REPLAN-VC: TARGET-SIDE SEMANTIC AND PROSODIC REPLANNING FOR CONTROLLABLE VOICE CONVERSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable voice conversion aims to modify speaker identity and expressive attributes while preserving linguistic content. Existing methods either retain source-derived content anchors while replanning selected factors or generate target-side states without explicitly modeling dependencies among functionally distinct states, limiting coordinated multi-state replanning. We propose **RePlan-VC**, a target-side semantic and prosodic replanning framework that regenerates semantic and prosodic trajectories under target speaker and emotion conditions. RePlan-VC employs a hierarchical autoregressive planner that generates semantic states before prosodic states, explicitly modeling their generative dependency, while target-condition modulation directly guides state prediction. To preserve linguistic content without enforcing source state reproduction, a frozen semantic to text mapper provides sequence level supervision over the replanned semantic trajectory. Experiments on timbre, emotion, and joint conversion demonstrate effective target attribute adaptation while preserving linguistic content. Under joint conversion, RePlan-VC achieves 7.68% WER, 0.5046 SECS, 0.6070 EmoSim, and 52.5% EmoAcc.

1 INTRODUCTION

Controllable speech generation aims to synthesize natural speech with desired attributes (Sisman et al., 2020; Wang et al., 2018). Voice conversion (VC) is an important branch of this field, modifying attributes such as speaker identity, emotion, or expressive style while preserving linguistic content (Bargum et al., 2023; Zhang et al., 2019; Yu et al., 2026).

As shown in Fig. 1, many VC approaches reuse source content representations and combine them with target speaker or style conditions (Qian et al., 2019; Wu et al., 2020; Tang et al., 2022). These representations provide a stable linguistic anchor while retaining source specific state trajectories and temporal structures, which can restrict target-side adaptation of duration, rhythm, and expressive realization. Other methods explicitly model target-side factors such as F0, duration, and prosody (Wang et al., 2021; Hwang et al., 2022; Choi et al., 2023). More recently, advances in self-supervised representations (Baevski et al., 2020; Hsu et al., 2021), neural audio modeling (Défossez et al., 2022; Zeghidour et al., 2021; Li et al., 2026a), and speech language models (Wang et al., 2023b;a) have extended target-side generation to high dimensional semantic and acoustic states. This shift from source-side reuse to target-side generation raises two key questions: *how to coordinate the generation of multiple target states with explicit dependencies* and *how to preserve linguistic content while replanning semantic states*.

Target-Side Multi-State Generation and Coordination. To model the target-side factors, factor-specific approaches replan only selected non-semantic factors, such as duration, pitch, or intonation (Chen et al., 2023; Zuo et al., 2025; Lee et al., 2022; Wang et al., 2021; Neekhara et al., 2023; Yao et al., 2025). Sequence-generative approaches produce new acoustic, content-style, or semantic token sequences (Huang et al., 2023; Baade et al., 2024; Zhang et al., 2025b; Cai et al., 2025; Li et al., 2025; Mehta et al., 2025). While both directions expand target-side adaptation, factor-specific approaches retain source derived content states, potentially constraining adaptation beyond the explicitly replanned factors through source specific temporal and expressive information; sequence-

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

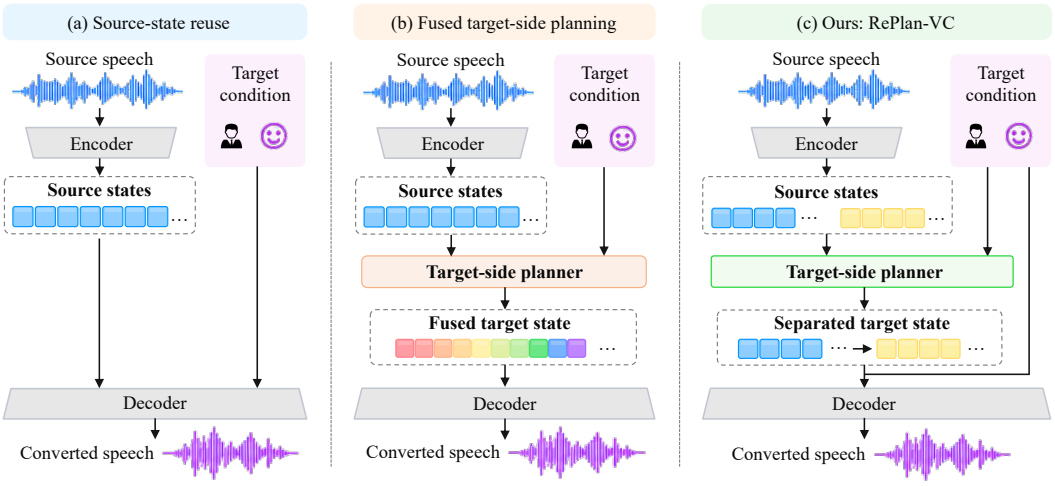


Figure 1: Comparison of source-side reuse, fused target-side planning, and RePlan-VC.

generative approaches do not necessarily distinguish and coordinate semantic and prosodic state generation. Consequently, the generative dependency between separately replanned semantic and prosodic trajectories remains insufficiently explored.

Linguistic Consistency in Semantic State Replanning. Semantic representations of speech primarily encode linguistic content but may retain speaker, prosodic, and temporal information (Qian et al., 2022; Zhang et al., 2024). Consequently, linguistically equivalent speech utterances need not share identical semantic state trajectories, as illustrated by $\mathrm{Text}(C_i) = \mathrm{Text}(C_j)$ while $C_i \neq C_j$. Existing approaches improve linguistic consistency at the representation level through CTC supervision, speaker disentanglement, or semantic alignment (Xinyuan et al., 2026; Hu et al., 2025; Akti et al., 2025; Wang et al., 2025; Xue et al., 2024). Other approaches promote linguistic consistency at the output level through ASR-based, perceptual, or reconstruction-consistency objectives (Mayrhofer et al., 2026; Liang et al., 2025; Yook et al., 2024; Ko et al., 2024; Li et al., 2023). These approaches constrain linguistic content either through representation learning or through the synthesized speech, rather than directly supervising regenerated semantic trajectories. Consequently, linguistic preservation during semantic replanning remains insufficiently addressed.

To address these challenges, we propose **RePlan-VC**, a target-side semantic and prosodic replanning framework for controllable voice conversion. Instead of directly reusing source semantic and prosodic states, RePlan-VC replans them into target-side trajectories under target speaker and emotion conditions. It explicitly models the generative dependency between semantic and prosodic states and directly constrains the linguistic content of the replanned semantic state trajectory. Therefore, RePlan-VC enables coordinated target-side state adaptation while preserving linguistic content without requiring source state reproduction.

Our main contributions are as follows:

- **A target-side multi-state replanning framework.** A unified replanning framework is introduced for controllable voice conversion, in which semantic and prosodic trajectories are regenerated under target speaker and emotion conditions before acoustic synthesis. This formulation enables target conditions to directly guide the replanning of multiple states and supports timbre, emotion, and joint conversion within a common framework.
- **Target-side hierarchical semantic and prosodic generation.** A hierarchical autoregressive planner is proposed to replan semantic and prosodic states while explicitly modeling their generative dependency. Semantic states are generated first and subsequently guide prosodic state prediction, with target conditions explicitly modulating the replanning process. This design enables coordinated multi-state generation under target conditions.
- **Sequence level linguistic consistency for semantic replanning.** A sequence level linguistic objective is introduced to constrain replanned semantic trajectories through a frozen semantic to text mapper. By supervising their decoded linguistic content rather than enforc-

ing token level or temporal matching with the source states, the objective promotes content preservation while retaining the freedom of semantic replanning.

