HoliTok:A Coutinuous Holistic Tokenization with Robust Dual Capabilities of Speech Generation and Understanding

Anonymous ACL submission

Abstract

Unified speech foundation models require a holistic tokenization space that is both learnable by language models and decodable into high-quality waveforms. Existing speech tokenizers, however, often fail to satisfy these requirements simultaneously, leading to increased architectural complexity and more involved training designs. We propose HoliTok, a continuous **Holi**stic speech **Tok**enization model designed for unified generation-understanding modeling. HoliTok encodes 48 kHz speech into a compact 25 Hz sequence of 128-dimensional latents. It is trained with a progressive strategy that jointly preserves signal-level fidelity, incorporates semantic information, and maintains strong latent learnability. Based on this tokenization, we build a unified AR+DiT model for speech synthesis and recognition, where the same latent sequence supports both generation-specific and unified generation-understanding tasks. Experiments show that HoliTok achieves competitive reconstruction fidelity, improves generative learnability for high-quality and controllable synthesis, and, among the evaluated representations, is the *only* one that operates robustly in our unified generation-understanding architecture without additional optimization tricks. These results suggest that HoliTok serves as an effective speech tokenizer and a foundational representation interface for unified spoken language modeling. The code is available at: https://anonymous.4open.science/r/HoliTok-2852.

1 Introduction

Recent progress in multimodal foundation models is moving toward unified understanding and generation (Zeng et al., 2024; KimiTeam et al., 2025; Ge et al., 2025; Fan et al., 2025; Xie et al., 2025, 2026). Rather than treating downstream tasks separately, emerging systems seek to build all-in-one architectures that can understand, reason over, and generate within a shared parameter space. In the speech domain, this direction places a stronger requirement on the tokenizer: speech should be represented in a continuous space that is simultaneously decodable, learnable, and informative, so that it can serve as the interface for unified generation-understanding modeling. However, such a holistic continuous speech tokenizer remains underdeveloped. In its absence, downstream models must compensate through incremental architectural designs, such as task-specific encoders, multiple token streams, or decoupled modules. Consequently, the burden of unification is shifted from the representation itself to increasingly complex model design (Xu et al., 2025a,b; Yan et al., 2025).

Conventional acoustic front-end features, such as mel spectrograms, Fbank features, and MFCCs (Abdul and Al-Talabani, 2022), retain local signal structure, but they produce dense frame-level sequences that are redundant and difficult to model for downstream understanding and generation. In contrast, self-supervised speech representations (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022) expose richer semantic information, but they are not naturally decodable into high-fidelity waveforms and often present a challenging target for generative modeling. Thus, existing representations typically satisfy only part of the requirements for unified continuous speech modeling, leaving a gap between semantic abstraction, acoustic fidelity, and model learnability.

Current speech tokenizers address this challenge only partially. Discrete codec-based tokenizers (Défossez et al., 2022; Ji et al., 2025; Du et al., 2024; Guo et al., 2026) compress speech into language-model-friendly symbols, but quantization and multi-codebook designs may introduce information loss and additional modeling complexity. Continuous tokenizers (Li et al., 2025; Niu et al., 2025; Cheng et al., 2026) avoid quantization and are favorable for generation, yet many

are optimized mainly for reconstruction or synthesis rather than as a shared tokenization space for unified generation-understanding models. Existing "unified" representations (Dinkel et al., 2026; Yang et al., 2026a) are also often evaluated in task-specific systems separately, leaving the consistency of the shared modeling space unclear.

Recent AR+DiT architectures offer a simple downstream framework for unified continuous speech generation and understanding. For example, Ming-UniAudio (Yan et al., 2025) proposes MingTok-Audio to connect a compact variational autoencoder (VAE) latent with richer semantic features via an additional semantic module. While this improves tokenizer usability, the low-level latent remains fixed as higher-level semantics are introduced, resulting in an inconsistent modeling space and limited generative capacity.

In this work, we propose **HoliTok**, a **Holi**stic **Tok**enization model for unified continuous speech generation and understanding. HoliTok encodes 48 kHz speech into a compact 25 Hz sequence of 128-dimensional continuous latents. Its training follows a progressive recipe that gradually shapes a learnable and semantically informative latent space. We first train an autoencoder to ground the representation in faithful waveform reconstruction. We then introduce a sequence-aware variational bottleneck to regularize the latent distribution, making the sequence smoother and easier to predict while preserving signal-level fidelity. Finally, we strengthen variational regularization and refine the latent space through high-level feature distillation and audio-language supervision, enabling the resulting tokenization to retain information useful for spoken language understanding while remaining highly learnable for diverse speech synthesis tasks.

We build a unified generation-understanding model based on an AR+DiT architecture to evaluate whether a continuous speech tokenizer can serve as a unified modeling interface. The latent sequence is first encoded into patch embeddings for autoregressive modeling by the LLM. For generation, the LLM predicts semantic hidden states, which condition a DiT-based flow-matching head to predict the next latent patch. For understanding, the LLM predicts the next text token through an LM head. This evaluation is intentionally downstream-aware: beyond measuring reconstruction quality, it examines whether the tokenizer facilitates unified AR+DiT modeling.

We evaluate HoliTok from three complementary perspectives: reconstruction, speech synthesis, and unified generation-understanding modeling. Empirically, HoliTok achieves competitive reconstruction fidelity with a highly compact latent sequence, while supporting high-quality, diverse, and controllable TTS. In unified spoken language modeling, instantiated with ASR and TTS, HoliTok-Base already provides a more modeling-friendly continuous latent space than existing alternatives. HoliTok-Unite further improves both synthesis and recognition by incorporating the causal semantic encoder trained in the final stage, demonstrating substantially better usability than the baselines. These results show that HoliTok is not only an effective speech tokenizer, but also a principled representation interface that bridges the modeling-space gap between unified continuous speech understanding and generation.

2 Related Work

Audio representation for unified generation and understanding. Audio tokenization for unified modeling has been studied through both discrete and continuous representations. Discrete codecs and speech tokenizers (Défossez et al., 2022; Ji et al., 2025; Du et al., 2024), provide compact language-model-friendly units. Continuous tokenizers avoid quantization and have been explored in (Li et al., 2025; Niu et al., 2025; Dinkel et al., 2026; Yang et al., 2026a) for speech synthesis or unified audio modeling. These works improve different aspects of acoustic fidelity, semantic accessibility, and downstream usability. Compared with these works, HoliTok emphasizes holistic evaluation of the tokenization space within a single unified generation–understanding model, directly testing whether the same continuous representation is modelable as a shared interface for both speech generation and understanding.

