while SemaVoice aligns continuous representations with a speech foundation model before patch-wise AR diffusion (Jia et al., 2025; Wang et al., 2026). Both address synthesis with content-consistent conditioning. Reference-conditioned lyric editing instead changes the requested content while reference-derived controls persist, so a source-lyric-biased generated patch can propagate through AR history. Self-supervised representations suggest separate phonetic-progress summaries (Li et al., 2024; Qian et al., 2022; Chen et al., 2022). The SCT history-access contract uses this separation: semantic feedback returns to the causal planner, whereas the previous AudioVAE latent patch remains local to the Transition.

C.2 Frozen Interfaces and Time Bases

Frozen feature contract. The AudioVAE encoder/decoder come from Ming-UniAudio (Yan et al., 2025) and remain frozen. The melody tokenizer takes 16-kHz mono audio, applies a 4096-point Hann STFT at 320-sample hop, and folds 65–2093-Hz energy to 24 chroma bins per octave. Each frame receives L1 normalization, $\log(1 + 10c)$, then L2 normalization; silent frames are zero. A three-layer residual Conv1D encoder (24/128/256/256 channels, kernel 5, stride 2) produces 128-D vectors at 6.25 Hz. A 512-by-128 VQ codebook uses EMA updates (decay .99); a symmetric decoder reconstructs chroma with L1 plus .25 commitment loss. It trains for 100K updates on broad singing (batch 128, 12.8-s crops, AdamW 2×10^{-4}, 2K warm-up), selecting lowest development reconstruction error. This chroma path limits direct lexical access but cannot guarantee its removal. Five 20-ms AudioVAE frames form one causal patch; a short final patch is zero-padded and retained.

Table 6: Frozen interfaces and training status. The feature encoders are frozen in Full; the feedback projection is trained.

Interface	Rate / output	Status / archive identifier
Melody tokenizer	∼6.25 Hz / 512-way token	Frozen / `melody_tok_v3`
Singer encoder	utterance / 192-D vector	Frozen / `campp_2025q1`
AudioVAE encoder	50 Hz / 64-D latent	Frozen / `avae64_50hz`
AudioVAE decoder	50 Hz / 16-kHz waveform	Frozen / `avae64_50hz`
Semantic encoder	32 layers, 1,280-D / trainable projection to 1,024-D	Frozen / `semantic-v1.0`
Singing ASR	utterance / Mandarin transcript	Frozen / FireRedASR-AED-L
Speaker verifier	utterance / cosine score	Frozen / WavLM-base-plus-SV

Tuple construction. For each retained recording, the pipeline (i) verifies a transcript and singer identifier; (ii) creates the reference waveform and target acoustic latent from the same performance; (iii) derives melody tokens and a singer vector; (iv) rejects benchmark song/recording overlap before splitting; and (v) stores duration, token counts, and component hashes. Song-family disjointness follows fingerprint and title–singer deduplication. Singer overlap is recorded rather than hidden: the main benchmark is song-held-out, whereas the separate 24-singer corpus was recorded only after checkpoint freeze and is singer-disjoint by construction. PSCG training uses same-recording reconstruction without paired counterfactual edit recordings. Edited tests thus measure generalization to unpaired lyric changes.

Inference procedure. The planner caches target lyrics and the full reference-melody sequence. Source-lyric-correlated evidence can enter only through frozen reference conditioning; the previous latent patch is generated, never copied from reference audio. At step k, the semantic encoder returns semantic feedback for the realized prefix, the previous latent patch supplies local acoustics, and the stop head predicts completion. Content-consistent reconstruction never presents revised lyrics against reference-derived controls, and a source-lyric-biased early patch can propagate through ordinary AR continuation. The SCT history-access contract therefore confines the previous latent patch to the one-patch Transition while returning only semantic feedback to the causal planner. For all reported quality results, Flow-DiT maps Gaussian noise to the next continuous latent patch with 24 Euler steps and guidance 2.0. Generation stops when the posterior is at least .50 or when the rollout reaches $\min(K_{\text{ref}} + 8, 256)$ patches. This budget is a safety guard rather than a duration-ratio rule or benchmark filter, and every produced candidate remains in evaluation.

Inference pseudocode. Given reference audio r, target lyric x, and sampling seed, the inference routine: (1) freezes singer vector v and reference-melody tokens m and caches the causal planner prefix; (2) initializes semantic feedback and the previous latent patch with zeros; (3) draws fresh patch noise from the indexed seed and applies 24 guided Euler updates; (4) appends the complete generated patch, updates both recurrent inputs, and records stop posterior; (5) repeats until the posterior is at least .50 or the $\min(K_{\text{ref}} + 8, 256)$ budget is reached; and (6) decodes once at 16 kHz and writes waveform, stop trace, seed, failure status, and all metric inputs. No candidate is redrawn.