2 Related Work

Factorized State Generation and Dependency Modeling. Recent speech language models increasingly organize generation through multiple semantic and acoustic representations. LM-VC (Wang et al., 2023b) retains source derived semantic tokens as linguistic conditions and generates target acoustic tokens through autoregressive coarse modeling followed by non-autoregressive fine modeling. Make-A-Voice (Huang et al., 2023) adopts a staged semantic to acoustic generation framework, autoregressively predicting target acoustic tokens from source semantic tokens under speaker prompting. Vevo (Zhang et al., 2025b) further introduces a content to content-style generation stage, autoregressively predicting target content-style tokens from source content tokens under style prompting. DisCo-Speech (Li et al., 2026b) factorizes speech into content, prosody, and timbre representations, but fuses content and prosody into a unified token stream for autoregressive generation. Overall, these approaches employ staged generation or fused target representations without explicitly modeling the generative dependency between separately replanned semantic and prosodic trajectories. In contrast, RePlan-VC separately replans semantic and prosodic states, with generated semantic states conditioning subsequent prosodic state prediction.

Linguistic Guidance for Speech Generation. Speech generation increasingly incorporates explicit linguistic supervision to improve content fidelity. CosyVoice (Du et al., 2024) introduces ASR-supervised semantic tokens to strengthen the alignment between speech representations and linguistic content. SemAlignVC (Mehta et al., 2025) combines CTC supervision with monotonic text-audio alignment to learn speaker independent semantic representations, reducing timbre leakage while preserving linguistic content. StarVC (Li et al., 2025) explicitly predicts text tokens before their corresponding acoustic tokens in a unified autoregressive framework, jointly optimizing transcription and acoustic generation to strengthen linguistic consistency. SemBridge (Xie et al., 2026) aligns continuous acoustic latents with semantic embeddings and directly supervises autoregressive LM states using discrete semantic token labels, strengthening linguistic modeling without introducing semantic token generation at inference. Overall, these methods incorporate linguistic information through supervised semantic representations, explicit text token prediction, or auxiliary semantic supervision. In contrast, RePlan-VC directly constrains the linguistic content of replanned semantic trajectories through a sequence level training objective.

3 RePlan-VC

3.1 Framework overview

To coordinate target-side state generation while preserving linguistic content, we propose RePlan-VC, a speech to speech voice conversion framework that replans semantic and prosodic trajectories under target speaker and emotion conditions before acoustic synthesis.

As shown in Fig. 2, the framework consists of three modules: **Factorized State Encoder (FSE)**, **Hierarchical Target State Planner (HTSP)**, and **Conditional Acoustic Generator (CAG)**. Given source speech x^s and a target reference utterance x^t_{ref}, FSE extracts discrete semantic and prosodic states (C_s, P_s) through $\mathcal{E}_{CP}$ and a global target speaker representation S_t through $\mathcal{E}_S$. HTSP then replans the source states into target-side semantic and prosodic trajectories $(\hat{C}_t, \hat{P}_t)$ under the target speaker and emotion conditions (S_t, E_t). Finally, CAG synthesizes the converted speech $\hat{x}^t$ from the replanned states $(\hat{C}_t, \hat{P}_t)$ and target speaker representation S_t, as shown in Eq. 1,

$$x^s \xrightarrow{\mathcal{E}_{CP}} (C_s, P_s) \xrightarrow[S_t, E_t]{\mathcal{F}_\theta} (\hat{C}_t, \hat{P}_t) \xrightarrow[S_t]{\mathcal{G}_\phi} \hat{x}^t. \tag{1}$$

The core of RePlan-VC is the Hierarchical Target State Planner, which connects the Factorized State Encoder and the Conditional Acoustic Generator. HTSP autoregressively generates target semantic and prosodic states and explicitly models their dependency. During training, a sequence level linguistic constraint promotes content preservation in replanned semantic trajectories.

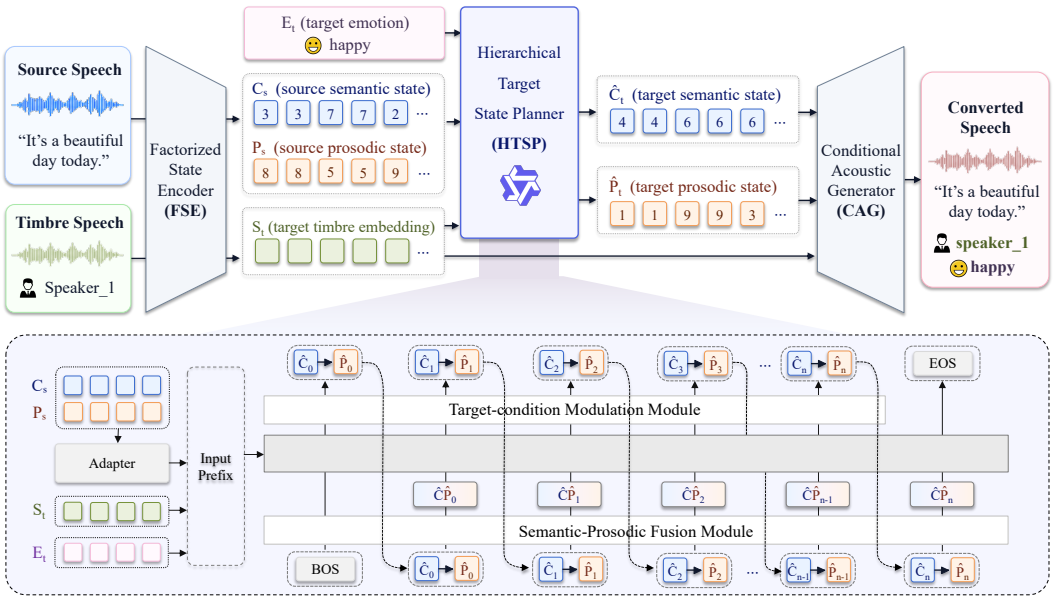


Figure 2: Overview of the RePlan-VC framework. **Top:** FSE extracts source semantic and prosodic states and the target speaker representation; HTSP replans target semantic and prosodic states under speaker and emotion conditions; and CAG synthesizes the converted speech. **Bottom:** HTSP takes the source states and target conditions as an input prefix and autoregressively generates target states through explicit target-condition modulation and hierarchical semantic-prosodic generation.

3.2 Hierarchical Target State Planner

Given source states (C_s, P_s) and target conditions (S_t, E_t), HTSP replans semantic and prosodic trajectories $(\hat{C}_t, \hat{P}_t)$ through three mechanisms: Target-condition prefixing, Target-conditioned modulation, and Hierarchical semantic-prosodic generation.

Target-Condition Prefixing. At the input, HTSP constructs an input prefix from the source semantic states, prosodic states and the target attributes. For each source frame i, the semantic and prosodic states are embedded using their respective codebooks and fused into a frame level source representation. Here, $\mathcal{E}_C$ and $\mathcal{E}_P$ denote the semantic and prosodic codebook embeddings, respectively, while $\Phi_s(\cdot)$ denotes the frame level fusion module that combines the two embeddings and projects them into the backbone hidden space. The resulting source representation is defined in Eq. 2,

$$f_i^s = \Phi_s([\mathcal{E}_C(c_i^s)\,;\,\mathcal{E}_P(p_i^s)])\,. \tag{2}$$

The target speaker and the target emotion are projected into condition embeddings S_t and E_t, respectively. These embeddings are combined with the source state sequence to construct the planning prefix. During autoregressive planning, this target-conditioned prefix is followed by a target beginning token and the target semantic-prosodic frames. $F^s = (f_1^s, \ldots, f_{T_s}^s)$ denotes the fused source-state sequence. $F_{<i}^t = (f_1^t, \ldots, f_{i-1}^t)$ denotes the representations of previously generated target frames. Under causal attention, the hidden state h_i captures the source states, target conditions, and preceding target state history for current state prediction, as shown in Eq. 3,

$$h_i = \mathcal{L}_\theta\big([F^s,\, S_t,\, E_t,\, \langle\text{bos}\rangle,\, F_{<i}^t]\big)\,. \tag{3}$$

Target-Conditioned Modulation Module. Although target speaker and emotion conditions are incorporated into the planning context through input prefixing, their influence on each prediction is mediated by the backbone attention mechanism and may be diluted by source state information and autoregressive history. To provide a more direct conditioning pathway, HTSP further modulates the hidden representation h_i immediately before state prediction. Specifically, the target conditions are mapped to shared FiLM style scale and shift parameters, bounded by a scaled hyperbolic tangent transformation to control the modulation strength. The resulting target conditioned representation is shared by the semantic and prosodic prediction branches, allowing both states to respond to target attributes while retaining their distinct generation roles.

