LEARNING TO ORGANIZE CONTINUOUS AUDIO REPRESENTATIONS FOR STREAMING GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming speech generation poses a representation challenge: acoustic latents must be compact and easy to predict incrementally, yet retain the detail needed to reconstruct natural speech and music. Systems that use an external semantic encoder must maintain a relationship between acoustic and semantic representations during generation. We propose a low-frame-rate continuous audio representation that incorporates semantic structure directly into reconstructable acoustic latents. The same representation serves as a generation target and supports autoregressive audio-history modeling, without requiring an external semantic encoder at inference. Block-causal processing makes the representation suitable for streaming, while soft semantic alignment provides a content prior without removing acoustic flexibility. We further organize channels into expandable ordered prefixes, allowing a single audio encoder–decoder to provide representations at different capacities and enabling the generator to adjust its learning priorities at a fixed capacity. Experiments on Seed-TTS Chinese speech generation and speech and music reconstruction show that this design offers a unified representation for streaming generation with semantic guidance, acoustic fidelity, and adjustable capacity.

1 INTRODUCTION

Continuous-latent speech synthesis predicts sequences of audio representations and decodes them into waveforms (Jia et al., 2025; Liu et al., 2024; Peng et al., 2025). These latents must preserve semantic and acoustic information while serving as targets that a generative model can predict. Semantic supervision can introduce content-related structure into these representations (Niu et al., 2026; Dai et al., 2026). During autoregressive synthesis, each new latent is conditioned on text and preceding audio history, so the representation also determines what information is available to the model as generation proceeds. High reconstruction fidelity alone, however, does not guarantee efficient learning of this conditional distribution within a finite training budget (Niu et al., 2026; Yao et al., 2025; Leng et al., 2025). Streaming synthesis imposes an additional restriction, requiring each segment to be produced before the full acoustic context is available (Lian et al., 2026; Rouard et al., 2026). Together, these demands motivate us to consider not only what information a representation retains, but how it is organized for generation.

Existing continuous audio systems introduce semantic structure by aligning features or incorporating semantic supervision into representation learning (Xie et al., 2026; Yan et al., 2025). Some systems feed distinct acoustic and semantic token streams to the autoregressive model, while others encode generated acoustic patches into semantic feedback (Peng et al., 2025; Lian et al., 2026). These designs show different ways semantics can support generation; when a separate semantic path remains active at inference, its output must be coordinated with the evolving audio history. At the same time, semantic targets emphasize content and need not preserve the acoustic variation required for faithful reconstruction; their geometry and statistics may also differ from those of acoustic representations (Chen et al., 2026; Dinkel et al., 2026; Cheng et al., 2026). This motivates a different design question: can a training-only semantic teacher shape a reconstructable acoustic latent that serves as a generation target and, after local aggregation, as autoregressive history, without a separate semantic encoder at inference?

A semantically structured and reconstructable latent may still be difficult to model in full. Generative models can require adaptation to a representation's dimensionality and statistics (Zheng et al., 2026),

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.

and visual studies show that greater capacity can improve reconstruction without improving generation under a fixed training budget (Yao et al., 2025). Ordered representations offer one response: compact prefixes can serve as usable targets, while longer prefixes add capacity. Nested Dropout and Matryoshka Representation Learning establish this principle in reconstruction and discriminative tasks (Rippel et al., 2014; Kusupati et al., 2022); STAR-VAE examines channel truncation for audio reconstruction and generation (Liu et al., 2026). We therefore ask how to train audio prefixes that remain useful for generation while allowing longer prefixes to improve speech and music reconstruction.

Our framework combines a low-rate semantic–acoustic interface with ordered-prefix learning. The representation operates at 10 Hz, and its two-frame blocks match the temporal units used by downstream generation. A frozen block-causal ASR teacher provides spherical semantic targets during codec training; soft directional alignment guides the acoustic latent while leaving waveform reconstruction to the acoustic path. We then train prefixes of different lengths with a shared encoder and decoder, using reconstruction and speech semantic prediction to organize content and acoustic detail. At inference, no external semantic encoder is needed: the selected acoustic prefix is the DiT prediction target and, after local aggregation, supplies the autoregressive audio-history condition. Downstream channel weighting provides an additional way to adjust learning priorities within a fixed prefix.

Our contributions are threefold:

- We present a low-rate, block-causal semantic–acoustic design in which a training-only ASR teacher guides a reconstructable acoustic latent. The same latent representation supports generation targets and autoregressive history without an external semantic encoder at inference.
- We organize acoustic information into expandable prefixes that share one encoder–decoder, and show how channel weighting can adjust generative learning priorities without retraining the codec.
- Experiments demonstrate effective speech generation from compact prefixes, improved speech and music reconstruction with longer prefixes, and better early content accuracy and UTMOS from tiered weighting at 96 channels.

2 Related Work

2.1 Semantic–Acoustic Representations

Semantic supervision aims to make continuous audio representations easier to model for generation without compromising the acoustic flexibility required for faithful reconstruction. Semantic-VAE and SemaVoice introduce content priors by aligning acoustic latents with pretrained speech features; Qwen-Audio-3.0 uses text-prediction supervision from a frozen language model (Niu et al., 2026; Wang et al., 2026; Dai et al., 2026). How these priors are incorporated affects the balance between semantic structure and acoustic fidelity. JMAS-VAE investigates this balance through different alignment objectives and loss weights, evaluating their effects on reconstruction and downstream performance (Cheng et al., 2026). SARA and DashengTokenizer complement frozen semantic features with trainable acoustic branches to capture additional detail (Chen et al., 2026; Dinkel et al., 2026). Streaming extends this design problem to temporal context. CALM incorporates semantic distillation into a causal VAE, and Ming-UniAudio integrates semantic learning into a streamable unified representation (Rouard et al., 2026; Yan et al., 2025).

At inference, VibeVoice uses acoustic and semantic token streams, while dots.tts feeds semantic features from generated patches to its language model. SemBridge applies semantic supervision to codec representations and autoregressive states during training (Peng et al., 2025; Lian et al., 2026; Xie et al., 2026). We ask whether a single reconstructable acoustic latent can carry semantic structure while serving both as the DiT prediction target and, after local aggregation, as autoregressive audio history.

2.2 Representations for Generation

Latent representations affect both reconstruction fidelity and the behavior of generative models trained on them. Prior work addresses this relationship through generator adaptation and representation learning. Feature alignment and architectural adaptations help generators use pretrained representations (Yu et al., 2025; Zheng et al., 2026), while representation learning objectives shape the latent space itself (Yao et al., 2025; Leng et al., 2025). These approaches highlight the need to evaluate representations through generation as well as reconstruction. This motivates a question of information organization: can a model use a compact prefix for generation while longer prefixes of the same representation retain additional acoustic detail?

Ordered representations offer one approach, but prior methods organize capacity along different axes. Nested Dropout and Matryoshka Representation Learning train nested representations, whereas FlexTok varies token-sequence length rather than per-frame channel width (Rippel et al., 2014; Kusupati et al., 2022; Bachmann et al., 2025). DC-AE1.5 studies channel-prefix reconstruction together with diffusion training (Chen et al., 2025). In audio, STAR-VAE regulates channel capacity and evaluates both reconstruction under channel truncation and downstream generation (Liu et al., 2026). Our work studies low-rate continuous-audio channel prefixes trained with a shared encoder–decoder, and tests tiered DiT loss weighting at a fixed 96-channel capacity.

3 Method

3.1 Overview

Codec learning proceeds in two stages after preparation of a block-causal ASR teacher with a spherical bottleneck. The teacher is trained with blocks corresponding to downstream generation units and then frozen. The first codec stage, illustrated in Figure 1, combines waveform reconstruction with soft directional alignment to its targets, introducing content-related structure while preserving acoustic flexibility. Building on this representation, the second stage trains channel prefixes of different lengths through reconstruction, auxiliary semantic prediction on speech, and group-specific scale constraints. The codec uses block-causal processing and limited decoder lookahead throughout. Semantic supervision shapes the latents during training but does not enter the waveform decoding path.

Once trained, the codec is frozen (Figure 2). Each downstream generative model selects a fixed-length prefix for both prediction and audio-history conditioning, and the shared decoder converts generated latents into waveforms. We next describe semantic–acoustic training (Section 3.2), ordered-prefix learning (Section 3.3), and downstream generation with selected prefixes and group-wise loss weighting (Section 3.4).

3.2 Semantically Anchored Acoustic Representation

We construct semantic targets with a controlled scale and temporal structure suited to blockwise generation. The ASR frontend subsamples input features to 10 Hz for a block-causal Transformer. A causal cross-frame module then reduces the channel dimension to 128 without changing the frame rate. Each bottleneck frame h_t is normalized as

$$s_t = \frac{h_t}{\sqrt{\frac{1}{C}\sum_{c=1}^{C} h_{t,c}^2 + \epsilon_s}}, \qquad C = 128.$$

A recognition head consumes s_t, and the ASR model is trained with the CTC objective (Graves et al., 2006). Normalization keeps the bottleneck norm approximately fixed, while recognition training encourages content information to be expressed through its direction. Once trained, the ASR model is frozen and supplies s_t as the target for codec training.