C.3 Optimization and Inference Configuration

Table 7: Default CLASVS configuration used in the reported experiments.

Field	Value
Planner	Qwen3-0.6B initialization
Semantic encoder (frozen)	32 layers, width 1,280; 1,024-D projection
Flow-DiT	8 blocks, width 1,024
Optimizer	AdamW, $\beta = (.9, .95)$
Peak learning rate	2.0×10^{-4}
Weight decay / clip norm	.01 / 1.0
Warm-up / schedule	5k / cosine
Global batch / updates	128 / 120k
Precision / sharding	BF16 / ZeRO-2
Hardware	8× NVIDIA H200 (training); 1× H200 (inference)
Median downstream wall time	71.8 h on 8× H200
Software	Ubuntu 22.04, PyTorch 2.3.1
Flow steps / guidance	24 / 2.0
Flow σ / temperature	0 / 1.5 (inactive at $\sigma = 0$)
Patch / history-drop rate	5 frames / .10
Joint condition-drop rate	.10
Stop threshold / budget	.50 / $\min(K_{\text{ref}} + 8, 256)$
Melody perturbation rate	.15
Loss weights	$1 : 10 : 1$

Losses and frozen-state supervision. The complete training objective is

$$\mathcal{L}_{\text{SVS}} = \mathcal{L}_{\text{text}} + 10\mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{stop}}.$$

Loss components. $\mathcal{L}_{\text{text}}$ is causally shifted cross-entropy on nonpadding target-lyric positions during prefix teacher forcing; melody tokens, separators, and recurrent semantic-feedback positions are excluded. Its gradients update shared planner parameters through lyric-prefix states. The supplied lyrics remain conditioning at inference, with no separate transcript decoding. $\mathcal{L}_{\text{FM}}$ directly supervises acoustic realization through planner-conditioned Flow-DiT: mean-squared vector-field error covers five-frame AudioVAE patches, including the zero-padded final patch. $\mathcal{L}_{\text{stop}}$ is inverse-frequency-weighted cross-entropy over valid patches: nonfinal patches are "continue" and the final patch is "stop." Thus the text objective trains the shared lyric-conditioning model, while flow and stop losses supervise the generated acoustic sequence. Character CTC and multi-scale log-mel losses pretrain the semantic encoder; the 512-way melody tokenizer uses 24-bin chromagrams. Both frontends remain frozen downstream.

Control corruptions. During PSCG Control Grounding, history dropout sets the previous latent patch to zero with probability $p = .10$; it never substitutes another utterance. Melody perturbation replaces one reference-melody span with the learned mask token ($p = .15$). Both choices are fixed before evaluating the locked final test set.

Checkpoint selection. CLASVS is trained with three independent seeds. Model selection uses a development manifest disjoint from the final 320 items. Each run independently selects the checkpoint minimizing development PER; no RTF threshold or edit-type-specific rule enters checkpoint selection. The reported quality configuration (24 steps, CFG 2.0), stopping threshold, patch size, loss weights, and perturbation rates are then locked before test generation. A 32-step option is retained for auxiliary internal runs but does not generate any reported table or listening result.

Semantic pretraining recipe. The custom encoder starts from random weights, with 32 layers, hidden size 1280, 20 heads, and FFN dimension 5120. A linear $64 \to 1280$ input adapter preserves the 50-Hz latent rate; sinusoidal positions and a frame-causal attention mask support incremental KV caching. Training uses 100K updates, global batch 128, BF16, and an 80%/20% broad-singing/Emilia-speech mixture excluding development and test songs. AdamW uses LR 10^{-4}, betas $(.9, .98)$, weight decay .01, and gradient clipping 1.0; 2K warm-up updates precede cosine decay to 10^{-5}. Validation runs every 2K updates; the lowest fixed-development character error rate selects the checkpoint, with earlier checkpoints breaking ties.

Pretraining targets. Character CTC is the ASR objective; its auxiliary head is used only in pretraining. The acoustic auxiliary head predicts multi-scale log-mel targets using FFT sizes 256/512/1024, hop 320 samples, and 80 mel bins; the loss averages L1 across scales. The total State objective is CTC plus five times this loss. Downstream Full freezes the encoder and input adapter, while the 1280-to-1024 feedback projection remains trainable outside the no-gradient encoder computation.

History and variant implementation. Both training histories use true latent patches. Local history dropout zeros only the predecessor input to Flow-DiT and does not corrupt the semantic-encoder history. There is no generated-history roll-in. Caches reset across samples and gradients are not truncated within a sample; joint variants do not detach semantic feedback. The singer vector passes through a shared adapter and each Flow-DiT block independently maps it to AdaLN scale/shift. It never enters the planner. CFG uses a learned null singer embedding.