Hierarchical Semantic-Prosodic Generation. HTSP generates target semantic and prosodic states sequentially within each frame. Given the target conditioned hidden representation h_i at frame i, the semantic branch first predicts the target semantic state. The predicted semantic state is then combined with h_i to guide the subsequent prediction of the target prosodic state. This semantic first design allows prosodic generation to build upon the semantic state predicted for the same frame. Let $\mathcal{H}_i$ denote the source state context, target conditions, and autoregressive target history available at frame i. $\hat{c}_i^t$ and $\hat{p}_i^t$ denote the replanned semantic and prosodic states at the current target frame, respectively. The hierarchical generation can be expressed as shown in Eq. 4,

$$p\big(\hat{c}_i^t, \hat{p}_i^t \mid \mathcal{H}_i\big) = p\big(\hat{c}_i^t \mid \mathcal{H}_i\big)\; p\big(\hat{p}_i^t \mid \mathcal{H}_i, \hat{c}_i^t\big)\,. \tag{4}$$

This factorization explicitly establishes a within frame semantic to prosodic dependency, allowing prosodic planning to be organized according to the semantic state determined for the same frame. Across frames, predictions are conditioned on the fused semantic prosodic representations of preceding target frames contained in $\mathcal{H}_i$. After generating the semantic-prosodic pair $(\hat{c}_i^t, \hat{p}_i^t)$, HTSP combines both states through the Semantic-Prosodic Fusion Module to obtain the target frame representation f_i^t. This representation is appended to the autoregressive history to predict the next frame. Meanwhile, a STOP predictor determines whether generation should terminate, allowing the target sequence length to be determined during replanning rather than fixed by the source length. In this way, HTSP preserves the functional distinction between semantic and prosodic states while explicitly coordinating their generation.

3.3 Sequence Level Linguistic Objective and Planner Optimization

The training objective of HTSP combines semantic and prosodic prediction, sequence termination, auxiliary emotion supervision, transition consistency, and anti-repetition losses. To preserve linguistic content during semantic replanning, we further introduce a sequence level linguistic objective. Progressive training with model predicted histories reduces the mismatch between training and autoregressive inference.

Sequence Level Linguistic Objective. We first train a semantic to text probe using semantic token sequences and their corresponding transcripts, which enables the probe to recover linguistic content from semantic trajectories. Then, the probe is frozen and used as a fixed linguistic mapper during HTSP training. The replanned semantic predictions are converted into differentiable semantic representations and passed through the frozen probe, with the source transcript serving as the supervision target. This directly constrains the linguistic content of the generated semantic trajectory. The corresponding loss is defined in Eq. 5,

$$\mathcal{L}_{\mathrm{LE}} = \mathrm{CTC}\left(\mathcal{G}_\psi\left(\left\{\sum_{k=1}^{K_C} \mathrm{Softmax}\left(\frac{z_i^C}{\tau_{\mathrm{LE}}}\right)_k \mathcal{E}_C(k)\right\}_{i=1}^{T_t}\right), Y^s\right). \tag{5}$$

z_i^C denotes the semantic prediction logits at target position i, τ_{LE} is the temperature applied to the semantic logits before computing the semantic token distribution, K_C is the semantic codebook size, and $\mathcal{E}_C(k)$ is the embedding of the k-th semantic token. The sequence enclosed by braces is the differentiable semantic trajectory of length T_t, $\mathcal{G}_\psi$ is the frozen probe, and Y^s is the transcript corresponding to the source utterance. By marginalizing over valid monotonic alignments, the loss promotes linguistic consistency without enforcing token level or temporal correspondence with the source semantic trajectory. The transcript and frozen semantic to text probe are used only during training, while inference remains fully speech to speech.

Joint Training Objective. The semantic and prosodic branches are optimized with transition weighted cross entropy losses $\mathcal{L}_C$ and $\mathcal{L}_P$, which assign larger weights to positions where the corresponding states change. This discourages unnecessary local repetition and places greater emphasis on state transitions. We further introduce transition consistency losses $\mathcal{L}_C^{\mathrm{tr}}$ and $\mathcal{L}_P^{\mathrm{tr}}$ to align predicted transition patterns with the ground truth trajectories. Margin based antirepetition losses $\mathcal{L}_C^{\mathrm{rep}}$ and $\mathcal{L}_P^{\mathrm{rep}}$ discourage persistence of the preceding state when a transition is required. The STOP predictor is trained with a positive class weighted binary cross entropy loss $\mathcal{L}_{\mathrm{stop}}$ to account for the strong imbalance between continuation and termination positions. When a target emotion is available, it is directly mapped to the emotion embedding; otherwise, an emotion predictor infers the emotion from

the source states. The auxiliary emotion classification loss $\mathcal{L}_{\text{emo}}$ is used only to train this predictor. The λ coefficients control the relative contributions of the corresponding objectives. Together with the sequence level linguistic loss, the overall training objective is given in Eq. 6,

$$\begin{aligned}\mathcal{L}_{\text{HTSP}} = \lambda_{\text{LE}}\mathcal{L}_{\text{LE}} + \lambda_C\mathcal{L}_C + \lambda_P\mathcal{L}_P + \lambda_C^{\text{tr}}\mathcal{L}_C^{\text{tr}} + \lambda_P^{\text{tr}}\mathcal{L}_P^{\text{tr}} \\ +\lambda_C^{\text{rep}}\mathcal{L}_C^{\text{rep}} + \lambda_P^{\text{rep}}\mathcal{L}_P^{\text{rep}} + \lambda_{\text{stop}}\mathcal{L}_{\text{stop}} + \lambda_{\text{emo}}\mathcal{L}_{\text{emo}}.\end{aligned} \tag{6}$$

To reduce exposure bias, HTSP progressively replaces a subset of ground truth semantic and prosodic pairs in the causal history. Rather than performing fully autoregressive roll-in, each candidate predicted pair is obtained under the preceding ground truth history, avoiding the cost of recursively generating the entire sequence. Ground truth semantic and prosodic states are jointly replaced by sampled predictions at the frame level, yielding a mixed history that gradually exposes the planner to its own predictions while preventing gradients from propagating through the sampled states.

3.4 Encoder and Generator

Factorized State Encoder extracts discrete semantic states C_s and prosodic states P_s from source speech, and a continuous global speaker representation S_t from target reference speech. The semantic branch is supervised by wav2vec2[1] representations to preserve linguistic content, the prosodic branch uses F_0, voicing, and energy supervision to model local prosodic variations, and the speaker branch adopts statistics pooling and an ArcFace objective to enhance speaker discriminability. Adversarial speaker disentanglement constraints are further applied to the semantic and prosodic branches, while acoustic reconstruction is used to preserve sufficient speech information. **Conditional Acoustic Generator** uses a conditional flow matching model to generate the target log-Mel spectrogram conditioned on the replanned semantic states $\hat{C}_t$, prosodic states $\hat{P}_t$, and target speaker representation S_t. The generated spectrogram is converted into a waveform using BigVGAN. Detailed architectures and training objectives of FSE and CAG are provided in Appendix A.