Faithful reconstruction also requires preserving acoustic variation, including timbre and prosody, that the semantic target may not fully capture. We therefore retain an independent acoustic path, $a = E(x)$ and $\hat{x} = D(a)$, without adding or concatenating teacher features to the decoder input. The acoustic encoder and decoder follow a MOSS-style Transformer architecture (Gong et al., 2026). Frame-rate studies in spoken dialogue motivate considering temporal granularity alongside

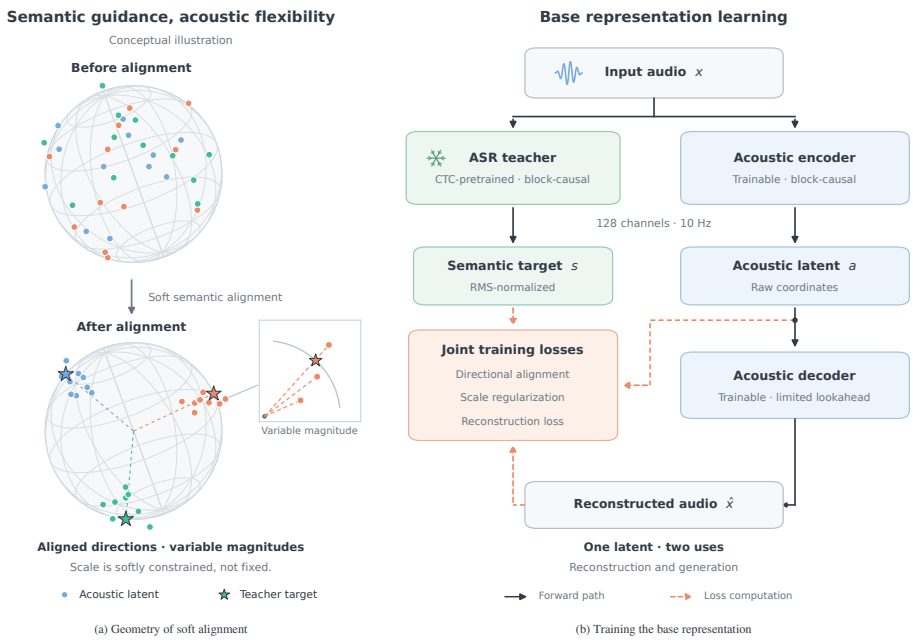


Figure 1: Base representation learning in TPAE. Left: conceptual illustration of soft semantic alignment with variable acoustic magnitudes; colors distinguish content-related groups and stars denote semantic targets. Right: a frozen block-causal ASR teacher supervises the acoustic representation. Normalization is used only for directional alignment; the decoder receives the original acoustic latent.

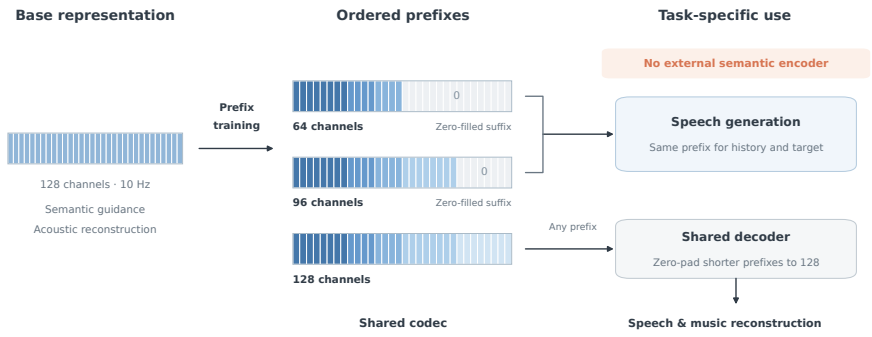


Figure 2: Ordered-prefix training and downstream use. Prefixes of the shared 128-channel, 10 Hz representation support reconstruction at different capacities. At inference, the selected prefix serves as both the DiT target and autoregressive audio history, without an external semantic encoder. Omitted channels are zero-padded only for decoding.

representation content (Ye et al., 2026). In our system, semantic and acoustic latents both operate at 10 Hz. The teacher Transformer and the lowest-rate acoustic encoder stage use two-frame blocks with six preceding blocks visible per layer. Each block spans 200 ms, matching the downstream DiT prediction unit. The remaining acoustic stages use finite history, and the decoder permits limited lookahead. This coordinates block-level context without requiring identical receptive fields.

We introduce content structure through soft directional alignment rather than requiring acoustic latents to reproduce the semantic targets exactly. Define the radius and normalized direction of a

vector v as $r(v)$ and $u(v) = v/r(v)$, respectively. The auxiliary objectives are

$$r(v) = \sqrt{\frac{1}{C}\sum_{c=1}^{C} v_c^2 + \epsilon_a^2}, \qquad u(v) = \frac{v}{r(v)},$$
$$\mathcal{L}_{\text{dir}} = \mathbb{E}_{b,t}\left[\frac{1}{C}\left\|u(a_{b,t}) - u(s_{b,t})\right\|_2^2\right],$$
$$\mathcal{L}_{\text{rad}} = \mathbb{E}_{b,t}\left[\left(\log r(a_{b,t})\right)^2\right],$$

where $\mathbb{E}_{b,t}$ averages over batch and temporal positions after length alignment. The acoustic latents are normalized only for the alignment loss; the decoder receives the unnormalized a. Acoustic information can consequently be expressed through both magnitude variation and directional deviations. The radius penalty discourages scale drift without fixing each frame to a sphere, allowing reconstruction and semantic supervision to jointly shape the latent space.

The first-stage objective combines these constraints with acoustic reconstruction:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{audio}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{rad}}\mathcal{L}_{\text{rad}},$$

where $\mathcal{L}_{\text{audio}}$ comprises multi-resolution mel reconstruction, adversarial, and feature-matching losses (Kumar et al., 2023). We jointly update the acoustic encoder and decoder, keep the teacher frozen, and optimize the discriminator with its own adversarial objective. Detailed configurations, numerical constants, and boundary handling are provided in the appendix. The resulting full-dimensional representation initializes the next stage, which learns content preservation and reconstruction across channel prefixes.

3.3 Learning Ordered Prefixes

The first stage establishes a full-dimensional representation with semantic structure and acoustic flexibility, but does not train shorter channel prefixes to function independently. We adopt nested-prefix training (Rippel et al., 2014; Kusupati et al., 2022) to support multiple capacities within this latent space. Given $a \in \mathbb{R}^{C\times T}$, with $C = 128$, the prefix operator is defined as

$$[P_K(a)]_{c,t} = \begin{cases} a_{c,t}, & c < K, \\ 0, & c \geq K, \end{cases} \qquad c = 0, \ldots, C-1.$$

The acoustic encoder and decoder are further trained under this masking scheme, shaping information allocation across a fixed channel order while retaining a shared decoding interface.

Reconstruction from sampled prefixes gives each selected capacity a direct training objective. For every example, a length K is drawn from a domain-dependent distribution $p_d(K)$ over a predefined set of sizes and held fixed throughout the sequence:

$$K \sim p_d(K), \qquad \hat{x}_K = D(P_K(E(x))).$$

The reconstructed audio is supervised using the audio reconstruction objective $\mathcal{L}_{\text{audio}}$ defined in Section 3.2. Speech training emphasizes compact prefixes while retaining longer ones for higher-capacity reconstruction; music and sound training predominantly uses the full representation. These sampling preferences reflect the intended capacity requirements without partitioning the channels into separate domain-specific representations. The supported sizes and sampling probabilities are specified in the appendix.

For speech, an auxiliary objective encourages content-related teacher features to remain predictable from the selected prefix. A shared block-causal semantic head H takes the same masked latent as the decoder and predicts the frozen teacher representation:

$$\mathcal{L}_{\text{sem}} = \mathbb{E}_{(b,t)\in\mathcal{V}_{\text{speech}}}\left[\left\|\nu(H(P_{K_b}(a_b))_t) - \nu(s_{b,t})\right\|_2^2\right].$$

Here, $\mathcal{V}_{\text{speech}}$ contains valid speech frames, and ν denotes numerically stabilized unit-length normalization across channels. This objective supervises what can be predicted from a prefix without requiring its acoustic coordinates to match the teacher directly. The acoustic channels can therefore

be reorganized under joint semantic and reconstruction supervision. The auxiliary head is used only during training, and its output does not enter the waveform decoder.

Prefix supervision governs information preservation, while a separate scale objective regulates channel amplitudes. Channel RMS is measured before masking over valid frames pooled across the current distributed batch, using shared statistics for speech, music, and sound. Each contiguous channel group has a soft RMS target, with targets decreasing across successive groups; deviations are penalized as part of the scale objective. Each channel contributes separately, with weights accounting for the loss weight and averaging factor. Mean penalties and soft upper bounds on channel and per-frame group RMS complete the scale objective $\mathcal{L}_{\text{scale}}$. Together, these terms regulate amplitude statistics without directly normalizing or clipping decoder inputs. The target profile specifies a scale preference rather than a measured importance ranking. Full definitions and numerical settings are provided in the appendix.

