

# LongCat-AudioDiT: High-Fidelity Diffusion Text-to-Speech in the Waveform Latent Space

Meituan LongCat Team  
longcat-team@meituan.com

## ABSTRACT

We present LongCat-AudioDiT, a novel, non-autoregressive diffusion-based text-to-speech (TTS) model that achieves state-of-the-art (SOTA) performance. Unlike previous methods that rely on intermediate acoustic representations such as mel-spectrograms, the core innovation of LongCat-AudioDiT lies in operating directly within the waveform latent space. This approach effectively mitigates compounding errors and drastically simplifies the TTS pipeline, requiring only a waveform variational autoencoder (Wav-VAE) and a diffusion backbone. Furthermore, we introduce two critical improvements to the inference process: first, we identify and rectify a long-standing training-inference mismatch; second, we replace traditional classifier-free guidance with adaptive projection guidance to elevate generation quality. Experimental results demonstrate that, despite the absence of complex multi-stage training pipelines or high-quality human-annotated datasets, LongCat-AudioDiT achieves SOTA zero-shot voice cloning performance on the Seed benchmark while maintaining competitive intelligibility. Specifically, our largest variant, LongCat-AudioDiT-3.5B, outperforms the previous SOTA model (Seed-TTS), improving the speaker similarity (SIM) scores from 0.809 to 0.818 on Seed-ZH, and from 0.776 to 0.797 on Seed-Hard. Finally, through comprehensive ablation studies and systematic analysis, we validate the effectiveness of our proposed modules. Notably, we investigate the interplay between the Wav-VAE and the TTS backbone, revealing the counterintuitive finding that superior reconstruction fidelity in the Wav-VAE does not necessarily lead to better overall TTS performance. Code and model weights are released to foster further research within the speech community.

**Github:**<https://github.com/meituan-longcat/LongCat-AudioDiT>

**HuggingFace:**

<https://huggingface.co/meituan-longcat/LongCat-AudioDiT-3.5B>

<https://huggingface.co/meituan-longcat/LongCat-AudioDiT-1B>

## 1 Introduction

Text-to-speech (TTS) synthesis is a fundamental task in content generation. Recent TTS systems, built upon either autoregressive (AR) or non-autoregressive (NAR) generative paradigms, have achieved impressive speech quality that approaches human-level naturalness [Wang et al., 2023, Le et al., 2024, Anastassiou et al., 2024, Ju et al., 2024, Du et al., 2025, Zhang et al., 2025]. Among these paradigms, NAR TTS—particularly diffusion-based models—stands out for its generation quality, architectural simplicity, and inference efficiency. Specifically, because NAR TTS can operate directly on continuous acoustic representations without relying on discrete audio tokenizers, it inherently bypasses complex system designs. Although early NAR systems heavily relied on auxiliary duration prediction modules to establish temporal alignment between text and audio [Ren et al., 2019, Le et al., 2024], recent advances have demonstrated that models can implicitly learn this alignment given sufficient training data [Eskimez et al., 2024a, Chen et al., 2024a, Lee et al., 2024], enabling further architectural simplification. Furthermore, by generating the entire speech sequence in parallel, NAR TTS exhibits a distinct speed advantage over its AR counterparts, especially as the sequence length increases. Despite these advantages, hybrid architectures that integrate both AR and NAR technologies have recently dominated the SOTA landscape [Betker, 2023, Anastassiou et al., 2024, Du et al., 2024a, Zhang et al., 2025], generally outperforming pure diffusion-based NAR models [Chen et al., 2024a, Lee et al., 2024]. An exception is the diffusion-based variant Seed-DiT, which reportedly surpasses its hybrid counterpart, Seed-ICL, within the Seed-TTS framework [Anastassiou et al., 2024]. However, the exact architecture and technical details