Unified generation-understanding architecture with continuous tokens. Continuous-token architectures have recently emerged for unified generation and understanding. In vision, recent works (Fan et al., 2025; Xie et al., 2025, 2026) perform autoregressive generation and understanding with continuous visual tokens. Similar in audio, DiTAR (Jia et al., 2025) uses an autoregressive backbone with a DiT-based flow-matching head for continuous speech patches, and Ming-UniAudio (Yan et al., 2025) extends this idea to unified speech understanding, generation, and editing. It shows

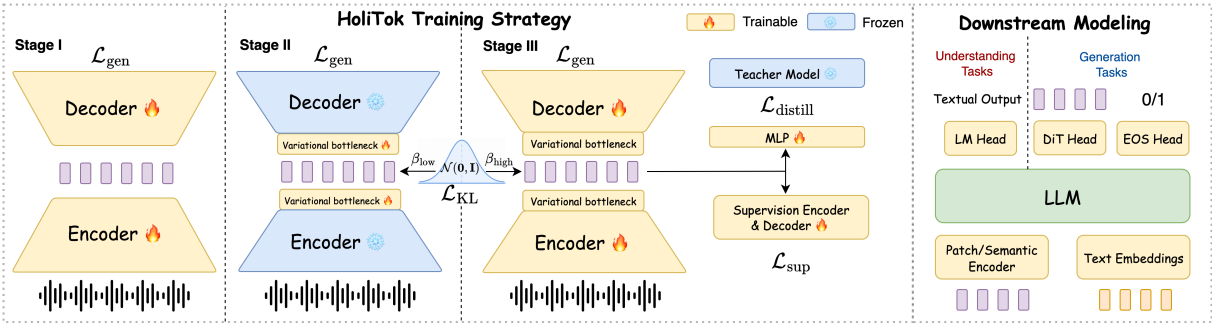

Figure 1: An Overview. Left side is the three-stage training strategy of HoliTok; Right side is our downstream architecture for unified generation-understanding tasks.

that continuous tokens can support unified modeling, but also make the representation space itself a bottleneck. Our work adopts the AR+DiT setting as a downstream-aware evaluation protocol and shows that HoliTok better balances generation and understanding under the same architecture.

3 Methodology

3.1 Main Architecture

HoliTok is a speech tokenizer built on a low-latency variational autoencoder backbone. We will introduce the model components in this section, and detailed configurations are posted in Appendix B.

Encoder. The encoder begins with a one-dimensional convolutional projection, followed by 6 strided causal convolutional downsampling blocks. Across these blocks, the channel width doubles from 12 to 768, with kernel sizes $4, 4, 4, 8, 12, 20$ and downsampling rates $2, 2, 2, 4, 6, 10$. This gives a total hop size of 1920, corresponding to a 25 Hz latent sequence for 48 kHz audio. Each downsampling block is followed by a residual stack of dilated causal convolutions. In our configuration, each stack contains 6 residual layers, which enlarge the receptive field while preserving causal processing. The final encoder projection maps the hidden sequence to a 128-dimensional acoustic representation. To improve reconstruction quality under a bounded-latency constraint, the encoder is causal except for a final 2-frame lookahead convolution.

Temporal variational bottleneck. On top of the convolutional encoder, we add bottleneck layers, consisting of a 4-layer LSTM block with project-in and -out linear layers. A 1×1 convolution then predicts the mean and log-scale of a diagonal Gaussian posterior, from which the latent sequence is sampled via the reparameterization trick. To increase the expressiveness of the latent distribution, we further apply a normalizing flow when computing the KL regularization against the standard normal prior. The sampled latent sequence is projected back to the model dimension and processed by a mirrored structure of encoder-side bottleneck before decoding.

Decoder. The decoder reconstructs the 48k Hz waveform from the 25 Hz latent sequence using a BigVGAN-style generator. Its upsampling module mirrors the encoder downsampling structure. Differently, following BigVGAN, each upsampling stage is refined by AMPBlocks with SnakeBeta activation. Similar to the encoder, the decoder introduces a 2-frame lookahead in its first convolutional net and is otherwise causal. The final projection maps the hidden features to a single-channel waveform.

Supervision network. The role of this component is detailed in Section 3.3. The supervision network follows an encoder–decoder design, consisting of a 0.6B Transformer encoder and a pretrained Qwen2.5-0.5B (Qwen et al., 2025) decoder. The encoder produces latent samples, then concatenated with task-label embeddings and fed into the language-model decoder.

3.2 Stage I&II: Progressive Training of High-fidelity Variational Latent Space

Empirically, imposing a strong KL constraint in VAE training can promote a more structured latent distribution, but it may also force the representation to discard acoustic details before the decoder has learned a high-fidelity reconstruction manifold. To mitigate this fidelity loss, we progressively shape the HoliTok latent space instead of learning it in a single stage. The overview is shown as Figure

1. Stage I trains a deterministic autoencoder to establish a high-fidelity acoustic autoencoding space. Stage II freezes the pretrained encoder and decoder, and converts this autoencoding space into a stochastic latent space by training only a temporal variational bottleneck with weak KL regularization. This staged procedure keeps the latent trajectory close to a reliable decoding region, providing a stable foundation for downstream-aware Stage III training. We further analyze this process as implicit fidelity transfer.

Stage I: reconstruction-oriented autoencoder pretraining. Given an input waveform $\mathbf{x}$, the encoder E_ϕ maps it to a low-rate acoustic representation, $\mathbf{z}_{\mathrm{AE}} = E_\phi(\mathbf{x})$, from which the decoder G_ψ reconstructs the waveform as $\hat{\mathbf{x}}_{\mathrm{AE}} = G_\psi(\mathbf{z}_{\mathrm{AE}})$. This stage is trained with a reconstruction-oriented generator objective:

$$\mathcal{L}_{\mathrm{I}} = \mathbb{E}_{\mathbf{x}} \left[\ell_{\mathrm{gen}}\big(\mathbf{x}, G_\psi(E_\phi(\mathbf{x}))\big) \right], \tag{1}$$

where ℓ_{gen} denotes the generator-side waveform generation loss, combining multi-scale spectral reconstruction, adversarial supervision, and discriminator feature matching:

$$\ell_{\mathrm{gen}} = \lambda_{\mathrm{spec}}\mathcal{L}_{\mathrm{spec}} + \lambda_{\mathrm{adv}}\mathcal{L}^{G}_{\mathrm{adv}} + \lambda_{\mathrm{fm}}\mathcal{L}_{\mathrm{fm}}. \tag{2}$$

Here, $\mathcal{L}_{\mathrm{spec}}$ is the multi-scale mel-spectral reconstruction loss, $\mathcal{L}^{G}_{\mathrm{adv}}$ is the generator-side adversarial loss, and $\mathcal{L}_{\mathrm{fm}}$ is the feature matching loss computed from discriminator intermediate activations. The discriminator objective is optimized in parallel and omitted for notational clarity. This stage establishes a high-fidelity reconstruction manifold before introducing variational regularization.