Once the transition is complete, the second-stage objective is

$$\mathcal{L}_{\text{ordered}} = \mathcal{L}_{\text{audio}}(x, \hat{x}_K) + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{scale}}.$$

The teacher remains frozen; the acoustic encoder, decoder, and semantic head are optimized jointly, and the discriminator retains its adversarial objective. During an initial warm-up, the head is trained on detached acoustic latents, with reconstruction training continuing throughout. Training then gradually shifts from full-dimensional inputs to domain-dependent prefix sampling, introduces semantic gradients into the encoder and scale regularization, and phases out direct directional supervision on valid speech frames. The first-stage direction and radius losses are not added to the objective above. The transition schedule is detailed in the appendix. After training, the codec is frozen for downstream use with a selected prefix size.

3.4 Continuous Generation with Selected Prefixes

As illustrated in Figure 2, with the codec frozen, each downstream model selects a fixed prefix length K for both prediction and audio-history conditioning. We adopt the patch-based autoregressive framework of DiTAR (Jia et al., 2025). Each patch contains two 10 Hz acoustic frames, spanning 200 ms. An aggregation encoder maps it to one audio embedding, yielding five audio positions per second. These embeddings are interleaved with text-token embeddings and processed by the LLM. A local DiT predicts the next patch conditioned on the LLM output, the preceding acoustic block, and speaker information. Whereas dots.tts uses a separately pretrained semantic encoder to transform generated latents into feedback embeddings (Lian et al., 2026), our aggregation encoder operates directly on the selected acoustic prefix, without an additional semantic extraction stage.

We train the local DiT to predict velocities along the standard linear flow-matching path (Lipman et al., 2023). The interpolation and target-velocity equations, together with timestep-sampling details, are given in Appendix E.2. The model predicts $\hat{v}$ from the noisy patch and its conditions, leaving the preceding acoustic block uncorrupted. Both targets and history retain the codec's original coordinates, without per-channel standardization, so the scale profile learned during representation training is preserved.

Selecting K determines which channels are modeled; weighting determines how their prediction errors contribute to optimization. To adjust this allocation at a fixed capacity, we apply group-specific channel weights to the velocity MSE and cosine-distance terms used in our implementation:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\text{valid}}\left[\frac{\sum_{c=0}^{K-1}\omega_c(\hat{v}_c - v_c^\star)^2}{\sum_{c=0}^{K-1}\omega_c} + 1 - \cos\big(\sqrt{\omega}\odot\hat{v},\ \sqrt{\omega}\odot v^\star\big)\right].$$

The velocity vectors correspond to individual frames, and the expectation averages over valid prediction positions, sampled timesteps, and noise. Positive weights ω_c are shared within each channel group. Setting all channel weights to one recovers the original objective. Assigning larger weights to earlier groups emphasizes their prediction errors while retaining supervision for later channels. Dividing the MSE by the weight sum prevents a common rescaling of the weights from directly amplifying this term. This adjustment changes the generator's objective while leaving the codec and its latent coordinates fixed. Section 4.3.2 specifies the channel weights; timestep sampling and reduction details are provided in the appendix.

The generative model is trained with this acoustic objective alongside text-token prediction and stop-prediction losses. Ground-truth acoustic patches provide history during training; generated patches replace them at inference. Each patch is sampled from Gaussian noise by integrating from $t = 1$ to $t = 0$, then fed back through the aggregation encoder and supplied as local history for the next DiT prediction. The sampled latents are zero-padded to 128 channels when necessary and decoded by the shared frozen codec under its streaming context constraints. Neither the semantic teacher nor the codec's auxiliary semantic head is required in this process.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Scope and representations. We evaluate speech generation and speech/music reconstruction with the shared Transformer Prior-Aligned Autoencoder (TPAE). The first-stage model is Base TPAE; Ordered TPAE is obtained after prefix training. Codec training uses speech, music, and other sounds with a separately trained, frozen ASR teacher, while downstream speech generation uses approximately 450,000 hours of speech. Data composition and preprocessing are given in Appendix F.1.

Comparisons. VAE references cover direct acoustic-latent conditioning, external semantic-encoder feedback, and semantic-fused latents. TPAE comparisons use the same frozen Ordered TPAE at 64 and 96 channels, with either uniform or tiered channel weighting at 96 channels. Table 1 summarizes the generation systems and their training and guidance settings; additional TPAE training details are in Appendix E.3.

Evaluation. Speech generation uses all 2,020 Seed-TTS test-zh utterances (Anastassiou et al., 2024), with non-continuation inference and 10 sampling steps. We report Paraformer-zh CER (Gao et al., 2022), WavLM-based speaker SIM (Chen et al., 2022), and UTMOS (Saeki et al., 2022). Reconstruction uses fixed sets of 100 speech and 100 music clips, comparing Ordered TPAE prefixes with the full-dimensional Base TPAE reference. We report PESQ (Rix et al., 2001), SIM, and SI-SDR; full protocols and additional metrics are in Appendix F.2. The prescribed Paraformer-zh protocol is retained for consistency, although ASR homophone substitutions can inflate CER (Wang et al., 2025).

4.2 SPEECH GENERATION PERFORMANCE

We first examine whether the learned representation supports intelligible, natural speech generation while preserving speaker identity. Table 1 compares the three Ordered TPAE configurations with the VAE reference systems defined in Section 4.1.

TPAE supports effective speech generation within a limited training budget without an external semantic encoder. After approximately one epoch, TPAE-64 achieves a CER of 1.457%, SIM of 0.718, and UTMOS of 3.140. Compared with VAE-64 without a semantic encoder at a similar training stage, speaker similarity is comparable, predicted speech quality is slightly better, and content accuracy is slightly lower. All three metrics improve over VAE-128, with TPAE using a larger DiT head in both comparisons. Among systems using 12-layer DiT heads with a hidden width of 1,536, TPAE-64 approaches the CER of VAE-128 with a semantic encoder (1.447%), achieves higher UTMOS (3.140 versus 2.837), and outperforms the semantic-fused VAE on all three metrics. TPAE-96-weighted reaches a UTMOS of 3.159. After approximately one epoch, TPAE thus matches or exceeds selected metrics of VAE reference systems trained for approximately two epochs, without requiring an external semantic encoder.

The three TPAE configurations exhibit different performance trade-offs: 64 channels yield the lowest CER, uniform weighting at 96 channels gives the highest SIM among TPAE configurations, and weighted training at 96 channels achieves the highest UTMOS. The VAE-128 semantic-encoder reference records a slightly higher SIM (0.730) than uniform TPAE-96 (0.727), while TPAE offers selectable prefixes and weighting without an external semantic encoder. Overall, prefix length and training weights provide alternative generation configurations without retraining the codec. The following sections examine reconstruction gains from longer prefixes and how channel weighting affects learning at a fixed dimensionality.

Table 1: Speech generation on Seed-TTS test-zh (2,020 utterances). Results use non-continuation inference with 10 sampling steps. DiT size, training duration, condition dropout, and guidance are listed. All TPAE results are from step 7,000; the best value in each metric is bold.

Representation	Channels	DiT	Epochs	Condition dropout	Guidance	CER (%) ↓	SIM ↑	UTMOS ↑
VAE	64	6 × 1024	≈1	LLM	CFG 2	**1.331**	0.716	3.097
VAE	128	6 × 1024	≈1	LLM	CFG 2	1.805	0.692	2.263
VAE + semantic encoder	128	12 × 1536	≈2	LLM	CFG 2	1.447	**0.730**	2.837
Semantic-fused VAE	128	12 × 1536	≈2	LLM	CFG 2	1.520	0.690	2.385
TPAE	64	12 × 1536	≈1	LLM	CFG 2	1.457	0.718	3.140
TPAE	96	12 × 1536	≈1	LLM	CFG 2	1.833	0.727	2.988
TPAE, weighted	96	12 × 1536	≈1	LLM	CFG 2	1.699	0.722	**3.159**

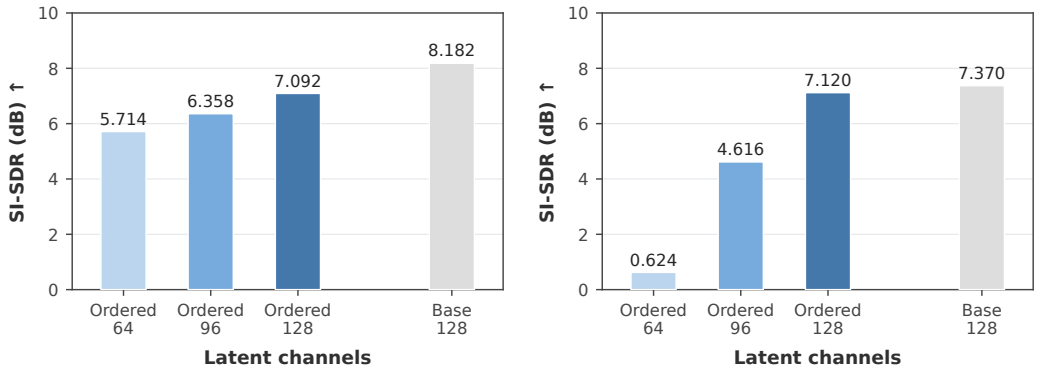
4.3 Capacity and Learning Priorities

4.3.1 Quality Across Prefix Dimensions

Having established that compact prefixes support speech generation, we examine what additional channels contribute to reconstruction. As detailed in Appendix C, speech PESQ and SIM improve consistently as the prefix expands from 64 to 96 and 128 channels within the same frozen codec. Moving from 64 to 96 channels raises PESQ from 3.4510 to 3.5773 and SIM from 0.9135 to 0.9359, with further gains at 128 channels.