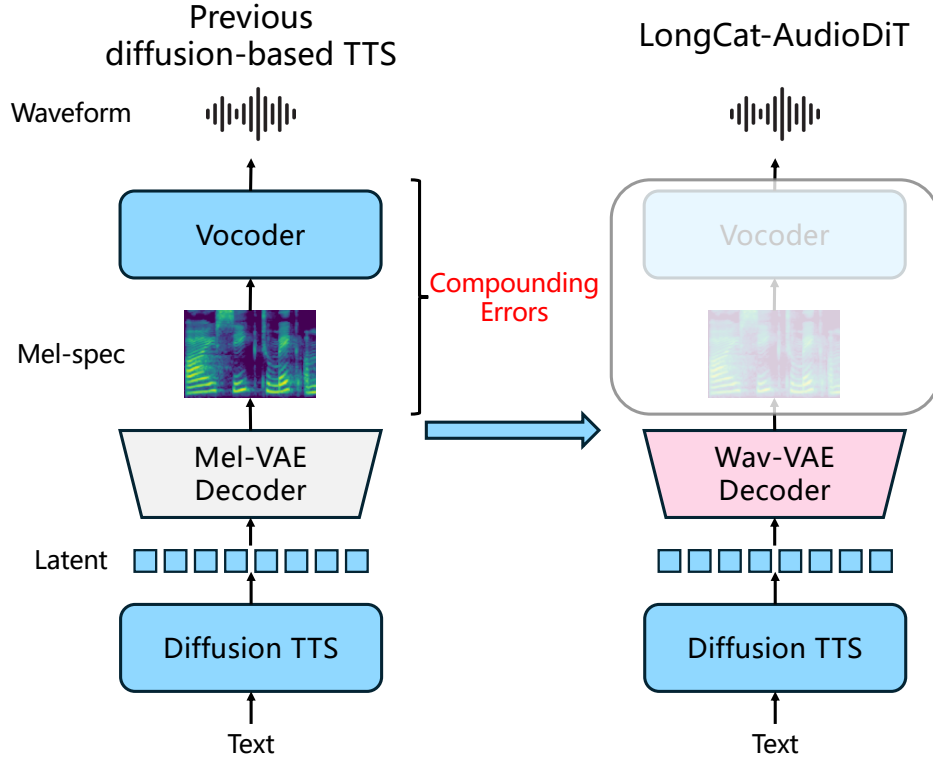


Figure 1 | Overview of LongCat-AudioDiT. Our architecture generates continuous waveform latents directly, thereby avoiding the compounding errors that inherently arise when predicting and subsequently converting intermediate representations (e.g., mel-spectrograms) into waveforms.

of Seed-DiT remain undisclosed, leaving a critical gap regarding how to construct a pure, highly performant diffusion-based TTS system.

In this paper, we present LongCat-AudioDiT, a diffusion-based NAR TTS model that achieves SOTA performance. A core finding of our work is that training the diffusion model directly in the waveform latent space yields substantial improvements over traditional paradigms that rely on intermediate acoustic representations, such as mel-spectrograms. Consequently, LongCat-AudioDiT consists of only two streamlined components: a waveform variational autoencoder (Wav-VAE) [Kingma and Welling, 2013] and a diffusion Transformer (DiT) [Vaswani et al., 2017, Peebles and Xie, 2023]. During training, the VAE encoder produces continuous latents for the DiT. During inference, the VAE decoder synthesizes raw waveforms directly from the latents sampled by the DiT, completely bypassing intermediate representations and eliminating the need for auxiliary vocoders heavily relied upon in previous studies [Chen et al., 2024a, Lee et al., 2024]. This end-to-end design mitigates the compounding errors typically incurred when predicting mel-spectrograms and subsequently converting them into waveforms. To support robust multilingual synthesis, we condition the model not only on the last hidden states but also on the raw word embeddings extracted from a pretrained language model. Furthermore, we introduce two critical improvements to the inference process: first, we identify and rectify a long-standing training-inference mismatch; second, we replace traditional classifier-free guidance with adaptive projection guidance to elevate generation quality. Finally, we explore the scalability of our architecture and observe a clear performance advantage when scaling up the model size. The final version of LongCat-AudioDiT, comprising 3.5B parameters and trained on 1 million hours of Chinese and English speech data, achieves SOTA performance on the Seed benchmark [Anastassiou et al., 2024]. To thoroughly validate our approach, we conduct comprehensive ablation studies on the proposed techniques. In addition, we systematically investigate the impact of latent dimensionality and compression rates on both the reconstruction fidelity of the Wav-VAE and the overall generation quality of the TTS model.

Our main contributions are summarized as follows:

- We propose LongCat-AudioDiT, a SOTA diffusion-based NAR TTS model. By operating directly in the waveform latent space, our approach effectively eliminates the compounding errors introduced by intermediate representations like mel-spectrograms.
- We propose two critical improvements to the inference process: first, we identify and rectify a long-standing training-inference mismatch; second, we replace traditional classifier-free guidance with adaptive projection guidance to elevate generation quality.
- We conduct systematic and comprehensive experiments to validate the effectiveness of our design choices. Notably, we provide empirical insights into the non-trivial relationship between the reconstruction quality of the Wav-VAE and the ultimate synthesis quality of the TTS backbone.
- We publicly release the source code and model weights of LongCat-AudioDiT to advance research and development within the community.

## 2 Related Work

### 2.1 Diffusion-based TTS

Early diffusion-based TTS models, such as Grad-TTS [Popov et al., 2021] and Diff-TTS [Jeong et al., 2021], adopted diffusion probabilistic models (DPMs) [Sohl-Dickstein et al., 2015, Song et al., 2020, Ho et al., 2020] governed by stochastic differential equations (SDEs). The fundamental concept of these approaches is to construct a bidirectional transformation between a simple Gaussian prior and the complex speech data distribution. While the forward process deterministically degrades speech data into Gaussian noise via continuous diffusion, the reverse denoising process lacks a closed-form solution and thus requires a neural network to approximate it.