4 Experiments

4.1 Experimental Setup

Dataset and Training Details. We train RePlan-VC on a mixed corpus comprising 1,796,625 training pairs drawn from multiple speech datasets. For evaluation, we construct independent conversion sets from VCTK and ESD for the timbre, emotion, and joint conversion tasks. All evaluation speakers are excluded from the training set. All experiments are conducted on seven NVIDIA RTX 4090 GPUs. Detailed datasets and training settings are provided in Appendix B.

Baseline Systems. We compare RePlan-VC with five representative zero-shot VC methods, grouped by whether they explicitly generate target-side intermediate states. *Source-side reuse:* FACodec (Ju et al., 2024) and X-VC (Zheng et al., 2026) retain source derived representations and realize target adaptation through downstream transformation or acoustic generation. *Target-side generation:* GenVC (Cai et al., 2025), Vevo (Zhang et al., 2025b), and Vevo2 (Zhang et al., 2025a) explicitly generate new target side states before acoustic synthesis, but do not separately replan semantic and prosodic states with an explicit inter state generative dependency.

Evaluation Metrics. We evaluate all systems using objective and subjective metrics. Speech naturalness is measured by UTMOS[2], and linguistic content preservation by WER using Whisper[3]. Speaker similarity (SECS) is computed as the cosine similarity between SpeechBrain ECAPA-TDNN embeddings[4], while emotion similarity (EmoSim) is computed between emotion2vec embeddings and emotion accuracy (EmoAcc) using emotion2vec[5]. The reference utterance for speaker and emotion metrics is selected according to the conversion task. For subjective evaluation, 15 listeners participated in MOS tests assessing naturalness (N-MOS), speaker similarity (S-MOS), and emotion similarity (E-MOS) on a five-point MOS scale.

[1] https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft

[2] https://github.com/tarepan/SpeechMOS/tree/v1.2.0

[3] https://huggingface.co/openai/whisper-large-v3

[4] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[5] https://modelscope.cn/models/iic/emotion2vec_plus_base

Table 1: Comparison under timbre, emotion, and joint conversion. Best and second best results are shown in **bold** and underlined, respectively.

Method	UTMOS ↑	WER (%) ↓	SECS ↑	EmoSim ↑	EmoAcc (%) ↑	N-MOS ↑	S-MOS ↑	E-MOS ↑
(a) Timbre Conversion								
FACodec	3.3222	7.51	0.3508	**0.7335**	77.0	4.47	4.47	4.27
X-VC	3.5841	7.79	**0.5143**	0.6611	68.0	4.33	4.27	4.07
GenVC	3.2993	8.48	0.2409	0.6201	64.0	4.40	3.87	4.27
Vevo	**3.8154**	**7.34**	0.4601	0.7237	**78.0**	4.40	4.07	3.93
Vevo2	3.6754	10.68	0.4668	0.6704	72.0	4.00	3.93	3.67
Ours	3.4629	8.34	0.4790	0.6616	69.5	**4.80**	**4.67**	**4.53**
(b) Emotion Conversion								
Vevo	3.6582	13.71	**0.6888**	0.1034	9.0	**4.47**	**4.60**	3.60
Vevo2	**3.7732**	**9.04**	0.6858	0.2346	18.0	4.27	4.40	3.67
Ours	3.2289	10.95	0.4732	**0.5781**	**50.0**	4.33	3.93	**4.07**
(c) Joint Timbre–Emotion Conversion								
Vevo	3.7641	9.79	0.4808	0.5442	48.0	4.40	4.20	4.07
Vevo2	**3.8945**	8.33	**0.5293**	0.5672	52.0	4.53	4.27	3.93
Ours	3.4428	**7.68**	0.5046	**0.6070**	**52.5**	**4.73**	**4.47**	**4.53**

4.2 Overall Performance Comparison

We evaluate all systems under three conversion settings: **Timbre conversion**, which changes only speaker identity; **Emotion conversion**, which modifies only emotional expression; and **Joint Timbre-Emotion conversion**, which changes both attributes simultaneously. These settings evaluate speaker conversion, emotion control, and coordinated multi attribute conversion, respectively, as shown in Table 1. The evaluation references depend on the conversion setting. For timbre conversion, SECS is computed against the target speaker, while EmoSim and EmoAcc measure source emotion preservation. For emotion conversion, SECS measures source speaker preservation, whereas the emotion metrics are evaluated against the target emotion. For joint conversion, both speaker and emotion metrics are evaluated against their corresponding targets.

Timbre Conversion. RePlan-VC achieves the second highest SECS (0.4790), behind X-VC (0.5143), while maintaining a low WER of 8.34%. This indicates that regenerating both semantic and prosodic states does not prevent effective target speaker adaptation or substantially disrupt linguistic content. However, the source emotion should remain unchanged for timbre conversion. RePlan-VC obtains lower EmoSim (0.6616) and EmoAcc (69.5%) than several baselines, indicating that source emotion preservation remains imperfect. One contributing factor is source emotion prediction error: replacing the predicted label with the ground truth label improves EmoSim from 0.6616 to 0.7049 and EmoAcc from 69.5% to 74.5%. Despite this, RePlan-VC achieves the highest subjective N-MOS (4.80), S-MOS (4.67), and E-MOS (4.53), indicating favorable perceptual quality, speaker similarity, and emotion preservation.

Emotion Conversion. Emotion conversion requires modifying emotional expression while preserving speaker identity and linguistic content. RePlan-VC achieves substantially higher EmoSim and EmoAcc (0.5781/50.0%) than Vevo2 (0.2346/18.0%) and Vevo (0.1034/9.0%), showing stronger adaptation to the target emotion. Its WER remains moderate at 10.95%, indicating that the emotion driven state replanning largely preserves linguistic content. However, the lower SECS (0.4732) shows that source speaker identity is not yet fully preserved when the semantic and prosodic trajectories are reorganized toward the target emotion. RePlan-VC also obtains the highest E-MOS (4.07), while its lower S-MOS (3.93) is consistent with the objective speaker similarity gap. The lower UTMOS (3.2289) further indicates room for improvement in acoustic quality.

Joint Timbre-Emotion Conversion. Joint conversion requires adaptation of speaker identity and emotion while preserving linguistic content. RePlan-VC achieves the lowest WER (7.68%), the highest EmoSim (0.6070), and the highest EmoAcc (52.5%), while obtaining the second highest SECS (0.5046), close to Vevo2 (0.5293). These results indicate that separately replanned semantic and prosodic states can support coordinated adaptation to both target attributes without substantial linguistic degradation. The subjective results favor RePlan-VC, which achieves the highest N-MOS

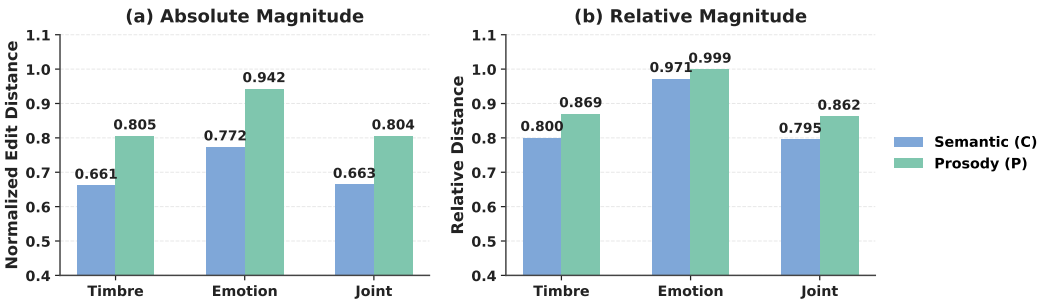


Figure 3: Target state replanning magnitude. (a) Absolute semantic and prosodic magnitudes. (b) Relative semantic and prosodic magnitudes.