Stage II: autoencoding-to-variational latent transfer. Starting from the pretrained autoencoder, we freeze the encoder E_ϕ and decoder G_ψ, and train only the temporal variational bottleneck. Given the deterministic acoustic representation $\mathbf{z}_{\mathrm{AE}} = E_\phi(\mathbf{x})$, the bottleneck defines a posterior $q_\eta(\mathbf{z}_{\mathrm{VAE}}|\mathbf{z}_{\mathrm{AE}})$ over stochastic latents, which are sampled with the reparameterization trick and decoded by the frozen decoder. We optimize a reconstruction-dominated VAE objective:

$$\begin{aligned}\mathcal{L}_{\mathrm{II}} = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{z}_{\mathrm{VAE}}\sim q_\eta(\cdot|\mathbf{z}_{\mathrm{AE}})} \left[\ell_{\mathrm{gen}}\big(\mathbf{x}, G_\psi(\mathbf{z}_{\mathrm{VAE}})\big) \right] \\ + \beta_{\mathrm{low}} D_{\mathrm{KL}} \left(q_\eta(\mathbf{z}_{\mathrm{VAE}}|\mathbf{z}_{\mathrm{AE}}) \| p(\mathbf{z}) \right),\end{aligned} \tag{3}$$

where $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The small KL weight encourages distributional regularity without forcing the bottleneck to discard reconstruction-critical acoustic details. Since E_ϕ and G_ψ remain fixed, Stage II transfers the deterministic autoencoding space into a variational latent space while keeping sampled latents close to the decoder's high-fidelity reconstruction region.

Implicit fidelity transfer. The progressive Stage-I/II design provides an implicit fidelity-transfer effect. As formalized in Appendix A, the frozen pretrained decoder and the reconstruction-dominated objective constrain Stage-II variational samples to stay near the high-fidelity autoencoding manifold, so their expected waveform distortion is controlled by the Stage-I autoencoder distortion and the AE-to-VAE latent shift. This supports our choice to first learn a reliable decoding space and then train only the temporal variational bottleneck with a small KL weight, using the pretrained decoder as a fixed fidelity-preserving reference.

3.3 Stage III: Downstream-aware Enrichment of the Tokenization Space

After Stages I–II, the latent space has acquired high-fidelity reconstruction ability and initial variational regularity. However, reconstruction alone does not guarantee that the latent sequence preserves information required by downstream understanding tasks. In Stage III, we further enrich the VAE latent space with pretrained speech representations and task-conditioned supervision, making the tokenization space both waveform-decodable and informative for downstream speech-language modeling.

Multi-granularity representation distillation. We introduce multi-granularity representation distillation to enrich the VAE latent space beyond waveform reconstruction. Given $\mathbf{z} \sim q_\theta(\mathbf{z}|\mathbf{x})$, we align the latent sequence with frozen teacher representations from pretrained speech models at both frame and utterance levels. For frame-level distillation, following (Niu et al., 2025), we use WavLM (Chen et al., 2022) as a contextual teacher and apply a prediction head to map the latent sequence to its 23rd-layer hidden representations, with temporal interpolation used when frame rates differ. For utterance-level distillation, we aggregate the latent sequence into an utterance-level representation and align it with an x-vector speaker embedding (Desplanques et al., 2020). The unified

distillation objective is

$$\mathcal{L}_{\text{distill}} = \sum_{r\in\mathcal{R}} \lambda_r \left[1 - \cos\left(H_r(A_r(\mathbf{z})), \text{sg}(F_r(\mathbf{x}))\right)\right], \tag{4}$$

where $\mathcal{R}$ denotes the set of teacher representations, F_r is a frozen teacher, A_r performs temporal alignment for frame-level teachers or pooling for utterance-level teachers, H_r maps the adapted latent representation to the teacher space, and $\text{sg}(\cdot)$ denotes stop-gradient. For frame-level teachers, the cosine term is computed after temporal alignment and averaged over time.

Multi-task language-modeling supervision. We further expose the latent representation to downstream supervision through the task-conditioned supervision network described in Section 3.1. Given a task type $\tau \in \mathcal{T}$ and its target output $\mathbf{y}^\tau$, we optimize a unified language-modeling objective:

$$\mathcal{L}_{\text{sup}} = -\mathbb{E}_{(\mathbf{x},\tau,\mathbf{y}^\tau)}\mathbb{E}_{\mathbf{z}\sim q_\theta(\cdot|\mathbf{x})}\left[\log p_\omega(\mathbf{y}^\tau \mid \mathbf{z}, \tau)\right]. \tag{5}$$

This formulation converts heterogeneous downstream annotations into a shared task-conditioned prediction interface, covering tasks including speech recognition, emotion recognition, audio captioning, and sound event detection. As a result, the latent space is encouraged to retain information that may be unnecessary for waveform reconstruction but is critical for speech and audio understanding.

Combining waveform reconstruction, variational regularization, representation distillation, and downstream supervision, the Stage-III objective is

$$\mathcal{L}_{\text{III}} = \mathcal{L}_{\text{gen}} + \beta_{\text{high}}\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{distill}} + \lambda_{\text{sup}}\mathcal{L}_{\text{sup}}. \tag{6}$$

Here, $\mathcal{L}_{\text{gen}}$ denotes the expected generator-side waveform generation loss, $\mathcal{L}_{\text{KL}}$ regularizes the VAE posterior toward the standard normal prior, and β_{high} is much larger than the weak KL weight used in Stage II.

Variational interpretation. Stage III can be interpreted as optimizing a downstream-aware variational surrogate. Let $\mathbf{u}_r = F_r(\mathbf{x})$ denote a frozen teacher representation. We view the latent variable $\mathbf{z}$ as jointly explaining the waveform, teacher representations, and task target:

$$p(\mathbf{x}, \{\mathbf{u}_r\}_{r\in\mathcal{R}}, \mathbf{y}^\tau \mid \tau) = \int p(\mathbf{z}) p_\psi(\mathbf{x} \mid \mathbf{z}) p_\omega(\mathbf{y}^\tau \mid \mathbf{z}, \tau) \prod_{r\in\mathcal{R}} p_r(\mathbf{u}_r \mid \mathbf{z})\, d\mathbf{z}. \tag{7}$$

With the variational posterior $q_\theta(\mathbf{z} \mid \mathbf{x})$, this gives the weighted ELBO-style objective

$$\mathcal{J}_{\text{III}} = \mathbb{E}_{\mathbf{z}\sim q_\theta(\cdot|\mathbf{x})}\Big[\log p_\psi(\mathbf{x} \mid \mathbf{z}) + \lambda_{\text{sup}} \log p_\omega(\mathbf{y}^\tau \mid \mathbf{z}, \tau) + \sum_{r\in\mathcal{R}} \lambda_r \log p_r(\mathbf{u}_r \mid \mathbf{z})\Big] - \beta_{\text{high}} D_{\text{KL}}\left(q_\theta(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z})\right). \tag{8}$$

Minimizing $\mathcal{L}_{\text{III}}$ can therefore be viewed as maximizing this surrogate with practical waveform, distillation, supervision, and KL terms.