More recently, flow matching paradigms [Lipman et al., 2022], built upon continuous normalizing flows (CNFs) [Chen, 2018], have become prevalent in diffusion-based TTS [Le et al., 2024, Mehta et al., 2024, Eskimez et al., 2024b, Chen et al., 2024a]. CNFs model the transformation as an ordinary differential equation (ODE) and can be efficiently trained using a simulation-free objective known as conditional flow matching (CFM) [Lipman et al., 2022]. Although recent studies have demonstrated that DPMs and CFM intrinsically belong to the same theoretical family [Albergo et al., 2025], CFM is often the preferred choice in practice. This is because it offers a simpler mathematical formulation [Liu et al., 2022a]—eliminating the need for complex noise scheduling—while delivering performance comparable or superior to traditional DPMs.

A parallel trajectory in the development of diffusion-based TTS focuses on text-to-speech alignment. While early systems addressed this challenge by incorporating explicit, auxiliary duration prediction modules [Popov et al., 2021, Shen et al., 2023, Le et al., 2024, Ju et al., 2024], recent advances have shifted towards fully end-to-end architectures. For instance, the representative E2-TTS [Eskimez et al., 2024a] framework, along with subsequent studies [Chen et al., 2024a, Lee et al., 2024, Zhu et al., 2025], demonstrated that the necessary alignment can be implicitly learned by the generative model without explicit supervision, provided there is sufficient training data.

LongCat-AudioDiT builds upon this modern trajectory by adopting both the CFM framework and an alignment-free architecture. However, we extend beyond these foundations by introducing several novel techniques designed to substantially improve the generation quality of diffusion-based TTS.

### 2.2 Latent Representations in Diffusion-based TTS

The choice of latent representation, which serves as the modeling target for the diffusion backbone, is critical in TTS systems. While it is feasible to train diffusion models directly on raw time-domain waveforms [Gao et al., 2023a], compressing the high-dimensional audio into a compact latent space has proven to be significantly more effective and computationally efficient [Rombach et al., 2022]. Specifically, the latent representation profoundly impacts both generation quality and synthesis speed, as it dictates the inherent trade-off between temporal compression rate and reconstruction fidelity. Most prior studies have adopted the mel-spectrogram as the default latent representation [Popov et al., 2021, Le et al., 2024, Eskimez et al., 2024b, Chen et al., 2024a], necessitating an auxiliary vocoder to invert the predicted mel-spectrograms back into audible waveforms. To achieve a higher compression rate and further accelerate inference, architectures like DiTTo-TTS [Lee et al., 2024] employ a Mel-VAE to encode the mel-spectrograms into an even lower-dimensional space. However, all these paradigms intrinsically suffer from potential compounding errors. These errors arise from the multiple

stages of data conversion—first predicting the intermediate acoustic features, and subsequently reconstructing the signal via a separate neural vocoder.

In LongCat-AudioDiT, we directly employ a waveform-based VAE (Wav-VAE) to encode raw audio into continuous latent representations. By unifying the acoustic modeling and waveform generation into a single continuous latent space, our approach elegantly bypasses intermediate transformations and mitigates the compounding error problem.

### 3 Wav-VAE

Compared to mel-spectrograms—which inherently discard phase information and fine-grained high-frequency details—compact variational autoencoder (VAE) representations retain essential acoustic characteristics while effectively eliminating redundant components. Consequently, they offer significantly greater potential for high-fidelity audio generation [Liu et al., 2022b, Lee and Kim, 2025, Qiang et al., 2024, Niu et al., 2025].

Motivated by these advantages, we develop a fully convolutional audio autoencoder that compresses raw waveforms into a compact, continuous latent representation. Operating directly in the time domain, the model consists of an encoder  $\mathcal{E}$ , a bottleneck module, and a decoder  $\mathcal{D}$ . Given an input waveform  $x \in \mathbb{R}^{1 \times T}$ , the encoder maps it to a latent sequence  $z \in \mathbb{R}^{D \times (T/R)}$ , where  $D$  denotes the latent dimensionality and  $R$  represents the temporal downsampling factor. Subsequently, the decoder reconstructs the waveform as  $\hat{x} = \mathcal{D}(z) \in \mathbb{R}^{1 \times T}$ .

#### 3.1 Model Architecture

**Encoder.** The encoder maps the input waveform to a low-dimensional latent sequence via hierarchical downsampling. The raw waveform is first projected into a high-dimensional feature space using a weight-normalized 1D convolution. The resulting representation is then processed by  $N$  cascaded Ooblock blocks [Evans et al., 2024]. The  $i$ -th block reduces the temporal resolution by a stride of  $s_i$  while expanding the channel dimension from  $C_i$  to  $C_{i+1}$ . The cumulative downsampling ratio is given by:  $R = \prod_{i=1}^N s_i$ .