(4.73), S-MOS (4.47), and E-MOS (4.53) under joint conversion. Although its UTMOS (3.4428) is lower than those of Vevo and Vevo2, this gap is not reflected in the subjective naturalness scores.

Overall. Across the three settings, RePlan-VC maintains effective speaker conversion under timbre conversion, improves target emotion adaptation under emotion conversion, and supports speaker and emotion control under joint conversion. The results also reveal task dependent attribute preservation errors: source emotion is not fully retained during timbre conversion, while source speaker identity degrades during emotion conversion. These effects suggest that further reducing cross attribute interference remains important. Subjective evaluation nevertheless shows favorable perceptual performance, particularly under joint conversion, while the gap between UTMOS and N-MOS also suggests that objective acoustic quality estimates do not fully reflect perceived naturalness.

4.3 Target-State Replanning Analysis

To investigate target state replanning, we measure the deviation of generated semantic and prosodic trajectories $(\hat{C}_t, \hat{P}_t)$ from their source counterparts (C_s, P_s) across the three conversion settings. Since state sequences may differ in length, we adopt normalized edit distance and further compare the replanning magnitude with the source and ground truth state difference. D quantifies sequence level state deviation, while R_C and R_P measure the replanning magnitude relative to the corresponding source and ground truth difference, as shown in Eq. 7,

$$D(A,B) = \frac{\mathrm{EditDist}(A,B)}{\max(|A|,|B|)}, \qquad R_C = \frac{D(\hat{C}_t, C_s)}{D(C_s, C_{gt})}, \qquad R_P = \frac{D(\hat{P}_t, P_s)}{D(P_s, P_{gt})}. \tag{7}$$

As shown in Fig. 3, both semantic and prosodic trajectories deviate substantially from their source states, confirming that HTSP performs target state replanning rather than simple source state reuse. Emotion conversion produces the largest changes, with $D(\hat{C}_t, C_s) = 0.772$ and $D(\hat{P}_t, P_s) = 0.942$. The corresponding $R_C = 0.971$ and $R_P = 0.999$ indicate that HTSP replans the states by nearly the same magnitude as the source and ground truth variation, rather than making only minor modifications to the source trajectory. Timbre and joint conversion exhibit similar relative replanning magnitudes, suggesting that HTSP can incorporate additional emotion control within a comparable degree of replanning without increasing the overall deviation from the source. We further compare the replanned and source states against their ground truth trajectories, with detailed results in Appendix C. These results show that substantial target-side replanning does not require reproduction of a unique ground-truth trajectory; under timbre and joint conversion, replanning can additionally move the states closer to the corresponding target realizations.

4.4 Linguistic Consistency Analysis

Speech derived semantic states are not unique, so the state deviations do not necessarily imply linguistic drift. We therefore examine whether semantic state deviation is associated with reduced linguistic decodability or degradation in the synthesized speech. Specifically, the source, ground truth, and replanned semantic trajectories C_s, C_{gt}, and $\hat{C}_t$ are decoded using the frozen probe and compared with the source transcript. The resulting word level error is denoted as A2T-Err. We further

Table 2: Linguistic consistency under semantic replanning.

Setting	Src-A2T-Err (%)	GT-A2T-Err (%)	Replanned-A2T-Err (%)	ρ_{probe}	ρ_{audio}
Timbre	9.79	14.57	25.29	0.322	0.037
Emotion	6.18	10.25	39.75	0.393	0.064
Joint	10.75	16.20	27.80	0.366	-0.024

Table 3: Ablation study under timbre conversion.

Method	UTMOS ↑	WER (%) ↓	SECS ↑	EmoSim ↑	EmoAcc (%) ↑
w/o Replanning	3.4293	10.31	0.3353	**0.8060**	**83.0**
w/o Modulation	3.2402	37.15	0.4587	0.6393	72.5
w/o Linguistic Constraint	3.4144	39.97	0.4644	0.5662	57.0
Parallel	3.3993	37.15	0.4462	0.6661	71.0
Full	**3.4629**	**8.34**	**0.4790**	0.6616	69.5

compute Spearman correlations between $D(\hat{C}_t, C_s)$ and changes in A2T-Err and WER, denoted as ρ_{probe} and ρ_{audio}, respectively. As shown in Table 2, replanned semantic states exhibit higher A2T-Err than source and ground truth states, with the largest increase under emotion conversion. The positive ρ_{probe} values indicate that larger semantic state deviations are associated with lower decodability under the frozen probe. However, ρ_{audio} remains close to zero (0.037, 0.064, and -0.024), indicating no clear monotonic association between larger state deviations and increased linguistic errors. Therefore, these results support our premise that linguistic content can be preserved without requiring the replanned semantic trajectory to reproduce the source state sequence.

4.5 Ablation Study

We evaluate four key components of the proposed framework. **w/o Replanning** bypasses the planner and feeds (C_s, P_s, S_t) directly into CAG. **w/o Modulation** retains the planner but removes explicit target-condition modulation. **w/o Linguistic Constraint** removes the sequence level linguistic objective. **Parallel** removes the current-frame semantic to prosodic dependency and predicts both states from the same autoregressive context. The **Full** model includes all components.

Table 3 reports the ablation results under timbre conversion, with additional results for emotion and joint conversion provided in Appendix D. Under timbre conversion, removing target state replanning causes the largest drop in speaker similarity, with SECS decreasing from 0.4790 to 0.3353, while WER rises moderately from 8.34% to 10.31%. This suggests that replanning mainly supports target condition adaptation. In contrast, source state reuse yields much higher EmoSim (0.8060) and EmoAcc (83.0%), as retaining source semantic and prosodic trajectories better preserves the original expression. Overall, replanning is important for target speaker adaptation, while source emotion preservation remains to be improved. Removing the linguistic constraint raises WER to 39.97% while retaining a relatively high SECS of 0.4644, indicating that it mainly preserves linguistic content rather than target speaker adaptation. Removing target condition modulation increases WER to 37.15% and decreases SECS to 0.4587, showing that prefix conditioning is insufficient for reliable state prediction. Replacing hierarchical semantic prosodic generation with parallel prediction raises WER to 37.15% and lowers SECS to 0.4462, supporting the importance of explicit semantic to prosodic dependency. Overall, the full model achieves the lowest WER and highest SECS and UTMOS, showing that these components play complementary roles in target state replanning.

5 Conclusion

We presented RePlan-VC, a target-side semantic and prosodic replanning framework for controllable voice conversion. RePlan-VC combines hierarchical semantic prosodic generation, target-condition modulation, and a sequence level linguistic objective to coordinate target state replanning while preserving linguistic content. Experiments on timbre, emotion, and joint conversion demonstrate effective target attribute adaptation across diverse controllable conversion settings and support explicit multi-state replanning without direct source state reuse. Future work will focus on improving naturalness and reducing cross attribute interference.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research methodology and experimental design, implementation and debugging of research code, interpretation of experimental results, and translation or refinement of technical writing. We also used generative AI tools for literature search and summarization, brainstorming, and manuscript editing. We did not use generative AI to generate synthetic experimental data, formulate or prove mathematical claims, or perform qualitative or thematic data analysis.

All AI-assisted content was reviewed and verified by the authors. Research ideas and related-work suggestions were checked against the original literature, generated or modified code was tested before use, and all reported experimental results were obtained from our own experimental pipelines. The authors take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

No specific ethical concerns were identified in this work.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproducibility of our results. Appendix A provides additional architectural and training objective details for the factorized state encoder and conditional acoustic generator. Appendix B describes the datasets, training-pair construction, training configurations, and inference settings.