3.4 Downstream Unified Spoken Language Modeling

To evaluate whether the learned speech representation can serve as a unified modeling space, we build a downstream spoken language model that supports both speech understanding and speech generation with a shared backbone. Inspired by (Jia et al., 2025), the model follows an AR+DiT design: an autoregressive language model processes mixed text–audio embedding sequences, while a DiT-based flow-matching module predicts continuous latent patches for speech generation. The architecture overview is on right side of Figure 1.

Speech understanding objective. Let $\mathbf{z}_{\text{audio}}$ denote the audio latent patches and $\mathbf{e}_{\text{audio}}$ denote their corresponding language-model embeddings. Given textual context $\mathbf{c}$ and target text $\mathbf{y}_{\text{text}}$, we optimize an autoregressive cross-entropy objective:

$$\mathcal{L}_{\text{understand}} = -\sum_j \log p_\theta\left(y_j \mid \mathbf{y}_{<j}, \mathbf{e}_{\text{audio}}, \mathbf{c}\right). \tag{9}$$

Speech generation objective. For speech generation, the autoregressive language model summarizes the available text and audio history into causal hidden states, and the DiT flow-matching module predicts each future latent patch conditioned on this previous hiddens and historical latents with an autoregressive pattern following (Liu et al., 2024). The conditional generation process is factorized as

$$p_\theta\left(\mathbf{z}_{1:K} \mid \mathbf{c}\right) = \prod_{k=1}^{K} p_\theta\left(\mathbf{z}_k \mid \mathbf{h}_{\le k}, \mathbf{z}_{<k}\right), \tag{10}$$

where $\mathbf{h}_{\le k}$ is the causal language-model hidden states sequence for k patches prediction, and $\mathbf{z}_{<k}$ denotes previously generated audio latent patches. Each conditional patch distribution is learned with a flow-matching objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{k,t}\left[\left\|v_\theta\left(\mathbf{z}_{k,t}, t \mid \mathbf{h}_{\le k}, \mathbf{z}_{<k}\right) - \mathbf{u}_{k,t}\right\|_2^2\right], \tag{11}$$

Model	CR	TPS	NB/WB PESQ ↑	STOI ↑	WER(%) ↓	SPKSIM ↑	EMOSIM ↑	UTMOS ↑
Ground Truth	1.00×	–	–	–	3.91	1.000	1.000	3.75
Mel Spectrogram	2.00×	86	4.15/4.05	**0.988**	**3.96**	0.957	0.988	3.75
SemanticVAE	2.73×	40	3.99/3.80	0.969	4.15	0.963	0.993	**3.76**
MingTok-Audio	2.19×	50	**4.23/4.12**	0.981	4.27	0.950	0.992	3.75
Vanilla VAE	**7.5**×	25	3.18/2.65	0.925	5.41	0.859	0.988	3.75
HoliTok			4.10/4.01	0.974	4.22	**0.968**	**0.995**	3.75

Table 1: Reconstruction evaluation results on LibriSpeech test-other.

where $\mathbf{z}_{k,t}$ is the interpolated noisy state of the k-th latent patch at timestamp t, and $\mathbf{u}_{k,t}$ is the corresponding target velocity. We further supervise audio termination with a binary cross-entropy EOS loss:

$$\mathcal{L}_{\text{generate}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{eos}}\mathcal{L}_{\text{eos}}. \tag{12}$$

The generated latent patches are assembled into a latent sequence and decoded into waveform audio by the frozen HoliTok decoder.

4 Experiments

4.1 Experimental settings and Baselines

Training datasets. We train HoliTok on a mixture of speech, environmental sound, and music data. The speech data include AISHELL-3 (Shi et al., 2021), HiFi-TTS (Bakhturina et al., 2021), VCTK (Yamagishi et al., 2019), HiFiTTS2 (Langman et al., 2025), and large-scale internal English and Chinese TTS corpora, totaling approximately 500K hours. To improve robustness beyond clean read speech, we further include emotional speech data, AudioSet (Gemmeke et al., 2017), VGGSound (Chen et al., 2020), VocalSound (Gong et al., 2022), FSD50K (Fonseca et al., 2022), MusicCaps (Agostinelli et al., 2023), and WavCaps (Mei et al., 2024).

Training settings. All audio is resampled to 48 kHz for HoliTok training. The generator is trained with a multi-period discriminator and a multi-scale sub-band CQT discriminator, following the BigVGAN v2 configuration (gil Lee et al., 2023). As described in Section 3.2, training proceeds in three stages. We first train the autoencoder backbone for 500K steps. We then train the variational bottleneck for 50K steps with $\beta_{\text{low}} = 0.1$. In the final stage, we train the full model with the supervision network for 200K steps using $\beta_{\text{high}} = 7$. Both the generator and discriminator are optimized with AdamW (Loshchilov and Hutter, 2019), using an initial learning rate of 1×10^{-4}, betas $(0.8, 0.99)$, and $\epsilon = 10^{-6}$. The learning rate is exponentially decayed to 1×10^{-6}. Additional configurations are provided in Appendix B.

Main baselines. We compare HoliTok with two representative continuous audio representations. Semantic-VAE (Niu et al., 2025) distills pretrained SSL representations into VAE latents and has shown strong performance for DiT-based speech synthesis over mel-spectrogram inputs. MingTok-Audio is a continuous speech tokenizer designed for AR+DiT-based unified speech understanding and generation. For MingTok-Audio, we use its unified feature as the input representation and its acoustic latent as the generation target, while keeping the semantic module fixed following its reported ablation protocol (Yan et al., 2025).

4.2 Reconstruction Evaluation

We evaluate reconstruction quality on LibriSpeech (Panayotov et al., 2015) test-other in terms of signal fidelity, linguistic preservation, and paralinguistic consistency. We report narrow-band and wide-band PESQ (Rix et al., 2001), STOI (Taal et al., 2010), and UTMOS (Saeki et al., 2022) for perceptual quality and intelligibility; WER on resynthesized speech for linguistic preservation; and speaker similarity(SPKSIM) (Desplanques et al., 2020) and emotion similarity(EMOSIM) (Ma et al., 2024) for paralinguistic consistency. Ground-truth waveforms are used as references for signal-level metrics. We compare HoliTok with BigVGAN v2 mel-spectrogram vocoding[1], directly trained VAE and main baselines.

We also report tokens per second (TPS) and compression ratio (CR). CR is computed as the ratio between the raw waveform bitrate and the latent representation bitrate:

$$\text{CR} = (f_s \lceil \log_2 f_s \rceil)/(f_z d_z b_{\text{float}}), \tag{13}$$

[1] BigVGAN v2 checkpoint

where f_s is the waveform sampling rate, $\lceil \log_2 f_s \rceil$ is the number of bits used for each waveform sample, f_z is the latent frame rate, d_z is the latent dimension, and $b_{\text{float}} = 32$ is the number of bits per floating-point latent value.