Unified routing controls. U-old appends $\mathbf{a} + \mathrm{Linear}_{320\to1024}(\mathrm{vec}(\bar{\mathbf{z}}))$ as one planner token; its local generator keeps five predecessor, five current, and one time/planner token. Frontends, freezing, exposure, optimizer, steps, selection, and inference match Full. The projection adds 328,704 parameters (.76533B trainable). U-old retains local history dropout $p_h = .10$ but leaves the planner's acoustic bypass unmasked. U-sync-add uses the same Bernoulli mask for the global acoustic projection and local predecessor. U-sync-gate retains synchronized masking and replaces direct addition with $g = \sigma(\mathbf{w}^\top[\mathbf{a}; \mathrm{LN}(W\,\mathrm{vec}(\bar{\mathbf{z}}))] + b)$ and $\mathbf{u} = \mathbf{a} + g\ \mathrm{LN}(W\,\mathrm{vec}(\bar{\mathbf{z}}))$. Both new variants retain the planner-token count and local generator layout. This progression changes masking and fusion together with access. The separate matched pair below isolates acoustic access within the gated recipe.

Table 8: Relationship among the routing variants. "Real" means the generated predecessor reaches the planner cache; "null" has the same input shape without predecessor information. All arms retain the local predecessor path. The matched pair isolates access only within its common gated configuration.

Arm	Planner acoustics	Dropout relation	Fusion
Full SCT + PSCG	None	Local only	Semantic token
U-old	Real	Planner unmasked	Additive
U-sync-add	Real	Shared mask	Additive
U-sync-gate	Real	Shared mask	Gated
Matched local-only	Null	Shared mask	Gated
Matched global-access	Real	Shared mask	Gated

Table 9: Operation-resolved PER for the synchronized unified controls on CLA-LyricEdit-320. Positive paired macro differences favor Full; brackets are 95% CIs for each variant minus Full.

Configuration	PSub	FSub	Del	Ins	Macro	Δ vs Full [CI]
Full	**.0505**	**.0442**	**.0260**	**.0298**	**.0376**	—
U-old	.0550	.0490	.0360	.0400	.0450	+.0074 [+.0048,+.0100]
U-sync-add	.0530	.0470	.0325	.0347	.0418	+.0042 [+.0019,+.0066]
U-sync-gate	.0522	.0462	.0320	.0328	.0408	+.0032 [+.0009,+.0055]

Seed and preservation audit. Macro-PER for seeds 0/1/2 is .0365/.0376/.0387 (Full), .0440/.0451/.0459 (U-old), .0408/.0418/.0428 (U-sync-add), and .0398/.0408/.0418 (U-sync-gate). U-sync-add FPC by seed is .936/.935/.934 and SrcRev is .101/.103/.105; U-sync-gate gives .938/.937/.936 and .094/.096/.098. All four operation-level point estimates improve from U-old

to each synchronized variant. Table 9 reports paired differences against Full; the matched pair in Table 10 directly tests acoustic access within the gated configuration.

Table 10: Matched planner-cache access ablation on CLA-LyricEdit-320. Both arms use the gated configuration and 576 training GPU-hours per seed. Only the planner-cache acoustic input changes from a fixed-null tensor to the real predecessor; parameters, planner tokens, projection, layer normalization, gate, local path, and dropout mask are identical.

Arm	PSub	FSub	Del	Ins	Macro	FPC	SrcRev
Matched local-only	.0515	.0454	.0293	.0298	.0390	.9395	.088
Matched global-access	.0522	.0462	.0320	.0328	.0408	.9370	.096

Global-access − local-only	Paired difference [95% CI]	Seed 0/1/2 macro-PER
Macro-PER	+.0018 [+.0004,+.0032]	Local: .0380/.0390/.0400
PSub	+.0007 [-.0011,+.0025]	Global: .0398/.0408/.0418
FSub	+.0008 [-.0010,+.0026]	
Del	+.0027 [+.0005,+.0049]	
Ins	+.0030 [+.0006,+.0054]	

Notes. Intervals use paired song-clustered bootstrap (10,000 resamples, seeds 0/1/2). Only the macro, Del, and Ins intervals exclude zero. Operation-level intervals are diagnostic and have no multiplicity adjustment.

C.4 Comparator Construction and Invocation

Table 11: Comparator eligibility audit and configurations under the shared one-candidate protocol (Zhang et al., 2022; Lei et al., 2024; Yang et al., 2025b; Park et al., 2026; Qian et al., 2026; Zhang et al., 2026a; Hao et al., 2026).