Prior to downsampling, each block employs a stack of dilated residual units to capture multi-scale temporal dependencies. A residual unit updates the hidden representation  $h$  as follows:

$$h \leftarrow h + \text{Conv}_{1 \times 1}(\sigma(\text{Conv}_{k,d}(\sigma(h)))), \quad (1)$$

where  $\text{Conv}_{k,d}$  denotes a weight-normalized 1D convolution with kernel size  $k$  and dilation rate  $d$ , and  $\sigma$  represents the Snake activation function [Ziyin et al., 2020].

Following [Wu et al., 2025], to stabilize the training process under aggressive downsampling, each encoder block incorporates a non-parametric shortcut path. Specifically, let the input to the  $i$ -th block be a tensor of shape  $[B, C_i, T]$  with a target stride of  $s_i$ . A space-to-channel reshape operation first folds the temporal dimension into the channel axis, transforming the tensor to  $[B, C_i \cdot s_i, T/s_i]$ , thereby matching the desired downsampled temporal resolution. Next, a channel-wise averaging operation groups adjacent channels to reduce the dimension to  $C_{i+1}$ , yielding a tensor of shape  $[B, C_{i+1}, T/s_i]$ . This parameter-free branch establishes a linear residual pathway that bypasses the nonlinear transformations of the main block, and its output is combined with the block’s main output via element-wise addition.

Finally, a convolutional projection layer—also equipped with an analogous shortcut mechanism—is applied to map the deepest features to the target latent dimension  $D$ . A VAE bottleneck is then applied to the encoder’s output, generating the mean  $\mu$  and log-variance  $\log \sigma^2$ . The continuous latent representation is sampled using the reparameterization trick:  $z = \mu + \sigma \odot \epsilon$ , where  $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$ .

**Decoder.** The decoder architecture closely mirrors that of the encoder in reverse. The sampled latent sequence  $z$  is initially projected into a high-dimensional feature space via a weight-normalized 1D convolution, and then progressively upsampled through  $N$  cascaded decoder blocks. Following each upsampling step, the same stack of dilated residual units used in the encoder is applied to model multi-scale temporal dependencies.

Furthermore, each decoder block incorporates a non-parametric shortcut branch symmetric to its encoder counterpart. For an input tensor of shape  $[B, C_{i+1}, T/s_i]$ , a channel-to-space rearrangement first restores the temporal resolution to  $T$ . This is followed by a channel replication step to match the main branch’s output shape of  $[B, C_i, T]$ . The shortcut and main branch outputs are then fused via element-wise addition. A final convolutional projection layer maps the reconstructed features back to the time-domain waveform  $\hat{x}$ .

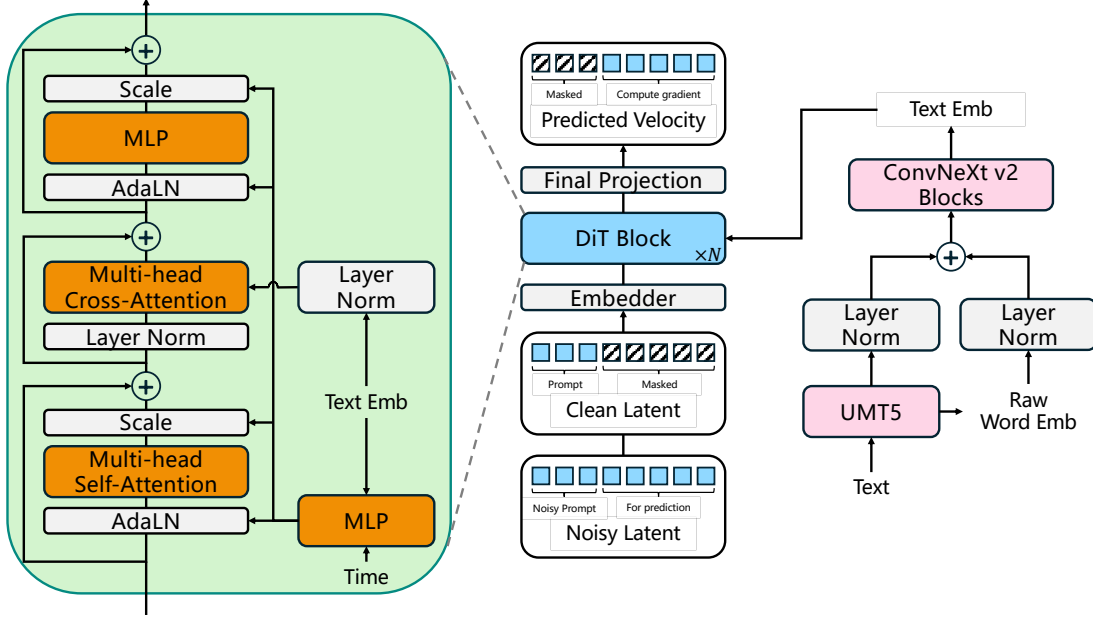


Figure 2 | Architecture of LongCat-AudioDiT. *Middle*: The overall architecture. *Left*: Detailed structure of the DiT block. *Right*: Detailed structure of the text encoder.