Figure 3 shows that both speech and music benefit from additional capacity, although the gains differ in magnitude. Extending the prefix from 64 to 96 channels improves SI-SDR by 0.644 dB for speech and 3.992 dB for music; the extension to 128 channels yields further gains of 0.733 and 2.504 dB, respectively. The shared codec thus supports quality improvements within both domains, with prefix length providing a way to accommodate their different capacity requirements.

At full dimensionality, the ordered model retains a reconstruction gap relative to the Base TPAE, whose decoder was trained exclusively on full-dimensional latents. Speech PESQ decreases by 0.072, and SI-SDR by 1.091 dB for speech and 0.250 dB for music. These differences document the trade-off observed across the two training stages. Viewed alongside the generation results in Section 4.2, the reconstruction gains from longer prefixes motivate a further question: can channel weighting improve generation at a fixed prefix length? We examine this question in Section 4.3.2.



(a) Speech (b) Music

Figure 3: Reconstruction SI-SDR for Base TPAE and Ordered TPAE on the same 100 speech and 100 music samples. Ordered prefixes share one frozen encoder and decoder; omitted channels are zeroed. Base TPAE is the full-dimensional reference before ordered-prefix training. Latents are decoded in FP32.

4.3.2 Channel Weighting at a Fixed Dimensionality

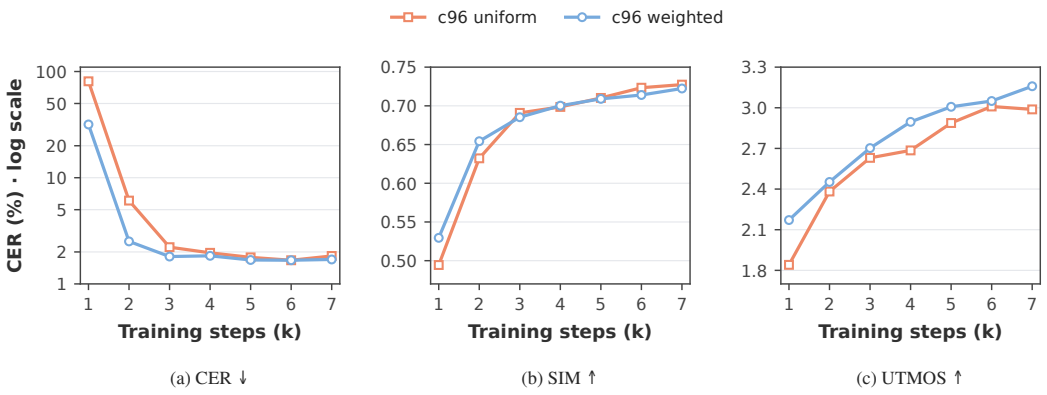
Longer prefixes improve reconstruction fidelity, but the generator must learn to use the additional capacity. We therefore ask whether channel weighting improves learning without reducing the 96-channel target. The comparison holds the Ordered TPAE tokenizer, 12-layer DiT with a hidden width of 1,536, training data, and inference settings fixed. Uniform weighting is compared with weights of 1.5, 1.0, 0.6, and 0.25 for channel groups $[0, 32)$, $[32, 48)$, $[48, 64)$, and $[64, 96)$, respectively. All 96 channels remain supervised. A uniformly weighted 64-channel model provides an additional compact-target reference; its full learning curve is reported in Appendix B.4. Because its prediction and audio-history input dimensions differ, it is not part of the fixed-dimensional comparison.

Channel weighting substantially improves content accuracy early in training (Figure 4). At step 2,000, CER falls from 6.090% under uniform weighting to 2.514%, approaching the 2.341% achieved by the smaller, uniformly weighted 64-channel reference (Appendix B.4). The CER advantage decreases from 3.576 percentage points at step 2,000 to 0.408 at step 3,000. The two models are nearly tied at step 6,000, with a smaller advantage of 0.133 points for weighted training at step 7,000. Emphasizing earlier channel groups thus improves early generation performance without discarding the additional channels.

The benefit extends to predicted speech quality: weighted training yields higher UTMOS at all seven evaluation checkpoints. At step 7,000, UTMOS rises from 2.9881 to 3.1589 and CER decreases by 0.133 percentage points, while SIM drops by 0.0050. Weighting therefore improves early content accuracy and consistently raises UTMOS over the evaluated training interval, but does not uniformly improve speaker similarity. Together with the preceding results, this comparison identifies two ways to adapt a shared representation for generation: prefix length selects the retained capacity, and channel weights adjust learning priorities within that capacity. Neither requires retraining the codec.

5 Conclusion

We present a framework that integrates semantic guidance, acoustic reconstruction, and selectable latent capacity within a single codec for streaming generation. Experiments show that compact prefixes support effective speech generation, while longer prefixes improve reconstruction fidelity for both speech and music. At 96 channels, tiered weighting improves early content accuracy and yields higher UTMOS at every evaluated checkpoint, although speaker similarity does not consistently improve. These findings indicate that latent capacity and the emphasis placed on different channels during training can be adjusted separately, allowing the same codec to serve different quality and training requirements. Our evaluation covers speech generation within a limited training budget and reconstruction in both domains. Future work will examine the approach at larger training scales and extend the evaluation to music generation.



(a) CER ↓ (b) SIM ↑ (c) UTMOS ↑

Figure 4: Learning curves for uniform and tiered channel weighting at 96 channels. Evaluation uses all 2,020 Seed-TTS test-zh utterances; CER is shown on a logarithmic axis.

REFERENCES

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-TTS: A Family of High-Quality Versatile Speech Generation Models. arXiv preprint arXiv:2406.02430, 2024. URL https://arxiv.org/abs/2406.02430.

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. FlexTok: Resampling images into 1D token sequences of flexible length. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2241–2292. PMLR, 2025. URL https://proceedings.mlr.press/v267/bachmann25a.html.

Junyu Chen, Dongyun Zou, Wenkun He, Junsong Chen, Enze Xie, Song Han, and Han Cai. DC-AE 1.5: Accelerating diffusion model convergence with structured latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Chen_DC-AE_1.5_Accelerating_Diffusion_Model_Convergence_with_Structured_Latent_Space_ICCV_2025_paper.pdf.

Peijie Chen, Wenhao Guan, Weijie Wu, Kaidi Wang, Daiyu Huang, Zhuanling Zha, Junbo Li, Jun Fang, Qingyang Hong, and Lin Li. SARA: A Dual-Stream VAE for High-Fidelity Speech Generation via Integrating Semantic and Acoustic Representations. arXiv preprint arXiv:2606.11611, 2026. URL https://arxiv.org/abs/2606.11611.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. arXiv preprint arXiv:2110.13900, 2022. URL https://arxiv.org/abs/2110.13900.

Changhao Cheng, Wei Wang, Wangyou Zhang, Dongya Jia, Jian Wu, Zhuo Chen, and Yanmin Qian. On the Distillation Loss Functions of Speech VAE for Unified Reconstruction, Understanding, and Generation. arXiv preprint arXiv:2604.12383, 2026. URL https://arxiv.org/abs/2604.12383.

Junyu Dai, Xiaoyue Duan, Xinyue Fan, Yihan Feng, Jingbei Li, Xiangang Li, Yunjia Li, Lejun Min, Yufei Shi, Xingchen Song, Yiran Wang, Cheng Wen, Menglin Wu, Bajian Xiang, Huaicheng Zhang, Han Zhao, and Ruichen Zheng. Qwen-Audio-3.0-Gen-Preview Technical Report. arXiv preprint arXiv:2607.27011, 2026. URL https://arxiv.org/abs/2607.27011.

Heinrich Dinkel, Xingwei Sun, Gang Li, Jiahao Mei, Yadong Niu, Jizhong Liu, Xiyang Li, Yifan Liao, Jiahao Zhou, Junbo Zhang, and Jian Luan. DashengTokenizer: One layer is enough for unified audio understanding and generation. arXiv preprint arXiv:2602.23765, 2026. URL https://arxiv.org/abs/2602.23765.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Interspeech 2022*, pp. 2063–2067, 2022. doi: 10.21437/Interspeech.2022-9996. URL https://www.isca-archive.org/interspeech_2022/gao22b_interspeech.html.