A. Candidate audit at the evaluation freeze

System	Checkpoint	Interface / protocol mismatch	Decision
EditSinger	Not located	Externally aligned phone durations/pitch and an edit mask/region	Context
MeloDISinger	Not located	Internally derived MFA durations/pseudo-MIDI; fixed edit span	Context
SoulX-Singer	Released	Auto metadata available; uncorrected four-operation call unverified	Context
SongCreator	Not located	Full-song/stem editing rather than the benchmark interface	Context
SongEditor	Not located	Sentence/span boundary; full-song/track editing	Context
Vevo2	Released	None	Included
YingMusic+	Released	None	Included
CLASVS	Ours	None	Evaluated

B. Included-system configuration

Attribute	Vevo2	YingMusic+	CLASVS
Representation	Discrete	Continuous	Continuous
Generation	AR	NAR	AR
Quality inference	Official recipe	Official recipe	24 Euler; CFG 2.0
Checkpoint status	Released	Released	Ours

Notes. Audit frozen September 20, 2026; "Not located" means no official runnable editing checkpoint appeared on the cited paper/project page. Internal automatic preprocessing is allowed. "Context" records an artifact or interface mismatch, not a manual-input requirement or lower system quality.

Comparator audit. Panel A of Table 11 records screened editors and whether an official, uncorrected four-operation invocation was verified at the freeze. SoulX-Singer's automatic preprocessing is allowed by our protocol; its context status reflects an unverified invocation. Panel B summarizes the included systems. Final inference settings appear below.

Model and training accounting. The trainable downstream modules have 0.765B parameters. The frozen AudioVAE has 1.350B parameters, giving a 2.115B generator–codec subtotal; the frozen se-

mantic encoder contributes approximately 0.630B more, with melody and singer frontends listed separately. Componentwise accounting distinguishes the trained editing model from the frozen audio representation and conditioning modules used for generation. Median downstream training takes 71.8 hours per run on eight NVIDIA H200 GPUs, excluding State pretraining. Appendix Table 5 includes State pretraining in the within-stack training-budget comparison.

Released-system protocol. External baselines are run with released checkpoints and official preprocessing where compatible. Permitted temperature, guidance, or decoding sweeps use the common development manifest, and Table 12, Panel B, reports the full range rather than only the selected point. Systems requiring human alignment, edit boundaries, or unavailable weights are retained as contextual comparators outside the headline benchmark ranking.

Development selection. Table 12 records the settings selected on development data before test evaluation. Comparator sweeps use the inference controls supported by their released checkpoints.

Table 12: Development search, locked settings, and comparator model-selection budgets.

A. CLASVS development program

Factor	Range	Count	Set	Selection criterion
Flow steps	8, 16, 24, 32	4	24	Dev. PER, then cost
Guidance scale	1.0–3.0	5	2.0	Dev. PER/FPC
History dropout	0–.30	5	.10	Macro-PER; SrcRev
Melody masking	0–.20	4	.15	Macro-PER/FPC
Patch size	5 frames	1	5	Fixed interface rate
Loss weights	1:10:1	1	1:10:1	Fixed before final sweep

B. Per-system quality-selection budget

System	Trials	Selected	Range
Vevo2	6	temp. .9	temp. .7–1.1; top-p .90–.98
YingMusic+	12	24 steps, CFG 2.0	steps 8–32; CFG 1.0–3.0
CLASVS	26	24 steps, CFG 2.0	disclosed development program

Notes. The 26 CLASVS settings form a structured, non-Cartesian program. The development matrix records every attempted combination, seed, checkpoint, and selection statistic.

Final per-system configurations. Vevo2 uses its released editing checkpoint, official reference/target preprocessing and stopping rule, temperature .9, and all remaining publisher defaults. YingMusic+ uses its released checkpoint and preprocessing at 24 flow steps and CFG 2.0, with other publisher defaults unchanged. CLASVS uses Table 7, including 24 Euler steps, CFG 2.0, serialized temperature 1.5 with flow $\sigma = 0$, posterior threshold .50, and the $\min(K_{\text{ref}} + 8, 256)$ rollout budget. The auxiliary cross-system runtime comparison uses separate recipes: Vevo2 retains its official editing recipe, YingMusic+ uses 32 NFE/CFG 3.0, and CLASVS uses 24 steps/CFG 2.0. These auxiliary timings are separate from the quality-configuration CLASVS measurement in Appendix G.

D Benchmark and Statistical Protocol

D.1 Benchmark Construction

Primary benchmark design. All training and evaluation data used in this work are Mandarin Chinese; no English or multilingual result is reported. CLA-LyricEdit-320 is newly constructed for this study and is not an existing YingMusic+ evaluation set. It contains 320 Mandarin edits from 160 reference excerpts, 80 songs, and 40 singers, with 80 items per operation. PSub replaces one contiguous 2–6-syllable span while retaining source context; FSub replaces every lexical syllable in one line. Deletion removes 2–6 syllables, and insertion adds 2–6 syllables at a word boundary. Two annotators verify target, pronunciation, boundary, operation, and plausibility; operation agreement is $\kappa = .93$, with disagreements adjudicated before generation. Neither target length nor the inference budget is used to select benchmark items.