### 3.2 Training Objective

The Wav-VAE is optimized via a two-stage adversarial training procedure. The generator (i.e., the autoencoder) minimizes a combined loss function formulated as:

$$\mathcal{L}_{\text{gen}} = \lambda_{\text{spec}} \mathcal{L}_{\text{spec}} + \lambda_{\text{mel}} \mathcal{L}_{\text{mel}} + \lambda_{\text{time}} \mathcal{L}_{\text{time}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{fm}} \mathcal{L}_{\text{fm}}. \quad (2)$$

The individual components of this objective are defined as follows:

- $\mathcal{L}_{\text{spec}}$  (Multi-resolution STFT loss [Zeghidour et al., 2021]): Incorporates perceptual weighting to encourage faithful reproduction of the time-frequency structure across various scales.
- $\mathcal{L}_{\text{mel}}$  (Multi-scale mel-spectrogram loss [Kumar et al., 2023]): Reduces spectral discrepancies across multiple FFT resolutions, ensuring perceptually natural synthesis.
- $\mathcal{L}_{\text{time}}$  (L1 time-domain loss): Directly minimizes the sample-level absolute error between the input and the reconstructed waveforms.
- $\mathcal{L}_{\text{KL}}$  (KL divergence loss): Regularizes the learned latent distribution towards a standard Gaussian prior, ensuring a smooth, continuous, and well-structured latent space suitable for the diffusion model.

The remaining two terms are derived from a multi-scale STFT discriminator, which is trained in parallel using a standard adversarial objective. Specifically, the adversarial loss  $\mathcal{L}_{\text{adv}}$  encourages the generator to synthesize waveforms that are perceptually indistinguishable from real audio. Meanwhile, the feature matching loss [Kong et al., 2020]  $\mathcal{L}_{\text{fm}}$  minimizes the L1 distance between the intermediate feature maps extracted by the discriminator for both real and reconstructed audio.

To ensure training stability, we employ an initial warmup phase. During this period, the adversarial and feature matching terms ( $\mathcal{L}_{\text{adv}}$  and  $\mathcal{L}_{\text{fm}}$ ) are disabled. This strategy allows the autoencoder to establish a stable and accurate reconstruction mapping before being subjected to the more challenging adversarial gradients.

## 4 Diffusion TTS

### 4.1 Overview

We adopt the Conditional Flow Matching (CFM) framework [Lipman et al., 2022] to model the TTS process as an Ordinary Differential Equation (ODE):  $dz_t = v_t dt$ , which deterministically transports random Gaussian noise  $z_0$  to target speech latents  $z_1$  along a velocity field  $v_t$ . Following the rectified flow formulation [Liu et al.,



2022a], we construct the noisy latent  $z_t$  via linear interpolation between the clean latent and the noise prior:

$$z_t = (1 - t)z_0 + tz_1. \quad (3)$$

The velocity field is estimated by a neural network parameterized by  $\theta_{\text{CFM}}$ , conditioned on the text sequence  $q$  and an audio context prompt  $z_{\text{ctx}}$ . Following VoiceBox [Le et al., 2024], we construct  $z_{\text{ctx}}$  by randomly masking continuous spans of the clean latent  $z_1$ , a strategy that inherently enables zero-shot voice cloning capabilities. The optimization objective for CFM is to minimize the mean squared error between the predicted velocity  $v_\theta$  and the ground-truth target velocity ( $z_1 - z_0$ ) over the masked regions:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,m,z_0,z_1} \left[ \left\| (1 - m) \odot ((z_1 - z_0) - v(z_t, t, z_{\text{ctx}}, q; \theta_{\text{CFM}})) \right\|^2 \right], \quad (4)$$

where  $m$  denotes the random binary mask used to generate  $z_{\text{ctx}}$ . Furthermore, to facilitate classifier-free guidance (CFG) [Ho and Salimans, 2021] during inference, we jointly drop the audio context  $z_{\text{ctx}}$  and the text condition  $q$  with a probability of 10% during training, thereby enabling the model to learn an unconditional distribution.

The overall architecture of our CFM network, illustrated in Fig. 2, is built upon the Diffusion Transformer (DiT) paradigm [Peebles and Xie, 2023]. It leverages a standard Transformer [Vaswani et al., 2017] backbone and employs Adaptive Layer Normalization (AdaLN) [Perez et al., 2018] to inject the timestep condition  $t$ . To stabilize the training dynamics, we incorporate QK-Norm [Henry et al., 2020] within the attention modules. While standard LayerNorm [Ba et al., 2016] is utilized throughout the network, RMSNorm [Zhang and Sennrich, 2019] is specifically applied for the QK-Norm operations. Following DiTTo-TTS [Lee et al., 2024], we utilize cross-attention mechanisms to implicitly learn the text-to-speech alignment, and apply Rotary Positional Embedding (RoPE) [Su et al., 2024] across all attention layers to capture relative positional dependencies.