As shown in Table 1, HoliTok achieves competitive reconstruction quality among continuous speech representations while using the most compact latent sequence, with a compression ratio of 7.5× and 25 TPS. Although mel-spectrogram vocoding and MingTok-Audio obtain slightly higher scores on some signal-level metrics, HoliTok preserves linguistic and paralinguistic information well, achieving strong WER, the best SPKSIM, and the best EMOSIM. Compared with the vanilla VAE using the same architecture and compression rate, HoliTok substantially improves PESQ, STOI, WER, and SPKSIM, validating the effectiveness of the progressive training strategy.

4.3 Evaluation on Speech Synthesis

Speech synthesis directly tests whether a representation is learnable as a generation target. We use the generation branch of the AR+DiT model in Section 3.4. A base TTS model is trained on 95K hours of filtered Emilia (He et al., 2025) data for 200k steps, and then further tuned 50k steps for controllable TTS with EmoVoice-DB (Yang et al., 2025), FCaps (Yang et al., 2026b), and PSCBase (Diwan et al., 2025). Details are given in Appendix B.

Zero-shot capability and synthesis diversity. We evaluate the base model in a zero-shot setting, synthesizing unseen-speaker speech from prompt speech and text. On Seed-TTS-Eval (Anastassiou et al., 2024), we report WER and speaker similarity for intelligibility and speaker preservation, respectively. We further evaluate the emotion and paralinguistic subsets of Emergent-TTS (Manku et al., 2026), reporting WER and win rate against GPT-4o-mini-TTS (OpenAI et al., 2024). Table 2 shows that HoliTok achieves competitive zero-shot TTS performance and obtains the highest win rates on both expressive dimensions, suggesting that its latent space is highly learnable and preserves expressive paralinguistic information.

Controllable TTS. We evaluate the fine-tuned TTS model on controllable synthesis, where speech is generated from explicit emotional or paralinguistic descriptions. On EmoVoiceDB-test, we report WER and EMOSIM for content consistency and emotion control. On FCaps-test, we report WER and CLSP (Yang et al., 2026b) score to measure alignment with fine-grained speaking-style descriptions. As shown in Figure 2, HoliTok achieves the best WER on both datasets, matches the best EMOSIM, and obtains the highest CLSP score, indicating stronger controllability without sacrificing intelligibility.

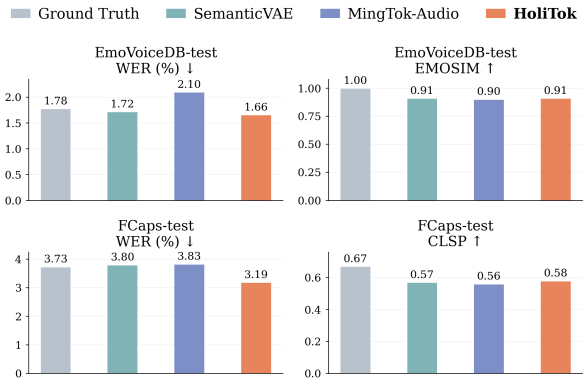

Figure 2: Controllable TTS evaluation on EmoVoiceDB-test and FCaps-test.

4.4 Evaluation on Unified Understanding and Generation

Settings. We use the AR+DiT architecture in Section 3.4 for unified spoken language modeling, instantiating understanding as ASR and generation as TTS. HoliTok-Base uses the learned VAE latents as the audio representation; its non-causal supervision encoder is used only during representation training. HoliTok-Unite uses the causal supervision encoder trained in Stage III as a built-in semantic encoder, replacing the downstream patch encoder and providing pre-modeled speech features, similar in spirit to MingTok-Audio. We train the unified model with Emilia for TTS and AISHELL-1/2 (Bu et al., 2017; Du et al., 2018), GigaSpeech (Chen et al., 2021), MLS (Pratap et al., 2020), Common Voice 20.0 (Ardila et al., 2020), FLEURS (Conneau et al., 2023), and LibriSpeech (Panayotov et al., 2015) for ASR, using a sampler that keeps the TTS-to-ASR ratio near 5:1. We evaluate TTS on Seed-TTS-Eval and ASR on LibriSpeech test-clean/test-other and AISHELL-1 test.

Evaluation results analysis. Table 3 shows that unified ASR–TTS training is substantially more demanding than task-specific modeling in Table 2. Under the same AR+DiT architecture, existing continuous representations degrade sharply on TTS, indicating that reconstruction quality or isolated downstream performance does not necessar-

Model	Seed-TTS-Eval						Emergent-TTS			
	Seed-TTS-en		Seed-TTS-zh		Seed-TTS-hard		Emotion		Paralinguistic	
	WER(%) ↓	SIM ↑	WER(%) ↓	SIM ↑	WER(%) ↓	SIM ↑	WER(%) ↓	Win-Rate ↑	WER(%) ↓	Win-Rate ↑
Semantic-VAE	1.42	**0.63**	**0.91**	0.70	**7.53**	**0.67**	**0.63**	14.3	**34.45**	44.2
MingTok-Audio	1.84	0.61	1.03	**0.71**	14.75	0.66	8.81	8.4	35.66	39.8
HoliTok	**1.33**	0.62	0.98	0.70	7.59	0.66	1.34	**25.5**	34.47	**53.6**

Table 2: Zero-shot TTS evaluation on Seed-TTS-Eval and Emergent-TTS."SIM" refers to the speaker similarity between synthesized and prompt speech.

Model	TTS						ASR		
	Seed-TTS-en		Seed-TTS-zh		Seed-TTS-hard		test-clean	test-other	AISHELL-1
	WER(%) ↓	SIM ↑	WER(%) ↓	SIM ↑	WER(%) ↓	SIM ↑	WER(%) ↓	WER(%) ↓	WER(%) ↓
Semantic-VAE	102.32	0.47	99.30	0.61	97.31	0.61	9.69	21.32	15.81
MingTok-Audio	51.06	0.42	18.17	0.61	50.35	0.55	**4.62**	**9.06**	**5.01**
HoliTok-Base	27.85	0.52	4.40	0.66	30.44	0.61	6.45	16.51	14.92
HoliTok-Unite	**7.20**	**0.55**	**1.78**	**0.67**	**16.79**	**0.62**	5.48	12.65	5.93

Table 3: Unified spoken language modeling evaluation on TTS and ASR tasks.

ily translate to usability in a shared generation-understanding model. Within this unified setting, HoliTok shows a better balance between generation and understanding. HoliTok-Base already outperforms the baselines on all TTS intelligibility and achieves comparable ASR results. This suggests that the proposed VAE latent space remains more learnable as a continuous generation target while preserving sufficient acoustic information. With the causal semantic encoder, HoliTok-Unite further reduces the average TTS WER from 20.90% to 8.59% and improves the average ASR WER from 12.63% to 8.02% over HoliTok-Base. These gains indicate that the Stage-III causal encoder provides useful pre-learning of HoliTok representations, rather than merely improving an isolated understanding branch. The comparison also reveals different failure modes of existing representations. Semantic-VAE obtains usable ASR performance but fails on TTS, suggesting that directly shaping the latent space toward semantic representations can weaken its generative learnability. MingTok-Audio achieves the best ASR WER, but its TTS performance remains much weaker than HoliTok-Unite, indicating an imbalance toward understanding. Overall, HoliTok better satisfies the joint requirements of unified spoken language modeling: acoustic preservation for generation, semantic accessibility for understanding, and latent learnability under a shared AR+DiT backbone.