Why these Mandarin data are needed. The three training slices have distinct roles. Approximately 8,000 hours of permission-documented Mandarin singing provide singer, genre, phrasing, and recording-condition coverage that available aligned corpora do not supply at comparable scale. The 2,000-hour Mandarin subset of the public Emilia corpus (He et al., 2024) broadens lexical and speaker coverage for semantic grounding, but speech alone cannot represent sustained pitch, vibrato, or singer-specific phrasing. Finally, 300 hours of curated open-source Mandarin singing emphasize clean vocals and reliable lyric–melody correspondence for editing fidelity. This broad-to-curated composition is motivated by coverage, semantic robustness, and singing-specific acoustics.

Table 13: Evaluation resources and benchmark construction.

Novel study set	Size	Split unit	Data-appendix contents
CLA-LyricEdit	320 edits	80 songs	Items, operations, targets, hashes, exclusions
Post-freeze transfer	192 edits	24 singers	Consent basis, items, outputs, scores
Listening set	240 outputs	80 base edits	Sampling seed, blind order, de-identified ratings

Dataset provenance. All datasets adopted from the literature in the reported experiments are cited and publicly accessible, including the Mandarin Emilia subset (He et al., 2024) and the released YingMusic+ lyric-editing benchmark (Hao et al., 2026). Permission-documented performances and post-freeze consented recordings are author-assembled sources; their provenance and access basis are recorded in the frozen manifests.

Disjointness and singer transfer. Duplicates are rejected at song level. Exact metadata and fingerprints reject overlap; all near-duplicate embedding candidates against train/development inventories are manually adjudicated before manifest freeze. The frozen internal and public manifests retain zero exact metadata/fingerprint matches. The main manifest is song-held-out, not singer-held-out. For a disjoint audit, 24 consented singers record two reference excerpts and eight balanced edits each (one per operation for each excerpt) only after checkpoint freeze, yielding 192 edits from absent identities. CLASVS/YingMusic+ overall PER is .0472 [.0416,.0533]/.0521 [.0461,.0586]; PSub/FSub/Del/Ins PER is .0580/.0510/.0380/.0418 versus .0250/.0300/.0960/.0574. CLASVS/YingMusic+ FPC and SIM are .934/.905 versus .928/.898; DUR-A/R1 and voiced-onset MAE are 1.18%/9.4 ms versus .21%/21.6 ms. The fixed post-freeze audit added no listening study.

Public Mandarin benchmark. The published LyricEditBench construction samples 300 items per operation and gender/language cell, yielding 600 Mandarin items per operation and 2,400 across PSub/FSub/Del/Ins (Hao et al., 2026). We evaluate this four-operation portion under one automatic-input and FireRedASR-based protocol. Vevo2 obtains PSub/FSub/Del/Ins PER .0664/.0607/.1123/.0649 (macro-PER .0761 [.0672,.0850]), compared with CLASVS .0508/.0435/.0398/.0357 (macro-PER .0425 [.0368,.0485]). Re-evaluated YingMusic+ gives .0220/.0230/.0930/.0440 (macro-PER .0455 [.0384,.0506]), distinct from the publisher's .0214/.0186/.0946/.0426 under its own ASR. CLASVS remains weaker on PSub/FSub but stronger on Del/Ins. The reported system-level intervals do not give a paired public-set difference interval, so the $-.0030$ macro difference is descriptive.

D.2 Metrics and Statistical Analysis

Lyric scoring. FireRedASR-AED-L first transcribes each generated waveform. Mandarin transcripts and target lyrics are converted to phone sequences with pypinyin, then PER is computed as phone Levenshtein distance divided by target-phone count; macro-PER is the unweighted mean of the four operation means. For the content audit only, NED divides by the larger compared sequence length. Text normalization standardizes Chinese characters and number readings, removes punctuation, and uses pypinyin initials and finals for primary PER. No manual transcript correction is applied.

Performance scoring. Reference-duration deviation (RefDur; DUR-A/R1) is $100|D_{\text{out}} - D_{\text{ref}}|/D_{\text{ref}}$ and measures adherence to the reference rather than an unavailable counterfactual target. It is a global endpoint measure and does not score local phrasing. FPC is common-voiced RMVPE log-F0 Pearson correlation; Onset is voiced-onset MAE. Singer similarity (SIM) is WavLM-base-plus-SV cosine between reference and output. FPC uses an absolute 10-ms grid without DTW or per-item stretching; common voiced support below 100 ms is marked missing and counted. Onsets

use voiced segments of at least 30 ms, bridge gaps up to 20 ms, and match monotonically within 200 ms; unmatched proportions accompany the metric. Appendix G gives the separate quality-configuration timing protocol.