REFERENCES

Seymanur Akti, Tuan Nam Nguyen, and Alexander Waibel. Towards better disentanglement in non-autoregressive zero-shot expressive voice conversion. *arXiv preprint arXiv:2506.04013*, 2025.

Alan Baade, Puyuan Peng, and David Harwath. Neural codec language models for disentangled and textless voice conversion. ISCA, 2024.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

Anders R Bargum, Stefania Serafin, and Cumhur Erkut. Reimagining speech: A scoping review of deep learning-powered voice conversion. *arXiv preprint arXiv:2311.08104*, 2023.

Zexin Cai, Henry Li Xinyuan, Ashi Garg, Leibny Paola García-Perera, Kevin Duh, Sanjeev Khudanpur, Matthew Wiesner, and Nicholas Andrews. Genvc: Self-supervised zero-shot voice conversion. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2025.

Li-Wei Chen, Shinji Watanabe, and Alexander Rudnicky. A unified one-shot prosody and speaker conversion system with self-supervised discrete speech units. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. *arXiv preprint arXiv:2311.04693*, 2023.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Desheng Hu, Yang Xiang, Jian Lu, Xinhui Hu, and Xinkang Xu. Speaker normalization and content restoration for zero-shot voice conversion with attention-enhanced discriminator. In *Proc. Interspeech 2025*, pp. 1403–1407, 2025.

Rongjie Huang, Chunlei Zhang, Yongqi Wang, Dongchao Yang, Luping Liu, Zhenhui Ye, Ziyue Jiang, Chao Weng, Zhou Zhao, and Dong Yu. Make-a-voice: Unified voice synthesis with discrete representation. *arXiv preprint arXiv:2305.19269*, 2023.

In-Seop Hwang, Sang-Hoon Lee, and Seong-Whan Lee. StyleVC: Non-parallel voice conversion with adversarial style generalization. In *26th International Conference on Pattern Recognition*, pp. 23–30, 2022.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 22605–22623, 2024.

Kyungdeuk Ko, Donghyeon Kim, Kyungseok Oh, and Hanseok Ko. Wavevc: Speech and fundamental frequency consistent raw audio voice conversion: K. ko et al. *Neural Processing Letters*, 56(3):166, 2024.

Sang-Hoon Lee, Hyeong-Rae Noh, Woo-Jeoung Nam, and Seong-Whan Lee. Duration controllable voice conversion via phoneme-based information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1173–1183, 2022.

Dayong Li, Xian Li, and Xiaofei Li. Dvqvc: An unsupervised zero-shot voice conversion framework. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Fengjin Li, Jie Wang, Yadong Niu, Yongqing Wang, Meng Meng, Jian Luan, and Zhiyong Wu. StarVC: A unified auto-regressive framework for joint text and speech generation in voice conversion. *arXiv preprint arXiv:2506.02414*, 2025.

Hui Li, Yangfan Gao, Junlin Shang, Changhao Jiang, Tao Gui, Qi Zhang, and Xuanjing Huang. Entanglecodec: A unified discrete audio tokenizer via semantic-acoustic entanglement. *arXiv preprint arXiv:2606.02739*, 2026a.

Tao Li, Wenshuo Ge, Zhichao Wang, Zihao Cui, Yong Ma, Yingying Gao, Chao Deng, Shilei Zhang, and Junlan Feng. Disco_speech: Controllable zero-shot speech generation with a disentangled speech codec. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18917–18931, 2026b.

Ziqi Liang, Xulong Zhang, Chang Liu, Xiaoyang Qu, Weifeng Zhao, and Jianzong Wang. Cycleflow: Leveraging cycle consistency in flow matching for speaker style adaptation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Benedikt Mayrhofer, Franz Pernkopf, Philipp Aichinger, and Martin Hagmüller. Lightweight and perceptually-guided voice conversion for electro-laryngeal speech. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17377–17381. IEEE, 2026.

Shivam Mehta, Yingru Liu, Zhenyu Tang, Kainan Peng, Vimal Manohar, Shun Zhang, Mike Seltzer, Qing He, and Mingbo Ma. SemAlignVC: Enhancing zero-shot timbre conversion using semantic alignment. *arXiv preprint arXiv:2507.09070*, 2025.

Paarth Neekhara, Shehzeen Hussain, Rafael Valle, Boris Ginsburg, Rishabh Ranjan, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley. Selfvc: Voice conversion with iterative refinement using self transformations. *arXiv preprint arXiv:2310.09653*, 2023.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. AutoVC: Zero-shot voice style transfer with only autoencoder loss. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5210–5219, 2019.

Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. ContentVec: An improved self-supervised speech representation by disentangling speakers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 18003–18017, 2022.

Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM transactions on audio, speech, and language processing*, 29:132–157, 2020.

Huaizhen Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. Avqvc: One-shot voice conversion by vector quantization with applying contrastive learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4613–4617. IEEE, 2022.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023a.

Disong Wang, Songxiang Liu, Lifa Sun, Xixin Wu, Xunying Liu, and Helen Meng. Learning explicit prosody models and deep speaker embeddings for atypical voice conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

Wenyu Wang, Yiquan Zhou, Jihua Zhu, Hongwu Ding, Jiacheng Xu, and Shihao Li. Avenet: Disentangling features by approximating average features for voice conversion. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, pp. 5180–5189. PMLR, 2018.

Zhichao Wang, Yuanzhe Chen, Lei Xie, Qiao Tian, and Yuping Wang. Lm-vc: Zero-shot voice conversion via speech generation based on language models. *IEEE Signal Processing Letters*, 30: 1157–1161, 2023b.

Da-Yi Wu, Yen-Hao Chen, and Hung-Yi Lee. VQVC+: One-shot voice conversion by vector quantization and U-Net architecture. *arXiv preprint arXiv:2006.04154*, 2020.

Hanke Xie, Haopeng Lin, Jiale Qian, Dake Guo, Yuepeng Jiang, Zhichao Wang, Wenxiao Cao, Jingbin Hu, Guobin Ma, Wenhao Li, Huakang Chen, Chengyou Wang, Ming Tao, Zhonghua Fu, Lei Xie, and Xinsheng Wang. SemBridge: Semantic token anchoring for continuous-latent autoregressive speech generation. *arXiv preprint arXiv:2608.07462*, 2026.

Henry Li Xinyuan, Zexin Cai, Lin Zhang, Leibny Paola García-Perera, Berrak Sisman, Sanjeev Khudanpur, Nicholas Andrews, and Matthew Wiesner. Universal speech content factorization. *arXiv preprint arXiv:2603.08977*, 2026.

Yuhang Xue, Ning Chen, Yixin Luo, Hongqing Zhu, and Zhiying Zhu. Clessr-vc: Contrastive learning enhanced self-supervised representations for one-shot voice conversion. *Speech Communication*, 165:103139, 2024.

Jixun Yao, Yang Yuguang, Yu Pan, Ziqian Ning, Jianhao Ye, Hongbin Zhou, and Lei Xie. Stablevc: Style controllable zero-shot voice conversion with conditional flow matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 25669–25677, 2025.

Dongsuk Yook, Geonhee Han, Hyung-Pil Chang, and In-Chul Yoo. Cyclediffusion: Voice conversion using cycle-consistent diffusion models. *Applied Sciences*, 14(20):9595, 2024.

Xinyue Yu, Youqing Fang, Pingyu Wu, Guoyang Ye, Wenbo Zhou, Weiming Zhang, and Song Xiao. MF-Speech: Achieving fine-grained and compositional control in speech generation via factor disentanglement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17966–17974, 2026.

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, 2021.