Ablation study. We provide complete ablation results in Appendix C, showing that the proposed training components and downstream modeling choices are complementary. Only high variational regularization alone is insufficient for a generation-friendly unified representation. Using representation distillation without supervision also severely degrades synthesis, consistent with the Semantic-VAE results, while supervision alone preserves much stronger TTS performance. On the downstream side, DiT initialization with TTS-only training consistently improves generation quality, and HoliTok-Unite performs best when the causal semantic encoder remains trainable rather than frozen.

5 Conclusion

We presented HoliTok, a holistic speech tokenizer for both generation-oriented and unified generation–understanding tasks. Through progressive training, HoliTok combines compact high-fidelity reconstruction, sequence-aware variational regularization, and downstream-aware semantic enrichment, yielding a tokenization that remains detokenizable, learnable, and informative. Experiments on reconstruction, zero-shot and controllable TTS, and unified ASR–TTS modeling demonstrate that HoliTok serves as an effective interface for speech compression, diverse speech synthesis, and unified spoken-language modeling. Comprehensive analyses further show that HoliTok achieves robust performance without relying on complex architectural modifications or incremental training mechanisms.

Limitations

This work has two main limitations, both of which point to natural directions for future research. First, our current study focuses on speech-centered generation and understanding. Although HoliTok is designed as a continuous audio representation, our experiments mainly cover speech reconstruction, text-to-speech synthesis, and automatic speech recognition. We have not yet systematically evaluated whether the same latent space can generalize to broader audio domains such as environmental sound and music. These domains may require different temporal abstractions, perceptual objectives, and semantic supervision signals. Extending HoliTok from speech to general audio and music modeling is therefore an important direction for future work. Second, our downstream evaluation is built on a unified AR+DiT architecture. This setting directly tests whether the learned representation can serve as a shared interface for both speech generation and understanding, but it does not exhaust all possible unified modeling paradigms. In particular, we have not explored pure DiT-based or fully non-autoregressive architectures for unified generation-understanding modeling. Future work can study how HoliTok interacts with different backbone designs, and whether the proposed representation remains robust across alternative generative and understanding architectures. Potential risks, artifact documentation, computational experiment details, and AI-assistant use are further discussed in Appendices B and D.

References

Zrar Kh. Abdul and Abdulbasit K. Al-Talabani. 2022. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. Musiclm: Generating music from text. *Preprint*, arXiv:2301.11325.

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, and 27 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *Preprint*, arXiv:2406.02430.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-Fi Multi-Speaker English TTS Dataset. In *Interspeech 2021*, pages 2776–2780.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Interspeech 2021*, pages 3670–3674.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725.

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. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Changhao Cheng, Wei Wang, Wangyou Zhang, Dongya Jia, Jian Wu, Zhuo Chen, and Yanmin Qian. 2026. On the distillation loss functions of speech vae for unified reconstruction, understanding, and generation. *Preprint*, arXiv:2604.12383.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, pages 3830–3834.

Heinrich Dinkel, Xingwei Sun, Gang Li, Jiahao Mei, Yadong Niu, Jizhong Liu, Xiyang Li, Yifan Liao, Jiahao Zhou, Junbo Zhang, and Jian Luan. 2026. Dashengtokenizer: One layer is enough for unified audio understanding and generation. *Preprint*, arXiv:2602.23765.

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. 2025. Scaling rich style-prompted text-to-speech datasets. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3639–3659, Suzhou, China. Association for Computational Linguistics.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. Aishell-2: Transforming mandarin asr research into industrial scale. *Preprint*, arXiv:1808.10583.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

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

Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, Michael Rubinstein, Michalis Raptis, Deqing Sun, and Radu Soricut. 2025. Unified autoregressive visual generation and understanding with continuous tokens. *Preprint*, arXiv:2503.13436.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. Fsd50k: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2025. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *Preprint*, arXiv:2404.14396.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. BigVGAN: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 151–155. IEEE.

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. 2026. Recent advances in discrete speech tokens: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4184–4204.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2025. Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4044–4054.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. 2025. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In *The Thirteenth International Conference on Learning Representations*.

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, and Yuxuan Wang. 2025. DiTAR: Diffusion transformer autoregressive modeling for speech generation. In *Forty-second International Conference on Machine Learning*.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

Ryan Langman, Xuesong Yang, Paarth Neekhara, Shehzeen Hussain, Edresson Casanova, Evelina Bakhturina, and Jason Li. 2025. HiFiTTS-2: A Large-Scale High Bandwidth Speech Dataset. In *Interspeech 2025*, pages 4778–4782.

Yixing Li, Ruobing Xie, Xingwu Sun, Yu Cheng, and Zhanhui Kang. 2025. Continuous speech tokenizer in text to speech. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3341–3347, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhijun Liu, Shuai Wang, Sho Inoue, Qibing Bai, and Haizhou Li. 2024. Autoregressive diffusion transformer for text-to-speech synthesis. *arXiv preprint arXiv:2406.05551*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand. Association for Computational Linguistics.

Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. 2026. EmergentTTS-eval: Evaluating TTS models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354.

Zhikang Niu, Shujie Hu, Jeongsoo Choi, Yushen Chen, Peining Chen, Pengcheng Zhu, Yunting Yang, Bowen Zhang, Jian Zhao, Chunhui Wang, and 1 others. 2025. Semantic-vae: Semantic-alignment latent representation for better speech synthesis. *arXiv preprint arXiv:2509.22167*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757–2761.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech 2022*, pages 4521–4525.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. In *Interspeech 2021*, pages 2756–2760.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2025. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations*.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. 2026. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

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, and 6 others. 2025. Ming-uniaudio: Speech llm for joint understanding, generation and editing with unified representation. *Preprint*, arXiv:2511.05516.