Objective uncertainty. The primary estimand is the song-level macro-PER difference on the fixed manifest, averaged over three CLASVS training runs and conditional on each released comparator checkpoint. Two-sided 95% percentile intervals use 10,000 hierarchical draws that resample the 80 song clusters and their edits, then resample the three CLASVS runs. A multiway song–singer pigeonhole bootstrap gives CLASVS–YingMusic+ $-.0034$ $[-.0060, -.0008]$. Four inference-noise repeats on a fixed stratified 80-item subset give macro-PER mean$\pm$SD .0701$\pm$.0007/.0412$\pm$.0004/.0379$\pm$.0006 for Vevo2/YingMusic+/CLASVS, preserving the system and operation ordering. Overall PER is tested first, followed by secondary objective and perceptual families. Per-operation, ratio-bin, intervention, and failure analyses describe the sources of the aggregate results. Listener contrasts use Holm correction across their prespecified family. Released-system comparisons evaluate complete editing pipelines; the gated matched pair estimates the effect of direct planner acoustic access under its shared training and fusion configuration.

Macro interpretation. The pre-specified equal-weight macro matches the benchmark design: each of the four editing capabilities contributes exactly 80 items, so no frequent or easier class can dominate the system score. It is a capability-balanced benchmark estimand rather than an assumed product-usage frequency. Operation cells reveal the contributions behind the aggregate score.

Listening-model specification. Paired MOS uses this mixed-effects structure:

$$\text{MOS} \sim \text{system} * \text{operation} + (1 \mid \text{listener}) + (1 \mid \text{song}) + (1 \mid \text{base_edit}) + (1 \mid \text{output}).$$

Random effects and sensitivity. A base edit is shared across systems, while output identifies a particular system's rendition. Listener, song, base-edit, and output intercepts represent repeated observations at these levels. Contrasts are expressed on the MOS score scale. The protocol uses 10,000 clustered bootstrap replicates that preserve system pairing, with Holm correction for the listener-contrast family. Leave-one-base-edit-out analysis is the specified content-sensitivity check. The tables report aggregate system contrasts and their uncertainty over the paired listening set.

How to read the intervals. Table 14 combines objective and listener contrasts with their respective sampling units. PER resamples songs and three trained CLASVS runs against the fixed released YingMusic+ checkpoint; MOS compares systems on the shared 80-edit set while accounting for listeners and content. These are fixed-checkpoint system comparisons. The public benchmark reports individual-system intervals, including [.0384,.0506] for re-evaluated YingMusic+. Its $-.0030$ CLASVS–YingMusic+ macro difference is a point-estimate comparison, since a paired public-set difference interval is not reported. Operation-level results show the same deletion/insertion advantage across the primary and public sets. Primary paired objective contrasts and paired listener contrasts provide the inferential results used in the main text.

Table 14: Fixed-checkpoint CLASVS–YingMusic+ contrasts.

Metric	CLASVS minus YingMusic+ [95% CI]	Holm p
PER	$-.0035$ $[-.0056, -.0013]$	.004
FPC	$+.0070$ $[+.0024, +.0116]$	.004
SIM	$+.008$ $[+.004, +.012]$	.003
N-MOS	$+.48$ $[+.34, +.62]$	$< .001$
M-MOS	$+.01$ (point estimate only)	.79
L-MOS	$+.32$ $[+.20, +.44]$	$< .001$

Notes. PER resamples three CLASVS runs; the released YingMusic+ checkpoint remains item paired. M-MOS is descriptive; its point estimate establishes no equivalence.

E Extended Evaluation Audits

Source-versus-target control. Audit NED is phone edit distance divided by the larger of the two compared sequence lengths. For PSub/FSub/Del/Ins, output–target NED is .0490/.0430/.0252/.0289 (overall .0365), output–source is .3107/.7639/.2500/.2954 (overall .4050), and the direct target–source requirement is .3020/.7600/.2380/.2860 (overall .3965). The corresponding target-preference

margin (output–source minus output–target) is .2617/.7209/.2248/.2665 (overall .3685). Every operation remains target-preferring when both distances use the shared maximum length of output, target, and source, and when raw edit distances are compared without normalization. Thus FSub exceeds PSub in both required and realized lexical change, while every operation retains a positive preference for the target. On 80 no-edit items, output–target/source NED is .018/.018, PER is .019, and target-preference margin and SrcRev are both zero because source and target coincide. These direct output comparisons audit editing behavior without relying on a learned representation-accessibility proxy; Figure 5 instantiates the lexical and acoustic views on one insertion case.