Yitian Gong, Kuangwei Chen, Zhaoye Fei, Xiaogui Yang, Ke Chen, Yang Wang, Kexin Huang, Mingshu Chen, Ruixiao Li, Qingyuan Cheng, Shimin Li, and Xipeng Qiu. MOSS-Audio-Tokenizer: Scaling audio tokenizers for future audio foundation models. arXiv preprint arXiv:2602.10934, 2026. URL https://arxiv.org/abs/2602.10934.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376, 2006. doi: 10.1145/1143844.1143891. URL https://www.cs.toronto.edu/~graves/icml_2006.pdf.

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

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, and Yuxuan Wang. 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.

R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pp. 125–128, 1993. doi: 10.1109/PACRIM.1993.407206. URL https://doi.org/10.1109/PACRIM.1993.407206.

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.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka Representation Learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af9993100e42-Abstract-Conference.html.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for End-to-End Tuning of Latent Diffusion Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Leng_REPA-E_Unlocking_VAE_for_End-to-End_Tuning_of_Latent_Diffusion_Transformers_ICCV_2025_paper.html.

Shi Lian, Changtao Li, Bohan Li, Hankun Wang, Da Zheng, Junfeng Tian, Yufeng Ma, Colin Zhang, and Kai Yu. dots.tts Technical Report. arXiv preprint arXiv:2606.07080, 2026. URL https://arxiv.org/abs/2606.07080.

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

Huadai Liu, Wen Wang, Kaicheng Luo, Qian Chen, Xiangang Li, and Wei Xue. STAR-VAE: Structured Topology-Aware Regularization for Audio Reconstruction and Generation. arXiv preprint arXiv:2606.23064, 2026. URL https://arxiv.org/abs/2606.23064.

Zhijun Liu, Shuai Wang, Sho Inoue, Qibing Bai, and Haizhou Li. Autoregressive Diffusion Transformer for Text-to-Speech Synthesis. arXiv preprint arXiv:2406.05551, 2024. URL https://arxiv.org/abs/2406.05551.

Zhikang Niu, Shujie Hu, Jeongsoo Choi, Yushen Chen, Peining Chen, Pengcheng Zhu, Yunting Yang, Bowen Zhang, Jian Zhao, Chunhui Wang, and Xie Chen. Semantic-VAE: Semantic-Alignment Latent Representation for Better Speech Synthesis. arXiv preprint arXiv:2509.22167, 2026. URL https://arxiv.org/abs/2509.22167.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, and Furu Wei. VibeVoice Technical Report. arXiv preprint arXiv:2508.19205, 2025. URL https://arxiv.org/abs/2508.19205.

Oren Rippel, Michael Gelbart, and Ryan Adams. Learning Ordered Representations with Nested Dropout. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32(2) of *Proceedings of Machine Learning Research*, pp. 1746–1754. PMLR, 2014. URL https://proceedings.mlr.press/v32/rippel14.html.

A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pp. 749–752, 2001. doi: 10.1109/ICASSP.2001.941023. URL https://doi.org/10.1109/ICASSP.2001.941023.

Simon Rouard, Manu Orsini, Axel Roebel, Neil Zeghidour, and Alexandre Défossez. Continuous Audio Language Models. arXiv preprint arXiv:2509.06926, 2026. URL https://arxiv.org/abs/2509.06926.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Interspeech 2022*, pp. 4521–4525, 2022. doi: 10.21437/Interspeech.2022-439. URL https://www.isca-archive.org/interspeech_2022/saeki22c_interspeech.html.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159. URL https://www.jstor.org/stable/1412159.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011. doi: 10.1109/TASL.2011.2114881. URL https://doi.org/10.1109/TASL.2011.2114881.

He Wang, Linhan Ma, Dake Guo, Xiong Wang, Lei Xie, Jin Xu, and Junyang Lin. ContextASR-Bench: A massive contextual speech recognition benchmark. arXiv preprint arXiv:2507.05727, 2025. URL https://arxiv.org/abs/2507.05727v2.

Huimeng Wang, Hui Lu, Jiajun Deng, Haoning Xu, Youjun Chen, Xueyuan Chen, Zhaoqing Li, Shuhai Peng, Shiyin Kang, and Xunying Liu. SemaVoice: Semantic-Aware Continuous Autoregressive Speech Synthesis. arXiv preprint arXiv:2605.16964, 2026. URL https://arxiv.org/abs/2605.16964.

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. URL https://arxiv.org/abs/2608.07462.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6194–6198, 2020. URL https://sewplay.github.io/files/papers/2020/icassp_000619 4.pdf.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, Hui Zhan, Jie Gao, Jing Peng, Jingdong Chen, Jun Zhou, Kaimeng Ren, Ming Yang, Mingxue Yang, Qiang Xu, Qin Zhao, Ruijie Xiong, Shaoxiong Lin, Xuezhi Wang, Yi Yuan, Yifei Wu, Yongjie Lyu, Zhengyu He, Zhihao Qiu, Zhiqiang Fang, and Ziyuan Huang. Ming-UniAudio: Speech LLM for Joint Understanding, Generation and Editing with Unified Representation. arXiv preprint arXiv:2511.05516, 2025. URL https://arxiv.org/abs/2511.05516.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. Generation: Taming Optimization Dilemma in Latent Diffusion Models. arXiv preprint arXiv:2501.01423, 2025. URL https://arxiv.org/abs/2501.01423.

Zhen Ye, Xu Tan, Yiming Li, Guangyan Zhang, Chimin Chan, Haohe Liu, Zhengxi Liu, Hongzhan Lin, Zheqi Dai, Xinshen Zhang, Peiwen Sun, Qiuqiang Kong, and Wei Xue. Which Speech Representation Better Matches Text-Native Reasoning? A Study of Speech-Text Alignment on Frame Rate and Representation. arXiv preprint arXiv:2606.12199, 2026. URL https://arxiv.org/abs/2606.12199.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion Transformers with Representation Autoencoders. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c4141c12660ad3625eb4ae845e0a6f9-Abstract-Conference.html.

A Design Exploration and Trade-offs

This section traces our choices of semantic supervision, temporal granularity, and acoustic backbone through early experiments. The results illustrate the benefits and costs of these alternatives for semantic alignment, acoustic reconstruction, and generative learning, informing the design of the current framework.

A.1 Semantic Supervision versus Direct Fusion

We first examine whether semantic features should contribute directly to the decoder input or serve only as supervision for the acoustic encoder. The former reconstructs from a combination of semantic and acoustic features; the latter requires acoustic latents to support reconstruction on their own. Early codec experiments compare these alternatives and examine the strength of semantic guidance.

With direction and radius weights fixed at 1.0 and 0.1, replacing the summed input $s+a$ with acoustic latents a improves PESQ from 2.991 to 3.114 and SI-SDR from 2.656 to 4.367 dB. Directional cosine similarity between acoustic latents and semantic targets also rises from 0.617 to 0.658. In this setting, removing semantic features from the decoder input improves both reconstruction and alignment, supporting their use as supervision for an independent acoustic reconstruction path.

Stronger constraints, however, do not consistently improve reconstruction. With the summed input, increasing the direction weight from zero to 0.1 improves SI-SDR but slightly reduces PESQ; raising it to 1.0 lowers both scores relative to 0.1. Holding that coefficient at 1.0, stronger radius regularization brings the acoustic-to-semantic norm ratio closer to one, at the cost of reconstruction fidelity and with little change in directional similarity. These observations support balancing semantic guidance with reconstruction, rather than optimizing alignment or scale in isolation.

Table 2: Early codec comparisons of decoder inputs and auxiliary constraint strengths. All runs share the stage-1 decoder initialization and are evaluated at epoch 100 on the same 100 speech clips. Here, s denotes semantic features and a acoustic latents. Directional cosine measures their alignment, and the norm ratio is acoustic norm divided by semantic norm. Dashes indicate values not reported here. These fixed-stage results inform design choices rather than establish converged performance or isolate component effects in the final Transformer model.

Decoder input	Direction weight	Radius weight	PESQ ↑	SI-SDR (dB) ↑	Directional cosine ↑	Norm ratio
$s + a$	0	0	3.207	2.720	—	—
$s + a$	0.1	0	3.155	4.307	—	—
$s + a$	1.0	0	3.139	3.182	0.620	18.98
$s + a$	1.0	0.1	2.991	2.656	0.617	2.25
$s + a$	1.0	1.0	2.936	1.754	0.619	1.09
a	1.0	0.1	3.114	4.367	0.658	0.93

A.2 Temporal Granularity and Alignment

After choosing an acoustic-only reconstruction path, we examine its temporal resolution. A frame rate of 10 Hz provides framewise correspondence with the semantic targets, with two frames forming each downstream generation block. Reducing this rate at a fixed channel dimension also leaves fewer latent vectors per second for reconstruction. We explore this trade-off using an earlier prototype with convolutional acoustic encoders and decoders; its relationship to the subsequent Transformer design is discussed in Appendix A.3. For 24 kHz audio, hop sizes of 2,400 and 2,048 samples yield frame rates of 10 Hz and approximately 11.72 Hz, respectively. The two settings retain the same channel dimension and are evaluated at nearby training checkpoints.