We also integrate two structural optimizations from DiTTo-TTS: long-skip connections and a global AdaLN formulation. The long-skip connection directly adds the network’s input to the final-layer hidden state, a modification that yielded slight but consistent improvements in our preliminary experiments. The global AdaLN mechanism, originally proposed in Gentron [Chen et al., 2024b], replaces individual AdaLN projections with a shared, global block for all DiT layers. We observe that this design significantly reduces the overall parameter count without degrading generation performance.

Additionally, we adopt Representation Alignment (REPA) [Yu et al., 2024] to ground the internal representations of the DiT to a robust, self-supervised semantic space. Specifically, we employ a pretrained mHuBERT model [Boito et al., 2024] and minimize the L1 distance between the outputs of the 8-th DiT layer and the corresponding mHuBERT features for the identical input speech. Our preliminary findings indicate that while REPA does not enhance the generation quality, it substantially accelerates the convergence during training.

In the next section, we detail our text encoder that supports multiple languages.

## 4.2 Multilingual Text Embedding

Our goal is to design a robust text encoder capable of supporting multilingual synthesis. Existing approaches typically either train a text encoder from scratch [Chen et al., 2024a] or leverage a pretrained language model, such as ByT5 [Xue et al., 2022, Lee et al., 2024]. However, training from scratch is highly resource-intensive and notoriously difficult to scale to new languages. Conversely, while ByT5 theoretically supports arbitrary languages, its byte-level tokenization results in prohibitively long sequence lengths for languages like Chinese, which empirically led to suboptimal performance and alignment difficulties in our preliminary experiments. To overcome these limitations, we propose utilizing UMT5 [Chung et al., 2023], a multilingual variant of T5, as our foundational text encoder. UMT5 supports 107 languages and employs a subword tokenizer that maintains reasonable sequence lengths across diverse languages, perfectly aligning with our architectural requirements. A standard practice when utilizing pretrained language models is to extract the last hidden state as the text representation  $q$ . However, we observed that relying exclusively on the final layer yields poor intelligibility in the TTS task. We hypothesize that while the last hidden state is rich in high-level semantic information, it abstracts away the low-level lexical and phonetic cues that are crucial for precise acoustic mapping. Motivated by this, we propose integrating the raw word embeddings (the initial embedding layer of UMT5) with the final hidden state. The resulting text representation  $q$  for LongCat-AudioDiT is formulated as:

$$q = \text{LayerNorm}(\text{last\_hidden\_state}) + \text{LayerNorm}(\text{raw\_word\_embedding}). \quad (5)$$

Here, non-parametric LayerNorm is applied to appropriately balance the distinct scales of the two representational spaces before summation. Although our empirical validation is conducted using UMT5, we posit that

this dual-embedding extraction strategy is model-agnostic and can be generalized to other large multilingual language models. We use UMT5-base<sup>1</sup> in all experiments.

Furthermore, following F5-TTS [Chen et al., 2024a], we pass the extracted text representation  $q$  through a lightweight sequence refinement module based on ConvNeXt V2 [Woo et al., 2023]. We empirically find that this localized convolutional refinement significantly accelerates the convergence of the text-to-speech alignment during training.

In the subsequent sections, we introduce two improvements to the inference process proposed in LongCat-AudioDiT that further elevate generation performance.

### 4.3 Mitigating the Training-Inference Mismatch in Noisy Latent

During inference, we employ the Euler method to solve the ODE. The number of function evaluations is set to 16. Initializing the process with randomly sampled Gaussian noise  $z_0$ , we iteratively update the latent  $z_t$  at each step as follows:

$$z_{t+\Delta t} = z_t + v(z_t, t, z_{ctx}, q; \theta_{CFM})\Delta t, \quad (6)$$

where  $\Delta t$  is the predefined integration step size.

By revisiting this sequential inference process, we identify a critical training-inference mismatch regarding the state of the noisy latent  $z_t$ . For clarity, we conceptually partition  $z_t$  along the temporal axis into two segments:  $z_t^{ctx} = z_t[:T_{ctx}]$  corresponding to the conditioning prompt, and  $z_t^{gen} = z_t[T_{ctx}:]$  corresponding to the target generation region, where  $T_{ctx}$  denotes the duration of the prompt latent  $z_{ctx}$ .