Matched continuous-AR and released-comparator comparisons. The unified-context comparison and compute budgets are specified in the main paper and Table 5. Released Vevo2 and YingMusic+ checkpoints provide system-level reference points under the same automatic-input evaluation. The original Full runs give macro-PER .0365/.0376/.0387 (mean .0376, SD .0011). Removing $\mathcal{L}_{\text{text}}$ gives PER/FPC/SrcRev .0458/.9382/.137, versus Full's .0376/.9410/.081. This control measures the auxiliary text objective's contribution to target accuracy, melody preservation, and source reversion. The access-matched comparison and representation probes address the corresponding routing and information-accessibility questions.

State intervention and progress probe. Each original/intervened pair shares item, controls, initial flow noise, and rollout budget. Only already-generated history is disturbed; reordered history requires rebuilding the planner cache. Replacement donors are fixed in advance and matched within 10% duration and .05 normalized position. The lyric margin is output–source NED minus output–target NED, with each NED normalized by the larger of its own two sequence lengths. The melody margin is output–intended-reference FPC minus output–fixed-donor-reference FPC. Retention divides aggregate intervention margin by aggregate Full margin. Semantic-history shuffling changes lyric margin .3685 to .2469 (retention .67); predecessor shuffling changes melody margin .200 to .170 (.85). Both interventions perturb only generated history under shared controls.

Progress-label and probe protocol. An independent ASR transcript and timestamps are monotonically aligned to target phones by edit distance. The last matched target position at a patch, divided by target-phone count, defines a reached-position label. The split is by song, 70%/15%/15%. The diagnostic comparison uses standardized features and L2-regularized linear regression: semantic feedback .082 MAE, acoustic latents .191, elapsed seconds .180, elapsed/reference duration .120, and normalized time plus semantic feedback .080. The semantic MLP supplement gives .078 and uses the separate 128-D projection/64-hidden-unit configuration. The linear comparison and the nonlinear supplement are not pooled. Reference duration makes the normalized-time baseline reference-informed.

Acoustic-side-channel probes. Capacity-matched frozen side-channel diagnostics use a song-disjoint split. Singer top-1 accuracy is 8.6% from semantic feedback versus 91.2% from AudioVAE latents (2.5% chance), while frame-level log-F0 R^2 is .09 versus .84. With semantic progress MAE .082, this suggests a progress-sensitive State with weak singer/pitch information.

Stopping protocol. Each current patch is generated and retained before checking stop posterior $\geq .5$. The cap is $\min(\lceil T_{\text{ref}}/.1 \rceil + 8, 256)$. Reaching this cap and being forced to stop without a positive learned stop decision are recorded separately. Endpoint deviations beyond ± 200 ms are flagged relative to reference duration; target-phone coverage at the end is a separate lyric-completion check. No failed output is redrawn.

Main-set stopping diagnosis. For PSub, FSub, deletion, and insertion, budget-hit rates are (0, 0, 0, 1.25)%, premature-stop rates are (1.25, 1.25, 0, 1.25)%, and delayed-stop rates are (1.25, 0, 1.25, 2.50)%. In PSub/FSub, substitutions account for 61/64% of phone errors, versus 23/20% deletions and 16/16% insertions. These measurements distinguish lexical replacements, phone omissions or additions, endpoint deviations, and budget-limited termination. The substitution error profile is dominated by lexical replacement, while the budget is reached infrequently in the evaluated edits.

Absolute pitch and local-preservation audits. Metrics that do not discard pitch offset or unvoiced errors give the same profile. Vevo2/YingMusic+/CLASVS semitone RMSE is 1.92/.84/.76, gross pitch error above two semitones is 14.1/5.6/4.8%, and framewise V/UV F1 is .908/.958/.963.

Table 15: Timing fidelity and auxiliary single-H200 runtime measurements.

A. Operation-resolved reference timing

System	PSub	FSub	Del	Ins	Mean
Vevo2	1.6118%	1.6603%	1.5319%	1.5939%	1.5995%
YingMusic+	.1522%	.1448%	.1512%	.1377%	**.1465%**
CLASVS	.9922%	1.0516%	1.0152%	.9034%	.9906%

Voiced-onset MAE: Vevo2 69.58 ms; YingMusic+ 18.80 ms; CLASVS **7.68 ms**.