The higher rate improves PESQ, STOI, MCD, and SI-SDR in both checkpoint comparisons examined; Table 4 presents the later pair. These results indicate that denser acoustic representations benefit reconstruction in this setting. The resulting sequence no longer corresponds framewise to the existing 10 Hz semantic targets, however, and would require additional temporal adaptation.

We therefore retain 10 Hz to preserve direct temporal correspondence with the semantic targets, accepting the reconstruction trade-off observed in these experiments.

Table 3: Reconstruction results from the later checkpoint pair in the CNN prototype experiments. Changing the hop size also changes the stride factorization, and the checkpoints differ in training duration. These results characterize the observed trade-off in this prototype rather than isolate the effect of frame rate in the final Transformer model.

Hop size (samples)	Frame rate (Hz)	Training step	PESQ ↑	STOI ↑	MCD ↓	SI-SDR (dB) ↑
2,400	10.00	324,235	3.406	0.9489	2.653	8.31
2,048	11.72	340,135	3.481	0.9547	2.437	9.29

A.3 Acoustic Backbone Selection

After establishing acoustic-only reconstruction and a shared temporal resolution, we examine whether the acoustic backbone can support both reconstruction fidelity and semantic alignment. Table 5 compares the CNN-based PAE explored earlier with Base TPAE on the same LibriTTS test-clean set. The Transformer configuration achieves better reconstruction scores and higher semantic cosine similarity, demonstrating its ability to support both objectives.

A representation that supports reconstruction and semantic alignment must also be tested in downstream generation. Both systems use acoustic latents as prediction targets and audio-history inputs without an external semantic encoder. In Table 6, CNN-PAE achieves a lower CER, whereas Base TPAE paired with a wider DiT obtains higher SIM and UTMOS. Lower content error rates thus do not necessarily coincide with better speaker similarity or predicted speech quality. Taken together, the reconstruction and generation results establish TPAE as a viable configuration that combines acoustic fidelity, semantic alignment, and direct use in speech generation.

Table 4: Reconstruction and semantic alignment on LibriTTS test-clean. PESQ and STOI cover 4,836 valid utterances; the remaining metrics cover all 4,837 utterances. WavLM SIM measures speaker similarity between reconstructed and reference audio. Semantic cosine measures directional alignment between acoustic latents and semantic targets.

Model	PESQ-WB ↑	STOI ↑	FDopenl3 ↓	WavLM SIM ↑	Semantic cosine ↑
CNN-PAE	3.6115	0.9703	5.8659	0.9157	0.3455
Base TPAE	**3.7937**	**0.9789**	**5.4572**	**0.9615**	**0.4959**

Table 5: Speech generation with CNN-PAE and Base TPAE. Both systems use full 128-channel acoustic latents as prediction targets and audio-history inputs without an external semantic encoder. Results cover all 2,020 Seed-TTS zh utterances with CFG 2.0, 10 sampling steps, and non-continuation inference. DiT sizes and training steps are listed for each system.

Representation	DiT	Training steps	CER (%) ↓	SIM ↑	UTMOS ↑
CNN-PAE	12×1536	13,488	**1.300**	0.6616	2.4970
Base TPAE	12×1920	13,574	1.730	**0.7353**	**2.9526**

B Representation Adaptation for Generation

This section examines three downstream choices for Base TPAE: semantic feedback, generator capacity, and channel scaling. The experiments explore how these choices affect generative learning while keeping the codec fixed.

B.1 Direct Acoustic History and Additional Semantic Guidance

Incorporating semantic structure into acoustic representations allows the same latent to serve as both a prediction target and a history condition. To assess the value of additional semantic guidance, we compare direct acoustic history with semantic feedback, training-only auxiliary alignment, and their combination. Feedback adds a feature extraction module to the history path, whereas auxiliary alignment supervises the generator's internal features without adding an inference module.

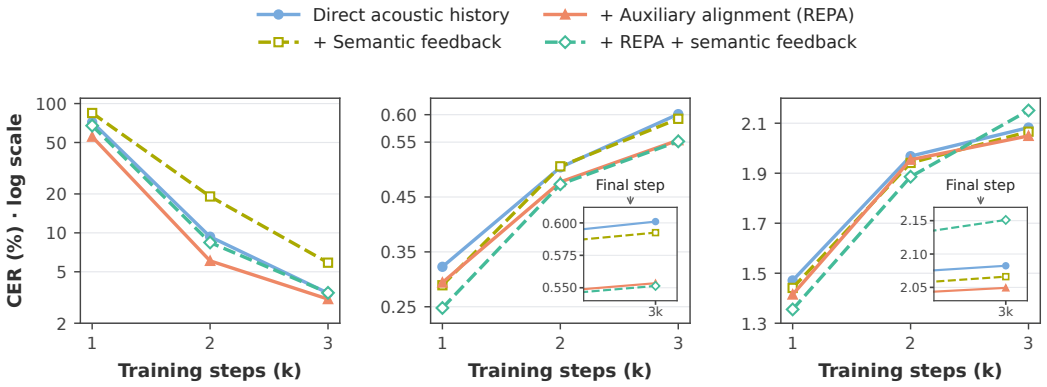
Direct acoustic history achieves lower CER and higher UTMOS than feedback alone at all three shared checkpoints (Figure 5). At step 3,000, adding feedback raises CER from 3.417% to 5.879%. Auxiliary alignment alone lowers CER at each checkpoint, but also reduces SIM and UTMOS relative to the direct path. Combining feedback and alignment yields the highest UTMOS at step 3,000, although CER remains slightly higher and SIM lower than with direct acoustic history. Together, these early-training results support using acoustic latents directly for history conditioning. The additional guidance introduces inference modules or training objectives without consistent gains across the reported metrics. We therefore retain the direct history path without either addition in the subsequent experiments.

B.2 Generator Capacity and Channel Normalization

Having retained direct acoustic history, we examine how to improve generative learning with the full 128-channel representation. We consider two adjustments: widening the DiT to increase model capacity and applying per-channel normalization to change the modeling scales. Normalization operates only within the DiT and is inverted before its output is used for history conditioning or waveform decoding, leaving the codec and history interface unchanged.

At the shared checkpoint of 8,000 steps, widening the DiT improves SIM and UTMOS but slightly increases CER, revealing a trade-off between content accuracy and acoustic quality. With the wider DiT fixed, per-channel normalization improves all three metrics at the training endpoint, reducing CER from 1.730% to 1.605%. These results indicate that learning from the full representation depends on both generator capacity and the modeling coordinates. For Base TPAE, a reversible change of scale improves generation while preserving all channels and the original decoding path.

The gains from normalization suggest that the original channel scales may not be well suited to generative learning. One possibility is that some low-amplitude channels contain useful information that is insufficiently learned; the next section examines this hypothesis through channel sensitivity analysis. This motivates a shift from rescaling an existing latent space to organizing it during codec training. Ordered TPAE follows this direction by learning decodable prefixes and a channel-scale



(a) CER ↓ (b) SIM ↑ (c) UTMOS ↑

Figure 5: Semantic guidance with full 128-channel Base TPAE latents on all 2,020 Seed-TTS zh utterances. CER uses a logarithmic axis. Insets enlarge the final evaluated checkpoint at step 3,000 for SIM and UTMOS. These shared early-training checkpoints characterize learning behavior rather than final convergence.

hierarchy. Its downstream models preserve these native scales and use prefix selection and tiered losses to adjust prediction dimensionality and learning priorities.

Table 6: All configurations use full 128-channel Base TPAE latents without an external semantic encoder. Evaluation covers all 2,020 Seed-TTS zh utterances with CFG 2.0, 10 sampling steps, and non-continuation inference. Capacity is compared at the shared checkpoint of 8,000 steps; normalization is compared at the common training endpoint. Bold values indicate the best score within each comparison.

Comparison	DiT	Channel normalization	Training steps	CER (%) ↓	SIM ↑	UTMOS ↑
Capacity	12×1536	No	8,000	**2.045**	0.7019	2.6404
Capacity	12×1920	No	8,000	2.138	**0.7153**	**2.7823**
Scaling	12×1920	No	13,574	1.730	0.7353	2.9526
Scaling	12×1920	Yes	13,574	**1.605**	**0.7401**	**3.0021**

B.3 From Channel Sensitivity to Ordered Representations

The gains from per-channel normalization motivate us to ask whether channel scales reflect their effects on decoded audio. We compute the mean absolute gradient of the MR-STFT reconstruction loss with respect to each Base TPAE latent channel, using 100 hours each of speech and music. On 2,000 speech clips, we also perturb one channel at a time by increasing its deviation from the channel mean by 20%, leaving the others unchanged.

Spectral gradient rankings are similar across speech and music, with a Spearman correlation (Spearman, 1904) of 0.887. However, channel standard deviation correlates negatively with gradient sensitivity in both domains (-0.336 and -0.290, respectively). Channels 0 and 1 illustrate this mismatch (Figure 6): their speech variance shares are below average, yet they rank highly in both spectral gradient sensitivity and speech UTMOS degradation under perturbation. These observations suggest that channel amplitude alone does not adequately reflect sensitivity to perturbations.