Configuration	HoliTok-Base	HoliTok-Unite	SemanticVAE	MingTok-Audio
Audio representation	VAE latent	semantic feature	VAE latent	semantic feature
Input dimension	128	1536	64	1280
Temporal aggregation	patch size 4	mean pooling size 4	patch size 4	mean pooling size 8
Audio-to-LLM mapping	8-layer PatchEncoder	linear projection	8-layer PatchEncoder	linear projection
Audio-to-LLM params	102M	1M	102M	1M
LLM backbone	Qwen2.5-0.5B	Qwen2.5-0.5B	Qwen2.5-0.5B	Qwen2.5-0.5B
LLM hidden size	896	896	896	896
LLM params	494M	494M	494M	494M
DiT layers	18	18	18	18
DiT hidden size	1024	1024	1024	1024
DiT attention heads	16	16	16	16
DiT FFN hidden size	4096	4096	4096	4096
DiT input/output dim.	1024 / 128	1024 / 128	1024 / 64	1024 / 64
DiT params	345M	345M	345M	345M
Bridge / EOS params	2M	2M	2M	2M
Total downstream params	942M	842(+680)M	942M	841M

Table 4: Downstream AR+DiT architecture configuration and parameter counts. Tokenizer-side modules in Table 6 are not included.

Guanrou Yang, Tian Tan, Qian Chen, Zhikang Niu, Yakun Song, Ziyang Ma, Yushen Chen, Zeyu Xie, Tianrui Wang, Yifan Yang, Wenxi Chen, Qi Chen, Wenrui Liu, Shan Yang, and Xie Chen. 2026a. Wavcube: Unifying speech representation for understanding and generation via semantic-acoustic joint modeling. *Preprint*, arXiv:2605.06407.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, Fan Yu, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2025. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 10748–10757, New York, NY, USA. Association for Computing Machinery.

Yifan Yang, Bing Han, Hui Wang, Wei Wang, Ziyang Ma, Long Zhou, Zengrui Jin, Guanrou Yang, Tianrui Wang, Xu Tan, and Xie Chen. 2026b. Towards fine-grained and multi-granular contrastive language-speech pre-training. *Preprint*, arXiv:2601.03065.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *Preprint*, arXiv:2412.02612.

A Implicit Fidelity Transfer Formulation

Proposition 1. *Let* $\epsilon_{\mathrm{AE}} = \mathbb{E}_{\mathbf{x}}[\|\mathbf{x} - G_\psi(\mathbf{z}_{\mathrm{AE}})\|_2^2]$ *denote the waveform reconstruction distortion of the Stage-I autoencoder. Assume that the frozen decoder* G_ψ *is locally* L_ψ*-Lipschitz in a neighborhood containing both* $\mathbf{z}_{\mathrm{AE}}$ *and the variational samples* $\mathbf{z}_{\mathrm{VAE}}$*. Define the AE-to-VAE latent shift as:*

$$\delta_{\mathrm{shift}} = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{z}_{\mathrm{VAE}}\sim q_\eta(\cdot|\mathbf{z}_{\mathrm{AE}})}\left[\|\mathbf{z}_{\mathrm{VAE}} - \mathbf{z}_{\mathrm{AE}}\|_2^2\right]. \tag{14}$$

Then the expected waveform distortion of the variational latent satisfies

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{z}_{\mathrm{VAE}}\sim q_\eta(\cdot|\mathbf{z}_{\mathrm{AE}})}&\left[\|\mathbf{x} - G_\psi(\mathbf{z}_{\mathrm{VAE}})\|_2^2\right] \\ &\le 2\epsilon_{\mathrm{AE}} + 2L_\psi^2\delta_{\mathrm{shift}}.\end{aligned} \tag{15}$$

Proof. For compactness, denote $\hat{\mathbf{x}}_{\mathrm{AE}} = G_\psi(\mathbf{z}_{\mathrm{AE}})$ and $\hat{\mathbf{x}}_{\mathrm{VAE}} = G_\psi(\mathbf{z}_{\mathrm{VAE}})$. For any input $\mathbf{x}$, by adding and subtracting $\hat{\mathbf{x}}_{\mathrm{AE}}$, we have

$$\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{VAE}} = \mathbf{x} - \hat{\mathbf{x}}_{\mathrm{AE}} + \hat{\mathbf{x}}_{\mathrm{AE}} - \hat{\mathbf{x}}_{\mathrm{VAE}}. \tag{16}$$

Using $\|\mathbf{a}+\mathbf{b}\|_2^2 \le 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$, we obtain

$$\begin{aligned}\|\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{VAE}}\|_2^2 \le{}& 2\,\|\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{AE}}\|_2^2 \\ &+ 2\,\|\hat{\mathbf{x}}_{\mathrm{AE}} - \hat{\mathbf{x}}_{\mathrm{VAE}}\|_2^2.\end{aligned} \tag{17}$$

By the local L_ψ-Lipschitz continuity of G_ψ,

$$\|\hat{\mathbf{x}}_{\mathrm{AE}} - \hat{\mathbf{x}}_{\mathrm{VAE}}\|_2^2 \le L_\psi^2\,\|\mathbf{z}_{\mathrm{AE}} - \mathbf{z}_{\mathrm{VAE}}\|_2^2. \tag{18}$$

Symbol	Template
$\mathcal{P}_{\mathrm{tts}}$	[text] $\mathbf{t}$ [text2speech] ⟨loss⟩$\mathbf{a}$⟨eos⟩⟨/loss⟩
$\mathcal{P}_{\mathrm{asr}}$	[speech] $\mathbf{a}$ [speech2text] ⟨loss⟩$\mathbf{t}$⟨eos⟩⟨/loss⟩
$\mathcal{D} \oplus \mathcal{P}_{\mathrm{tts}}$	[desc] $\mathbf{d} \oplus \mathcal{P}_{\mathrm{tts}}$

Table 5: Symbolic task templates used in downstream training. $\mathbf{t}$, $\mathbf{a}$, and $\mathbf{d}$ denote text, audio latent sequence, and description instruction, respectively.

Component	HoliTok-Base	HoliTok-Unite
Encoder	36M	36M
Decoder	128M	128M
Variational bottleneck	17M	17M
Semantic encoder	–	680M
Total	181M	861M

Table 6: Parameter counts of tokenizer-side representation modules used in downstream modeling.

Taking expectation over $\mathbf{x}$ and $\mathbf{z}_{\mathrm{VAE}} \sim q_\eta(\cdot|\mathbf{z}_{\mathrm{AE}})$ gives

$$\mathbb{E}_{\mathbf{x}}\mathbb{E}_{q_\eta}\left[\|\mathbf{x} - G_\psi(\mathbf{z}_{\mathrm{VAE}})\|_2^2\right] \leq 2\epsilon_{\mathrm{AE}} + 2L_\psi^2\delta_{\mathrm{shift}}.$$

□

B Experimental Setting and Responsible Use Details

Tokenizer configuration and training settings. Table 6 reports the tokenizer-side parameters. And Table 7 summarizes the optimizer, scheduler, and loss weights used for tokenizer training. All audio is resampled to 48 kHz. The generator is trained with a multi-period discriminator and a multi-scale sub-band CQT discriminator, following the BigVGAN V2 configuration. In Stages I–II, training uses 9.6-second cropped audio segments; in Stage III, the per-GPU batch size is set to 1 to support downstream supervision.