Recall that during training, the exact trajectory of the entire  $z_t$  is constructed via linear interpolation (Eq. 3), acting as the ground truth (GT) noisy latent. During inference, however, an asymmetry emerges. Because the flow matching objective (Eq. 4) penalizes velocity prediction errors only on the masked target region ( $v^{gen}$ ), the iterative update successfully yields a valid approximation of the GT trajectory for  $z_t^{gen}$ . Conversely, because no loss is computed over the prompt region, the model’s velocity predictions for  $z_t^{ctx}$  are essentially unconstrained and arbitrary. Consequently, accumulating these unconstrained updates causes  $z_t^{ctx}$  to drift away from its theoretical GT trajectory, thus introducing a training-inference mismatch that has been overlooked in prior work [Le et al., 2024, Chen et al., 2024a]. We resolve this discrepancy by forcibly overwriting  $z_t^{ctx}$  with its GT value at every inference step:

$$z_t^{ctx} \leftarrow tz^{ctx} + (1-t)z_0^{ctx}, \quad (7)$$

where  $z_0^{ctx}$  is the initial Gaussian noise of the prompt part.

Furthermore, on the basis of this problem, we propose a corollary for CFG. To obtain a truly unconditional velocity estimate, it is insufficient to merely drop  $z_{ctx}$ ; the explicitly constructed noisy prompt latent  $z_t^{ctx}$  must also be dropped, as it inherently leaks acoustic information about the prompt.

In Section 5.3.3, we empirically demonstrate that mitigating this mismatch and isolating the conditional information yields substantial improvements in overall synthesis performance.

### 4.4 Replacing CFG with Adaptive Projection Guidance

Following standard practice, we first utilize classifier-free guidance (CFG) [Ho and Salimans, 2021] to steer the predicted velocity at each integration step:

$$v_t^{CFG} = v_t + \alpha(v_t - v_t^u), \quad (8)$$

where  $v_t^u = v(z_t^u, t, \emptyset, \emptyset; \theta_{CFM})$  represents the unconditional velocity;  $\alpha$  denotes the CFG scale. By default, we set  $\alpha = 4.0$ . As established in Section 4.3, to accurately compute the unconditional velocity, we compute the noisy latent  $z_t^u$  by dropping the prompt part  $z_t^{ctx}$  to avoid information leakage, i.e.,  $z_t^u = \text{concat}(\emptyset, z_t^{gen})$ .

In our preliminary experiments, while standard CFG effectively improved synthesis quality, it occasionally introduced audible artifacts, and increasing the guidance scale  $\alpha$  further exacerbated the degradation. We hypothesize that a large CFG scale induces an *oversaturation* phenomenon, a widely recognized issue in diffusion-based image generation [Kynkäänniemi et al., 2024]. To alleviate this problem, we incorporate Adaptive Projection Guidance (APG) [Sadat et al., 2024]. The core intuition of APG is to decompose the guidance residual,  $v_t - v_t^u$ , into two geometrically orthogonal components: one parallel to the conditional

<sup>1</sup><https://huggingface.co/google/umt5-base>

prediction  $v_t$  and the other orthogonal to it. APG theorizes that the parallel component is the primary cause behind oversaturation; thus, the issue can be resolved by selectively dampening this term.

To integrate APG into our flow matching framework, we first project the model’s output from the velocity domain into the data sample domain (i.e., predicting  $z_1$ ), as suggested by [Sadat et al. \[2024\]](#):  $\mu_t = z_t + (1-t)v_t$ . Let the guidance term in this sample domain be denoted as  $\Delta\mu_t = \mu_t - \mu_t^u$ . The parallel component  $\Delta\mu_t^\parallel$  with respect to  $\mu_t$  is calculated as:  $\Delta\mu_t^\parallel = \frac{\langle \Delta\mu_t, \mu_t \rangle}{\langle \mu_t, \mu_t \rangle} \mu_t$ , and the corresponding orthogonal term is  $\Delta\mu_t^\perp = \Delta\mu_t - \Delta\mu_t^\parallel$ . The APG-adjusted prediction in the sample domain is then formulated as:

$$\mu_t^{\text{APG}} = \mu_t + \alpha \Delta\mu_t^\perp + \eta \Delta\mu_t^\parallel, \quad (9)$$

where  $\eta$  acts as a dampening factor for the parallel component and is set to 0.5 by default. Subsequently, we map the adjusted sample prediction back to the velocity domain to proceed with the ODE solver:

$$v_t^{\text{APG}} = \frac{\mu_t^{\text{APG}} - z_t}{1-t}. \quad (10)$$

Furthermore, we adopt the reverse momentum trick proposed in APG [\[Sadat et al., 2024\]](#), which maintains a moving average  $\Delta\mu_t \leftarrow \Delta\mu_t + \beta \Delta\mu_t$ . Applying a negative momentum ( $\beta < 0$ ) forces the guidance to focus more on the current update direction rather than accumulating past momentum. By default, we set  $\beta = -0.3$ .

As demonstrated in Section 5.3.3, APG effectively eliminates artifacts and significantly elevates synthesis quality.

## 5 Experiments

### 5.1 Experimental Setup

**Data** For the training of the Wav-VAE, we employ a curated internal corpus comprising 200K hours of Chinese and English speech. Audio clips are segmented to approximately 3 seconds.