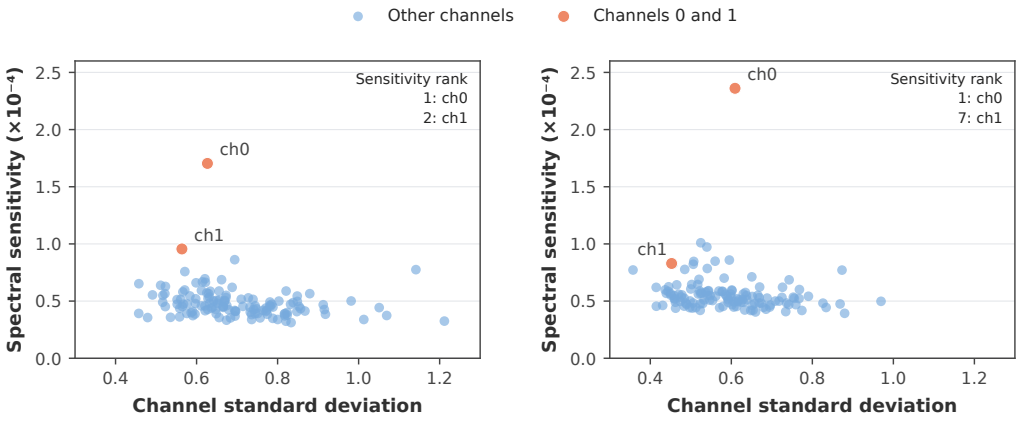
Because sensitivity depends on the metric and decoder, we do not directly adopt the probe rankings as a channel order. Instead, Ordered TPAE uses prefix reconstruction to train earlier channels to support reconstruction under restricted capacity, shaping channel organization through the reconstruction task itself. Tiered losses then adjust downstream learning priorities across channel groups. The probes motivate this design; the main-text experiments evaluate prefix usability and the effects of weighting.

B.4 Compact 64-Channel Generation Reference

Table 7 provides the per-checkpoint values for the uniformly weighted 64-channel compact-target reference discussed in Section 4.3.2. It differs from the 96-channel runs in both prediction and audio-history dimensions, so it is included for context rather than as a controlled weighting comparison.

Table 7: Learning curve of the uniformly weighted 64-channel compact-target reference discussed in Section 4.3.2. Evaluation uses all 2,020 Seed-TTS test-zh utterances and the same inference protocol as the 96-channel runs.

Training steps	CER (%) ↓	SIM ↑	UTMOS ↑
1,000	22.793	0.5387	2.1737
2,000	2.341	0.6525	2.6209
3,000	1.612	0.6895	2.7023
4,000	1.493	0.7030	2.8597
5,000	1.505	0.7095	3.0645
6,000	1.447	0.7167	3.1065
7,000	1.457	0.7182	3.1402



(a) Speech (b) Music

Figure 6: Channel scale and spectral sensitivity of Base TPAE. Each panel shows all 128 channels, with channels 0 and 1 highlighted. Sensitivity is the mean absolute gradient of the MR-STFT reconstruction loss with respect to each latent channel. Rank 1 denotes the greatest sensitivity within each domain; both panels use the same axis ranges.

C Additional Reconstruction Results

This section extends the prefix reconstruction comparison to 32 and 48 channels and adds intelligibility and spectral error metrics to examine the quality cost of further compression. All prefixes come from the same Ordered TPAE and share a decoder, without training a separate codec for each dimensionality. Evaluation uses the same fixed sets of 100 speech and 100 music clips, with latents stored in BF16 and decoded in FP32. Metric definitions and evaluation settings are provided in Appendix F.2.

The full results show that decodable prefixes do not make compression cost-free. Reducing the speech prefix from 64 to 32 channels lowers STOI from 0.9523 to 0.9057 and SIM from 0.9135 to 0.7684, whereas expanding from 32 to 48 channels improves every reported speech metric. Music also benefits from longer prefixes, with substantial reconstruction gains between 64 and 128 channels. MCD and MR-STFT decrease consistently with dimensionality in both domains, showing improvements in spectral fidelity alongside the waveform metrics. Shorter prefixes therefore provide compact representations with measurable quality costs rather than lossless substitutes for the full representation. Prefix selection should combine these reconstruction results with downstream generation performance to meet the requirements of the task and the desired quality.

Table 8: All prefix sizes use the same 100 speech clips, with omitted suffix channels set to zero. SIM uses the original recordings as references. Bold values indicate the best score in each column.

Channels	PESQ ↑	STOI ↑	SIM ↑	SI-SDR (dB) ↑	MCD ↓	MR-STFT ↓
32	2.9085	0.9057	0.7684	-0.4482	3.0047	1.0092
48	3.2545	0.9349	0.8678	3.5291	2.6459	0.8839
64	3.4510	0.9523	0.9135	5.7144	2.3967	0.8077
96	3.5773	0.9615	0.9359	6.3581	2.2678	0.7726
128	**3.6767**	**0.9682**	**0.9495**	**7.0916**	**2.1575**	**0.7442**

Table 9: All prefix sizes use the same 100 music clips, with omitted suffix channels set to zero. Bold values indicate the best score in each column.

Channels	SI-SDR (dB) ↑	MCD ↓	MR-STFT ↓
32	-8.1764	6.1436	1.3358
48	-2.8347	5.5922	1.1896
64	0.6237	5.1842	1.0634
96	4.6157	4.7305	0.9238
128	**7.1200**	**4.2392**	**0.8355**

D Sampling and Guidance Settings

To establish a common inference configuration, we compare sampling steps, guidance strength, and solver-state precision on a fixed Base TPAE generation model using 200 Chinese utterances from Seed-TTS. Classifier-free guidance (CFG) combines conditional and unconditional predictions (Ho & Salimans, 2022). With CFG fixed at 2.0 and native bfloat16 (BF16) solver states, increasing sampling from 5 to 10 steps improves all three metrics, but further increases to 15–30 steps provide no additional gains. With the sampling budget fixed at 10 steps, we next examine guidance strength and numerical precision.

Reducing CFG to 1.2 yields the highest speaker similarity but substantially increases the character error rate. A setting of 1.5 instead improves similarity with a similar error rate, at the cost of lower UTMOS. Paired bootstrap resampling over utterances (Efron, 1979) supports the similarity gain and UTMOS decrease, while the 95% confidence interval for the error-rate difference includes zero. Accumulating solver states in 32-bit floating point yields no statistically detectable improvement at either 10 or 30 steps. These results support 10 steps, CFG 2.0, and native BF16 solver states as a balanced configuration, with CFG 1.5 as an alternative when speaker similarity is prioritized.

Table 10: Sampling and guidance comparisons at training step 12,000 under non-continuation inference. The FP32 variants retain BF16 model forward passes and leave codec decoding precision unchanged. Bold indicates the best point estimate in each metric.

Sampling steps	CFG	Solver state	CER (%) ↓	SIM ↑	UTMOS ↑
5	2.0	BF16 (native)	1.0460	0.7172	2.2521
10	2.0	BF16 (native)	**0.8178**	0.7203	2.3724
15	2.0	BF16 (native)	0.9677	0.7171	2.3644
20	2.0	BF16 (native)	0.8840	0.7150	2.3342
30	2.0	BF16 (native)	1.2728	0.7117	2.2940
10	1.0	BF16 (native)	2.7214	0.7219	2.1713
10	1.2	BF16 (native)	1.6575	**0.7298**	2.2419
10	1.5	BF16 (native)	0.8277	0.7277	2.3398
10	2.0	FP32	0.9188	0.7199	**2.3849**
30	2.0	FP32	1.1774	0.7115	2.2952

E Architecture and Training Details

This section supplements the main text with implementation details. Tables 11 and 12 summarize the architecture and optimization settings.

E.1 Architecture and Temporal Context

The acoustic backbone follows MOSS (Gong et al., 2026). The teacher's recognition head trains its spherical bottleneck through CTC (Graves et al., 2006); the shared prediction head instead predicts frozen teacher targets from acoustic prefixes. Two-frame blocks and six-block history windows

coordinate the teacher backbone, the lowest-rate acoustic encoder stage, and the prediction head. Table 11 lists per-layer attention windows rather than full-network receptive fields.

Table 11: The acoustic path converts 24 kHz audio to 128-dimensional latents using a downsampling factor of $300 \times 2 \times 2 \times 2$.