Downstream configuration and training settings. Table 4 summarizes the downstream AR+DiT configuration. HoliTok-Base maps VAE latent patches to the LLM hidden space using an 8-layer PatchEncoder. HoliTok-Unite mean-pools semantic features over each patch and uses a lightweight linear projection before the shared LLM backbone and DiT predictor. For downstream AR+DiT training, all settings use AdamW with learning rate 1×10^{-4}, betas $(0.9, 0.99)$, $\epsilon = 1\times10^{-6}$, bf16 precision, and gradient clipping of 2. The learning rate follows a cosine scheduler with 5000 warmup batches and a minimum learning rate of 1×10^{-5}. The TTS-only setting uses $\mathcal{P}_{\mathrm{tts}}$, controllable TTS prepends a description instruction as $\mathcal{D} \oplus \mathcal{P}_{\mathrm{tts}}$, and unified ASR–TTS uses $\mathcal{P}_{\mathrm{tts}}$ for generation and $\mathcal{P}_{\mathrm{asr}}$ for recognition. The symbolic templates are defined in Table 5.

Setting	Value
Optimizer	
Optimizer	AdamW
Initial learning rate	1×10^{-4}
Betas	$(0.8, 0.99)$
Epsilon	1×10^{-6}
Gradient clipping	500
Scheduler	
Scheduler	exponential decay
Final learning rate floor	1×10^{-6}
Warmup steps	1
Decay rate	0.9999996
Loss weights	
Adversarial generator loss	1.0
Feature matching loss	2.0
Multi-scale mel loss	45.0
KL loss, Stage II	$\beta_{\mathrm{low}} = 0.1$
KL loss, Stage III	$\beta_{\mathrm{high}} = 7.0$
WavLM distillation loss	1.0
X-vector distillation loss	1.0
Supervision CE loss	1.0

Table 7: Tokenizer optimization settings and loss weights. HoliTok-Base and HoliTok-Unite use the same recipe, except that the supervision encoder is non-causal for HoliTok-Base and causal for HoliTok-Unite.

Potential risks. Because HoliTok supports high-quality speech generation, it may be misused for voice impersonation, spoofing, or misleading synthetic speech if deployed without safeguards. The intended use of the released artifacts is research on speech tokenization and unified spoken language modeling. Practical deployments should include consent-aware data policies, provenance or watermarking mechanisms for generated audio when appropriate, and restrictions against impersonation or deceptive use.

Scientific artifacts, licenses, and intended use. This work uses public speech and audio datasets, pretrained model components, baseline tokenizers, and evaluation tools as scientific artifacts[2][3][4]. and artifact creators are cited in Section 4, and the training data mixture and descriptive statistics are summarized there. Third-party datasets and models should be used according to their original licenses

[2] emotion2vec checkpoint.
[3] speaker embedding checkpoint.
[4] CLSP checkpoint.

Model	Ablation setting	TTS						ASR		
		Seed-TTS-en		Seed-TTS-zh		Seed-TTS-hard		test-clean	test-other	AISHELL-1
		WER ↓	SIM ↑	WER ↓	SIM ↑	WER ↓	SIM ↑	WER ↓	WER ↓	WER ↓
SemanticVAE	default	102.32	0.47	99.30	0.61	97.31	0.61	9.69	21.32	15.81
	DiT init	34.12	0.60	4.10	0.70	29.63	0.66	10.41	25.52	16.58
MingTok-Audio	default	51.06	0.42	18.17	0.61	50.35	0.55	4.62	9.06	5.01
	DiT init	14.53	0.58	3.14	0.69	28.21	0.64	4.32	10.41	5.25
HoliTok-Base	default	27.85	0.52	4.40	0.66	30.44	0.61	6.45	16.51	14.92
	DiT init	6.63	0.59	1.71	0.69	16.97	0.66	6.39	15.92	14.32
	w/o distill	22.80	0.51	3.65	0.65	19.89	0.61	6.27	16.04	15.23
	w/o supervise	110.05	0.39	97.56	0.60	96.81	0.60	6.44	16.59	16.63
	w/o. both	103.14	0.37	98.11	0.58	97.24	0.57	6.12	16.00	14.99
HoliTok-Unite	default	7.20	0.55	1.78	0.67	16.79	0.62	5.48	12.65	5.93
	DiT init	3.64	0.59	1.13	0.69	12.53	0.65	5.32	13.12	7.17
	freeze semantic encoder	8.46	0.56	1.86	0.66	15.60	0.63	6.44	15.92	7.42

Table 8: Complete ablation results for unified spoken language modeling. TTS is evaluated by WER and SIM on Seed-TTS subsets, and ASR is evaluated by WER on LibriSpeech test-clean/test-other and AISHELL-1. "default" denotes the standard unified training setting for each representation. "DiT init" initializes the DiT predictor from a TTS-specialized checkpoint. For HoliTok-Base, "w/o distill", "w/o supervise", and "w/o both" remove representation distillation, multi-task supervision, and both objectives in Stage III, respectively. For HoliTok-Unite, "freeze semantic encoder" keeps the causal semantic encoder fixed during downstream training.

and terms of use; internally collected corpora are used only for training and are not redistributed. The released code and checkpoints are intended for research use and will include documentation describing model usage, expected inputs and outputs.

C Complete Ablation Results

Table 8 provides a detailed ablation of the unified spoken language modeling setting. The DiT initialization rows show that a TTS-oriented initialization substantially improves generation across all representations, especially for the two baseline tokenizers whose default unified training yields high TTS WER. This confirms that the downstream DiT head is a major factor for continuous-latent speech generation, but it does not by itself guarantee balanced understanding performance: for example, the initialized HoliTok-Unite improves TTS WER but degrades AISHELL-1 ASR compared with its default setting.

For HoliTok-Base, removing distillation improves several TTS WER scores but slightly weakens speaker similarity and does not improve ASR consistently, suggesting that representation distillation mainly contributes semantic and paralinguistic information rather than pure generation ease. Removing supervision severely degrades TTS while increasing AISHELL-1 WER, indicating that downstream supervision is important for both generation robustness and cross-lingual understanding ability. Removing both distillation and supervision similarly weakens TTS performance, further confirming that downstream-aware enrichment is necessary for a holistic tokenizer. For HoliTok-Unite, freezing the semantic encoder weakens ASR on all three test sets and gives mixed TTS changes, showing that adapting the semantic interface during unified training is important for balancing generation and recognition. Overall, the ablations indicate that strong unified modeling requires both a learnable continuous latent space and task-aware adaptation of the semantic and DiT components.

D Use of AI Assistants

AI assistants were used to support writing and editing tasks, including grammar checking, wording refinement, and LaTeX formatting. The authors reviewed and edited all AI-assisted text and retained responsibility for the scientific claims, experimental design, analysis, and final manuscript.