B. Auxiliary runtime recipes

System	Cold RTF↓	Steady RTF↓	ΔGiB↓
Vevo2	**2.649**	.744	8.62
YingMusic+	3.392	**.241**	6.74
CLASVS	3.840	1.395	**5.82**

Notes. Panel A reports absolute output/reference duration deviation; each mean averages four equally sized operation cells. Onset MAE reports local timing. Panel B uses separate runtime recipes: Vevo2 uses its official recipe, YingMusic+ 32 steps/CFG 3.0, and CLASVS 24 steps/CFG 2.0. Timing boundaries were not fully specified, so these cannot rank quality-matched efficiency.

On PSub, the reference outside the edited span (excluding a 200-ms boundary collar) is directly observable: energy-normalized log-spectral distance is .184/.112/.097, SIM is .899/.912/.923, and edit-boundary onset MAE is 55.1/20.2/8.0 ms.

Failure labeling and held-out audit. The 64 largest standardized metric failures are selected automatically before annotation. Two annotators assign one primary label ($\kappa = .81$): lexical substitution (23), stopping (12), pitch/rhythm drift (16), singer drift (8), or artifact (5). The target-margin/stop-entropy/budget-proximity union marks 11.6% of outputs with .71 precision/.68 recall and never rejects or regenerates a candidate; none of its components exceeds 3/8 recall on singer drift. On the post-freeze 192-edit, 24-singer audit, CLASVS/YingMusic+ obtains PER .0472 [.0416,.0533]/.0521 [.0461,.0586]. On 80 no-edit items, CLASVS gives PER .019, RMVPE FPC .960, and SIM .932.

F Listening Protocol

Paired study design. The study uses 80 base edits, 20 from each operation, with Vevo2, YingMusic+, and CLASVS generated for every edit (240 audios). A pool of 20 native Mandarin listeners supplies eight ratings per audio, yielding 1,920 audio ratings and 5,760 individual MOS values. The independent content unit is a base edit, not a rating. Shared base edits provide system pairing; individual listeners need not rate every audio. Blind assignment records permit an audit.

Stimulus and scoring protocol. The checkpoint follows the development-PER rule and is frozen before rendering; ratings do not select seeds or candidates. Audio is mono 16 kHz, EBU-R128 normalized to -23 LUFS and true-peak limited to -1 dBTP. Listeners see the target lyric, hear the reference once and the output up to twice, with randomized output order and hidden system identities. N-MOS rates naturalness, clarity, and artifacts; M-MOS melody and phrasing; L-MOS scores the intelligibility and correctness of the target lyrics.

Table 16: Paired listening study: 80 edits, three systems, eight ratings per audio.

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 p	$< .001$	.79	$< .001$

Paired inference. The listening analysis specifies system-by-operation fixed effects and random intercepts for listener, song, base edit, and output. Ten thousand clustered bootstrap replicates preserve

system pairing; leave-one-base-edit-out is the specified sensitivity analysis, and reported contrasts use Holm correction. N-MOS improves by .48 [.34,.62], adjusted $p < .001$; L-MOS improves by .32 [.20,.44], adjusted $p < .001$. The study reports aggregate means and paired contrasts. Per-operation means and individual-system intervals are descriptive. The .01 M-MOS gap ($p = .79$) does not establish equivalence across systems.

Listener safeguards. Participants provide informed consent, may stop without penalty, and may withdraw before anonymization. Attention checks evaluate instruction compliance rather than agreement with a preferred system.

G Quality-Configuration Efficiency Protocol

Matched quality setting. The quality-configuration measurement keeps the reported quality configuration (24 Euler steps, CFG 2.0), uses one H200 at batch one, warms up on 10 items, then evaluates a fixed stratified 64-item subset for three rounds. Cold-start measurements use five new processes. Timing synchronizes GPU work and separates cold end-to-end, warm end-to-end, and generation-only execution. Warm end-to-end RTF is 1.45; incremental peak memory is 5.8 GiB and total resident peak is 9 GiB. Only warm end-to-end RTF is reported for this quality configuration. Auxiliary Table 15 uses different recipes and cannot support a valid ranking of quality-matched efficiency.

H Run and Evaluation Protocol

Run identifiers and randomness. Training uses seeds 0/1/2 and run identifiers of the form configuration–data-version–seed–checkpoint-step. Initialization, data order, dropout, and flow noise use separate seeds. Evaluation seeds derive from item and run identifiers; separate seeds govern the 10,000 bootstrap draws and listener sampling and order. Caches reset between samples, and each reference–target–seed tuple produces one candidate.

Item-level accounting. Evaluation items are indexed by split, song family, singer, operation, source and target transcripts, reference, and exclusion status. Output identifiers include system, configuration, data version, seed, checkpoint, waveform hash, and stop trace. These keys join generation, ASR transcription, phone conversion, scoring, and song-clustered resampling. Missing metrics retain reasons and denominators; no candidate is replaced after scoring. Training-audio source records retain permission categories, work families, and exclusion status.