Component	Rate (Hz)	Width	Layers / heads	Context or operation
Teacher backbone	10	1,280	26 / 20	Two-frame blocks; six-block history
Teacher bottleneck	10	$1{,}280 \to 640 \to 320 \to 128$	—	Causal temporal pyramid; RMS normalization
Teacher recognition head	10	1,280	8 / 20	Two-frame blocks; six-block history
Acoustic encoder / decoder, upper scales	80 / 40 / 20	768	12 / 12 at each scale	History: 2 / 4 / 8 s
Acoustic encoder, lowest scale	10	1,280	32 / 20	Two-frame blocks; six-block history
Acoustic decoder, lowest scale	10	1,280	32 / 20	10 s history; two-frame lookahead
Shared prediction head	10	768	8 / 12	Two-frame blocks; six-block history; FFN width 2,048

E.2 Objective Implementation

Reconstruction uses the seven-scale log-mel L1 objective of DAC (Kumar et al., 2023), summed across power-of-two windows from 32 to 2,048 samples, with corresponding mel-bin counts of 5, 10, 20, 40, 80, 160, and 320 and quarter-window hops. Acoustic RMS normalization uses $\sqrt{\mathrm{mean}_c(a_c^2) + 10^{-6}}$, solely for auxiliary losses; the decoder receives raw latents. The teacher uses an additive RMS stabilizer of 10^{-5}. Prefix semantic loss sums squared differences between L2-normalized vectors ($\epsilon = 10^{-3}$) over channels and averages over valid speech frames across workers. Valid lengths are $\lceil L/2400 \rceil$, where L counts audio samples; base-stage alignment instead averages over complete aligned tensors. For ordered-stage scale regularization, each channel RMS is computed before prefix masking over valid frames pooled across workers, $q_c = \sqrt{\mathbb{E}_{(b,t)\in\mathcal{V}}[a_{b,c,t}^2]}$. The profile penalty is $\mathcal{L}_{\mathrm{profile}} = \sum_c w_c(\log q_c - \log \tau_{g(c)})^2$, where $g(c)$ maps a channel to its group. The final continuation phase applies

$$\mathcal{L}_{\mathrm{scale}} = 0.1\,\mathbb{E}_{c<64}(\log r_c - \log \tau_c)^2 + 0.02\,\mathbb{E}_{c\geq 64}(\log r_c - \log \tau_c)^2 + 0.01\,\mathbb{E}_c \mu_c^2 + 0.1 \max_c [r_c - 1]_+^2 + 0.01\,\mathbb{E}_{i,g}[r_{i,g} - 1]_+^2.$$

Here, μ_c, r_c are channel means and RMS values, τ_c denotes the target, $r_{i,g}$ is frame-group RMS, and $[x]_+ = \max(x, 0)$. Expectations are uniform over the indicated indices. Moments pool valid, untruncated speech and music/sound latents across workers; second moments are floored at 10^{-8}. Batch-moment penalties require at least two valid frames. The downstream DiT uses the standard linear flow-matching path for target patch z:

$$z_t = (1-t)z + t\epsilon, \qquad v^\star = \epsilon - z, \qquad \epsilon \sim \mathcal{N}(0, I),$$

where $t = 0$ denotes data and $t = 1$ denotes noise. The preceding acoustic block remains uncorrupted, and losses are averaged over valid prediction positions.

E.3 Optimization

Base training initializes the acoustic backbone from MOSS weights and freezes the teacher. Ordered training updates both acoustic modules and introduces the prediction head. Auxiliary gradients train only the head for 100 steps, while reconstruction continues. Over the next 1,000 steps, prefix sampling, encoder semantic gradients, and active scale penalties ramp up as direct directional supervision fades. The final shared scale objective is introduced later. Downstream weighted MSE

is normalized by the sum of weights; cosine distance uses square-root-weighted vectors. Both terms are averaged over valid frames.

The TPAE generation comparisons share Flash-3B initialization, a 12-layer DiT with width 1,536, and global batch size 128. Checkpoints are evaluated every 1,000 steps through step 7,000 (approximately one epoch); Table 1 reports this common endpoint alongside the VAE reference settings.

Table 12: Training settings. Ordered-stage values refer to the final phase.

Setting	Value
Mel / adversarial / feature-matching weights	15 / 0.5 / 1.5
Waveform L1 / MR-STFT reconstruction weights	0 / 0
Base directional / radial weights	1.0 / 0.1
Prefix semantic prediction weight	1.0
Prefix lengths	32 / 48 / 64 / 96 / 128
Speech / music-sound prefix probabilities	0.05, 0.10, 0.55, 0.20, 0.10 / 0, 0, 0, 0.10, 0.90
Group boundaries / RMS targets	0, 32, 48, 64, 96, 128 / 0.90, 0.60, 0.45, 0.30, 0.20
Ordered AdamW: betas / weight decay	(0.8, 0.99) / 0.001
Configured LR: acoustic modules / prediction head / discriminator	2.5×10^{-5} / 10^{-4} / 5×10^{-5}
Ordered global batch / precision	1,536 (128 GPUs $\times$ 1 $\times$ 12 accumulation steps) / BF16 AMP
Downstream optimizer: betas / epsilon / weight decay	(0.9, 0.95) / 10^{-16} / 0.1
Downstream LR / global batch	$3 \times 10^{-4} \rightarrow 3 \times 10^{-5}$, linear / 128
Warmup / planned sample visits	17,375 / 1,737,538
Time sampling	$t = \mathrm{sigmoid}(\xi),\ \xi \sim \mathcal{N}(0, 1)$
LLM-condition / speaker-condition dropout	0.2 / 0
Audio / stop / text loss weights	1.0 / 1.0 / 0.01

F Data and Evaluation Protocols

F.1 Data and Preprocessing

The three training components use internal data for semantic learning, cross-domain acoustic reconstruction, and speech generation. The semantic teacher is trained on transcribed multilingual speech. Codec data include multispeaker and overlapping speech, unaccompanied singing, music, and other sounds. During ordered training, the speech sampling probability increases to emphasize speech learning with compact prefixes while retaining the other audio categories. The three downstream TPAE configurations use the same speech data and preprocessing pipeline.

F.2 Evaluation Protocols

Table 14 summarizes the evaluation protocols. Generation SIM uses prompt audio as the reference, whereas reconstruction SIM uses the original audio. All reconstruction prefixes are evaluated on the same fixed samples. Ordered TPAE reconstruction uses latents stored in BF16 and an FP32 decoder, with valid lengths used to exclude batch padding.

STOI (short-time objective intelligibility) estimates speech intelligibility (Taal et al., 2011); we use its non-extended implementation. Mel-cepstral distortion (MCD) measures cepstral differences (Kubichek, 1993), computed with pymcd in $\mathrm{dtw_sl}$ mode, which combines dynamic time warping with a length-ratio penalty. The implementation uses 13th-order mel-cepstral analysis, a 5 ms frame period, and a warping coefficient of 0.65. MR-STFT averages spectral convergence and log-magnitude L1 distance across resolutions (Yamamoto et al., 2020). Our evaluator uses 24 kHz audio and FFT/hop/window lengths of (1024, 120, 600), (2048, 240, 1200), and (512, 50, 240) samples, with rectangular windows. STOI and MR-STFT inputs are truncated to their common length when needed. Higher STOI and lower MCD and MR-STFT indicate better scores; these metrics are computed per clip and then averaged.

Table 13: Dataset proportions are calculated by entry count, not audio duration, and differ from the sampling probabilities used during training.

Training component	Data and scale	Sampling and preprocessing
Semantic teacher	Internal transcribed multilingual speech	ASR supervision using transcripts
Ordered-training data pool	Approximately 18.43 million internal audio entries; 53.87% speech and 46.13% music and other sounds under the training categorization	24 kHz; crops of up to 12 seconds, preserving the actual duration of shorter clips
Sampling schedule	The mixed pool above	Speech / music-and-other-sounds sampling shifts from the original mixture to 90/10, then 99/1; unaccompanied singing is assigned to the speech group
Downstream speech generation	Approximately 450,000 hours of internal speech	Identical packing across the three configurations; audio sequences are capped at 500 latent frames, or 50 seconds

For the paired comparisons in Appendix D, we resample utterance-level score differences with replacement 10,000 times, preserving the pairing between configurations. We report percentile confidence intervals using the 2.5th and 97.5th percentiles of the bootstrap mean differences.

Table 14: A dash indicates that the metric is not reported for that evaluation.

Item	Speech generation	Speech reconstruction	Music reconstruction
Evaluation samples	Seed-TTS test-zh, 2,020 utterances (Anastassiou et al., 2024)	100 fixed speech samples	100 fixed music samples
CER (Gao et al., 2022)	Paraformer-zh transcription; utterance-averaged CER	—	—
WavLM-based SIM (Chen et al., 2022)	Generated versus prompt audio	Reconstructed versus original audio	—
UTMOS (Saeki et al., 2022)	Predicted quality of generated speech	—	—
Reconstruction fidelity	—	PESQ, STOI, SI-SDR, MCD, and MR-STFT distance	SI-SDR, MCD, and MR-STFT distance
Inference settings	Non-continuation; 10 sampling steps; CFG 2.0 and a uniform time schedule for TPAE	Encode the original audio and decode the selected prefix	Same as speech reconstruction

G LLM USE DISCLOSURE

Paper writing. Commercial large language models (LLMs) assisted with drafting and revising the manuscript, including grammar, phrasing, clarity, readability, and Chinese–English translation. The authors reviewed, edited, and approved the final text.

Research assistance. LLMs assisted with brainstorming research framing and experimental designs, suggesting analysis approaches, searching and summarizing prior work, discussing result interpretation, and creating or revising figures and plotting code. The authors provided the research context, checked cited literature against the source papers, verified analyses and reported results against project records, and take full responsibility for the final manuscript and figures.