Mingyang Zhang, Xin Wang, Fuming Fang, Haizhou Li, and Junichi Yamagishi. Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet. *arXiv preprint arXiv:1903.12389*, 2019.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified speech tokenizer for speech language models. In *International Conference on Learning Representations*, 2024.

Xueyao Zhang, Ju-Nan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu. Vevo2: A unified and controllable framework for speech and singing voice generation. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2262–2275, 2025a.

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

Qixi Zheng, Yuxiang Zhao, Tian-Rui Wang, Wenxi Chen, Ke Xu, Yi-Kang Li, Qinyu Chen, Xipeng Qiu, Kai Yu, and Xie Chen. X-vc: Zero-shot streaming voice conversion in codec space. *ArXiv*, abs/2604.12456, 2026.

Jialong Zuo, Shengpeng Ji, Minghui Fang, Mingze Li, Ziyue Jiang, Xize Cheng, Xiaoda Yang, Chen Feiyang, Xinyu Duan, and Zhou Zhao. Rhythm controllable and efficient zero-shot voice conversion via shortcut flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16203–16217, 2025.

A Encoder And Generator Details

This section provides additional architectural and training objective details for the Factorized State Encoder (FSE) and Conditional Acoustic Generator (CAG). Numerical hyperparameters and training schedules are reported in Appendix B.

A.1 Factorized State Encoder

Acoustic Backbone. FSE decomposes speech into discrete semantic and prosodic trajectories and a continuous global speaker representation. A pretrained SEANet encoder first converts the input waveform into a frame-level acoustic latent sequence. The encoder remains frozen during representation learning, while separate semantic, prosodic, and speaker branches are optimized to extract their corresponding factors.

Semantic and Prosodic Representations. The semantic branch applies temporal convolutional encoding followed by vector quantization, producing the discrete semantic trajectory $C = (c_1, \ldots, c_T)$ and its quantized features Q^C. A frozen speech-recognition teacher provides frame-level pseudo-phoneme labels, which are temporally aligned with the semantic representation. H_C denotes the semantic prediction head and Y^{ph} represents the aligned pseudo-phoneme labels. The semantic objective is defined as shown in Eq. 8,

$$\mathcal{L}_C = \mathrm{CE}\left(H_C(Q^C), Y^{ph}\right). \tag{8}$$

The prosodic branch independently encodes and quantizes the acoustic latent sequence, producing $P = (p_1, \ldots, p_T)$ and Q^P. It is supervised by log-F0, voiced/unvoiced information, and relative log-energy. The F0 and energy terms use smooth-L1 losses, while the voiced/unvoiced term uses binary cross entropy, as shown in Eq. 9,

$$\mathcal{L}_P = \lambda_{F0}\mathcal{L}_{F0} + \lambda_{UV}\mathcal{L}_{UV} + \lambda_E\mathcal{L}_E. \tag{9}$$

Speaker Representation and Disentanglement. The speaker branch extracts a continuous utterance level representation through temporal residual blocks and statistics pooling. The pooled mean and standard deviation are projected by an MLP and ℓ_2-normalized to obtain the global speaker embedding. Speaker discriminability is supervised using an ArcFace classifier. To reduce speaker leakage into the semantic and prosodic streams, separate speaker adversaries are attached to Q^C and Q^P through gradient reversal layers. The adversaries are optimized to predict speaker identity, whereas the reversed gradients encourage the two state encoders to suppress speaker discriminative information.

Reconstruction and Training Objective. The quantized semantic and prosodic features are projected and combined by a temporal residual mapper. The global speaker embedding is injected through speaker-conditioned FiLM residual blocks to construct a new acoustic latent, which is decoded by a newly initialized SEANet decoder. The reconstruction objective combines waveform domain and multi-resolution Mel-spectrogram losses, as shown in Eq. 10,

$$\mathcal{L}_{\text{rec}} = \lambda_{\text{wav}}\|\hat{x} - x\|_1 + \lambda_{\text{mel}}\mathcal{L}_{\text{mel}}. \tag{10}$$

The complete representation-learning objective combines semantic, prosodic, speaker, vector-quantization, adversarial, and reconstruction losses. The speaker representation remains continuous and is not vector quantized.

A.2 Conditional Acoustic Generator

CAG synthesizes the converted speech from the replanned semantic and prosodic states and the target speaker representation. Emotion is not supplied as an additional direct condition to this module; its effect is represented by the target-side states generated by HTSP.

Condition Construction. The predicted semantic and prosodic token IDs are converted into their corresponding frozen codebook embeddings, Q^C and Q^P, and temporally aligned with the Mel-spectrogram resolution. Together with the continuous target speaker representation S^t, they form the acoustic condition, as shown in Eq. 11,

$$\mathcal{K} = (Q^C, Q^P, S^t). \tag{11}$$

CAG employs a DiT-based conditional flow-matching decoder to predict the velocity field in normalized log-Mel space. Semantic and prosodic conditions are projected independently and injected at the decoder input and throughout the DiT blocks. The target speaker embedding is projected into the hidden space and additionally used to condition adaptive normalization within the decoder. Flow time is encoded using a sinusoidal embedding.

Flow Matching Objective. Let x_1 denote the normalized target log-Mel spectrogram and $x_0 \sim \mathcal{N}(0, I)$ denote Gaussian noise. For a sampled flow time $t \sim \mathcal{U}(0, 1)$, the interpolation and target velocity are defined as shown in Eq. 12,

$$x_t = (1 - t)x_0 + tx_1, \qquad v^* = x_1 - x_0. \tag{12}$$

The decoder is trained with the conditional flow matching objective as shown in Eq. 13,

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{x_0, x_1, t}\left[\|v_\theta(x_t, t \mid \mathcal{K}) - v^*\|_2^2\right]. \tag{13}$$

Finally, the generated log-Mel spectrogram is converted into the waveform using a pretrained BigVGAN vocoder.

Table 4: Composition of the training corpus.

Dataset	Utterances	Ratio (%)
Common Voice	556,301	70.31
LibriTTS	68,046	8.60
AISHELL-3	50,521	6.39
VCTK	35,209	4.45
ESD	28,000	3.54
MEAD	25,362	3.21
Effect AI	8,787	1.11
IEMOCAP	8,027	1.01
MELD	6,636	0.84
CSEMOTIONS	3,324	0.42
CASIA	960	0.12
Total	791,173	100.00

B Experimental Details

This section provides the dataset composition, preprocessing procedures, training configurations, and inference settings used in our experiments.

Training Data. We construct the training corpus from 11 speech datasets, comprising 791,173 utterances in total. The dataset composition is summarized in Table 4. All waveforms are converted to mono and resampled to 16 kHz for state extraction. The representation model is trained on 3-second segments, obtained by random cropping or zero-padding, while validation uses deterministic center crops.

Planner Pair Construction. We construct 1,796,625 planner training pairs from the multi corpus speech pool based on linguistic content, speaker identity, and emotional expression. Transcripts are normalized for content matching, and emotion labels are obtained using emotion2vec. We consider seven emotion classes: angry, disgusted, fearful, happy, neutral, sad, and surprised. Utterances used as sources or acoustic targets must have a predicted emotion confidence of at least 0.80 and a margin of at least 0.10.

We construct four types of training pairs to cover different target state adaptation requirements:

- **P1 (timbre conversion):** Source and target utterances have matched linguistic content and emotion but different speakers. No explicit target emotion label is provided, encouraging the planner to adapt speaker identity while preserving the source emotion.
- **P2 (identity preservation):** The source utterance also serves as the acoustic target, while a different utterance from the same speaker provides the speaker reference. Approximately 5% of the reliable-emotion utterances are selected to provide identity preservation supervision. No explicit target emotion label is provided.
- **E1 (joint conversion):** Source and target utterances have matched linguistic content but different speakers and emotions. The target emotion is explicitly specified to supervise joint timbre-emotion adaptation.
- **E2 (emotion conversion):** Source and target utterances have matched linguistic content and the same speaker identity but different emotions. The target emotion is explicitly specified to supervise emotion adaptation while preserving speaker identity.

For all four pair types, the speaker reference is selected from the target speaker and must be a different utterance with different linguistic content from both the source and acoustic target. The reference is used only to extract the target speaker representation; its emotion label is not used as the target emotion condition. This prevents the reference from directly supplying the linguistic content of the acoustic target.

To reduce data imbalance, candidate pairs are dynamically downsampled based on emotion classes, source to target emotion transitions, and source and target speaker frequencies. Additional constraints limit the dominance of individual pair types and task modes without oversampling minority classes.

The resulting pairs are augmented with pre-extracted semantic and prosodic trajectories and target speaker embeddings to form the planner training records. Each record contains the source and target state sequences, speaker condition, emotion labels, and source transcript. The source transcript is used exclusively by the training time linguistic consistency objective and is never provided as sample specific input to the planner.

Evaluation Data. We construct three fixed evaluation sets from VCTK and ESD, containing 200 timbre conversion pairs, 100 emotion conversion pairs, and 200 joint conversion pairs. The same evaluation pairs are used across all applicable systems and ablation variants to ensure consistent comparisons.

Training Configuration. The representation model operates at 16 kHz with a 50-Hz state frame rate. The semantic and prosodic branches use separate vector-quantization codebooks, each containing 2,048 entries with 256-dimensional embeddings. The global speaker representation is also 256-dimensional. The pretrained SEANet encoder and semantic and pitch teachers remain frozen during representation learning. The model is trained with a global batch size of 126. Following the initial representation-learning stage, gradient-reversal refinement is applied to reduce speaker information in the semantic and prosodic states.

The planner uses Qwen2.5-1.5B as its backbone and is fine-tuned with LoRA of rank 16 and alpha 32. The backbone and newly introduced modules use learning rates of 5×10^{-5} and 1×10^{-4}, respectively. We use AdamW with a per-GPU batch size of 2 and gradient accumulation of 8, resulting in an effective global batch size of 112. Training is conducted for three epochs in bfloat16.

The sequence level linguistic objective is disabled during the first 2,000 optimizer steps. Its weight is then linearly increased to 0.10 over the following 2,000 steps and remains fixed thereafter. To reduce exposure bias, training initially uses ground truth autoregressive histories. From step 4,000, one contiguous 16-frame history segment per sample is replaced with detached model predictions.

The acoustic generator is trained using the frozen representation model and a DiT-based conditional flow matching decoder. The decoder has a hidden dimension of 768, 12 layers, and 12 attention heads. Training targets are 100-band log-Mel spectrograms extracted at 24 kHz, with an FFT size of 1,024 and a hop size of 256. We use AdamW with a learning rate of 2×10^{-4} and a global batch size of 126. Only the flow decoder is optimized during this stage. Emotion is not directly supplied to the acoustic generator; its effect is represented by the replanned semantic and prosodic states.

Inference Configuration. At inference, source semantic and prosodic states are extracted from the source speech, while the speaker embedding is obtained from the target reference speech. For emotion and joint conversion, the specified target emotion is supplied to the planner. For timbre conversion, the planner predicts the source emotion when no explicit emotion target is available.

The planner generates semantic and prosodic states autoregressively using greedy decoding. The minimum and maximum generation lengths are 10 and 600 frames, with an additional length constraint of 0.5-1.8 times the source state length. The STOP threshold is set to 0.5. The resulting target states and speaker embedding condition the acoustic generator. We solve the conditional flow ODE using the Heun method with 50 integration steps and synthesize the final waveform using the pretrained BigVGAN vocoder.

C Ground Truth Proximity Analysis

To further investigate target state replanning, we compare the replanned semantic and prosodic trajectories with their corresponding ground truth realizations. For each state, we compute the normalized edit distances between the source and ground truth trajectories and between the replanned and ground truth trajectories, using the distance function defined in Eq. **??**. We define the ground-truth proximity gain as the reduction in distance to the ground truth after replanning. A positive gain indicates that the replanned trajectory is closer to the ground truth than the source trajectory. We also report the percentage of samples with positive gains. The gain is formulated as Eq. 14,

$$\Delta_X^{\mathrm{GT}} = D(X^s, X^{gt}) - D(\hat{X}^t, X^{gt}), \quad X \in \{C, P\}. \tag{14}$$

As shown in Table 5, timbre and joint conversion exhibit modest improvements in ground truth proximity for both semantic and prosodic states, with the percentage of samples showing reduced

Table 5: Ground-truth proximity of replanned semantic and prosodic states under different conversion settings.

Setting	State	Src-GT	Pred-GT	Δ_X^{GT}	Improved (%)
Timbre	Semantic	0.8264	0.8003	+0.0261	61.5
	Prosody	0.9258	0.9021	+0.0237	61.5
Emotion	Semantic	0.7953	0.8176	-0.0223	43.0
	Prosody	0.9432	0.9442	-0.0010	46.0
Joint	Semantic	0.8350	0.8015	+0.0335	64.5
	Prosody	0.9335	0.9028	+0.0307	71.0

Table 6: Additional ablation results under emotion and joint timbre-emotion conversion.

Method	UTMOS ↑	WER (%) ↓	SECS ↑	EmoSim ↑	EmoAcc (%) ↑
Emotion Conversion					
w/o Modulation	3.2205	44.93	0.4536	0.5326	50.0
w/o Linguistic Constraint	3.3318	36.54	0.5347	0.1594	10.0
Parallel	3.2814	38.73	0.4583	0.5641	49.0
Full	3.2289	10.95	0.4732	0.5781	50.0
Joint Timbre-Emotion Conversion					
w/o Modulation	3.2957	45.55	0.4762	0.5785	47.0
w/o Linguistic Constraint	3.3490	40.79	0.5067	0.5814	51.5
Parallel	3.4309	38.70	0.5097	0.6348	53.5
Full	3.4428	7.68	0.5046	0.6070	52.5

distances ranging from 61.5% to 71.0% across states and settings. In contrast, emotion conversion produces the largest state deviations without corresponding improvements in ground truth proximity. This task dependent behavior reveals that the magnitude of state replanning is not directly coupled with token level agreement with a specific target trajectory. These observations support the distinction between target conditioned state replanning and ground truth trajectory reproduction: RePlan-VC can adapt intermediate states toward target attributes while allowing different semantic and prosodic realizations rather than enforcing a unique target state sequence.

D Additional Ablation Results

To complement the ablation study under timbre conversion in Section 4.5, we further evaluate the planner related variants under emotion and joint timbre emotion conversion. The w/o Replanning variant is not included because CAG does not directly accept an emotion condition; emotion adaptation is realized through the replanned target states.

As shown in Table 6, all ablated variants exhibit substantially higher WER than the full model under both emotion and joint conversion. Removing target-condition modulation increases WER from 10.95% to 44.93% under emotion conversion and from 7.68% to 45.55% under joint conversion, highlighting its role in reliable target state prediction. Removing the linguistic constraint additionally causes a pronounced degradation in emotion adaptation under emotion conversion, reducing EmoSim from 0.5781 to 0.1594 and EmoAcc from 50.0% to 10.0%, despite improved source-speaker similarity. Under joint conversion, parallel generation achieves slightly higher SECS, EmoSim, and EmoAcc than the full model, but increases WER from 7.68% to 38.70%. These results indicate that target attribute similarity alone does not guarantee linguistic preservation and support the complementary roles of explicit target conditioning, linguistic supervision, and hierarchical semantic prosodic generation in coordinated state replanning.