For the TTS backbone (DiT), we utilize a curated internal dataset containing 100K hours of Chinese and English speech for all baseline and ablation experiments. For the large-scale scaling experiments, this training corpus is further expanded to 1M hours. The transcriptions for all utterances are obtained by a speech recognition model. We sample all audio data at 24 kHz. The maximal audio duration-TTS training is 60 seconds.

**Training Details** The Wav-VAE contains 157M parameters and is optimized on 32 NVIDIA H800 GPUs with a global batch size of 384. By default, the model is configured with a latent dimensionality of 64 and operates at a temporal frame rate of 11.72 Hz.

For the diffusion backbone, we train two variants with 1B and 3.5B parameters, respectively. The 1B model is trained on 16 GPUs with a global batch size of 256, whereas the 3.5B model utilizes 64 GPUs with a global batch size of 1024. Both models are optimized using AdamW [\[Loshchilov and Hutter, 2018\]](#), with moving average coefficients set to  $\beta_1 = 0.9$  and  $\beta_2 = 0.95$ . We apply a linear learning rate decay schedule, gradually decreasing the learning rate from  $1e-4$  to  $1e-5$  following an initial 1K warmup steps.

**Evaluation Metrics** We benchmark the Wav-VAE on the LibriTTS `test-clean` subset [\[Zen et al., 2019\]](#), and evaluate the full TTS pipeline on the Seed benchmark [\[Anastassiou et al., 2024\]](#).

To evaluate the Wav-VAE reconstruction fidelity, we adopt standard objective metrics including PESQ [\[Rix et al., 2001\]](#) for assessing perceptual quality and STOI [\[Taal et al., 2011\]](#) for measuring speech intelligibility.

The generative capabilities of the TTS models are evaluated across four primary dimensions: intelligibility, zero-shot voice cloning, naturalness, and overall acoustic quality. We measure these using the following metrics:

- **Character/Word Error Rate (CER/WER):** To quantify intelligibility, we transcribe the synthesized speech using Whisper large-v3 [\[Radford et al., 2023\]](#) for English and Paraformer [\[Gao et al., 2023b\]](#) for Chinese, subsequently calculating the respective CER or WER.
- **Speaker Similarity (SIM):** To evaluate voice cloning accuracy, we compute the cosine similarity between the speaker embeddings of the reference prompt and the synthesized speech. This formulation is mathematically equivalent to the SIM-O metric proposed in VoiceBox [\[Le et al., 2024\]](#). Following Seed-TTS [\[Anastassiou](#)



Table 1 | Objective evaluation results of LongCat-AudioDiT on the Seed benchmark [Anastassiou et al., 2024]. The results of other methods are taken from the original paper or, if open-sourced, evaluated by us. **Bold** indicates the best score. Underline indicates the second-best score.

Model	ZH		EN		ZH-Hard	
	CER (%) ↓	SIM ↑	WER (%) ↓	SIM ↑	CER (%) ↓	SIM ↑
GT	1.26	0.755	2.14	0.734	-	-
<b>NAR Models</b>						
Seed-DiT [Anastassiou et al., 2024]	1.18	0.809	1.73	<b>0.790</b>	-	-
MaskGCT [Wang et al., 2024]	2.27	0.774	2.62	0.714	10.27	0.748
E2 TTS [Eskimez et al., 2024b]	1.97	0.730	2.19	0.710	-	-
F5 TTS [Chen et al., 2024a]	1.56	0.741	1.83	0.647	8.67	0.713
F5R-TTS [Sun et al., 2025]	1.37	0.754	-	-	8.79	0.718
ZipVoice [Zhu et al., 2025]	1.40	0.751	1.64	0.668	-	-
<b>AR/Hybrid Models</b>						
Seed-ICL [Anastassiou et al., 2024]	1.12	0.796	2.25	0.762	7.59	0.776
SparkTTS [Wang et al., 2025]	1.20	0.672	1.98	0.584	-	-
Qwen2.5-Omni [Xu et al., 2025]	1.70	0.752	2.72	0.632	7.97	0.747
CosyVoice [Du et al., 2024a]	3.63	0.723	4.29	0.609	11.75	0.709
CosyVoice2 [Du et al., 2024b]	1.45	0.748	2.57	0.652	6.83	0.724
FireRedTTS-1S [Guo et al., 2025]	1.05	0.750	2.17	0.660	7.63	0.748
CosyVoice3-1.5B [Du et al., 2025]	1.12	0.781	2.21	0.720	<u>5.83</u>	0.758
IndexTTS2 [Zhou et al., 2025a]	1.03	0.765	2.23	0.706	7.12	0.755
DiTAR [Jia et al., 2025]	1.02	0.753	1.69	0.735	-	-
MiniMax-Speech [Zhang et al., 2025]	0.99	0.799	1.90	0.738	-	-
VoxCPM [Zhou et al., 2025b]	<u>0.93</u>	0.772	1.85	0.729	8.87	0.730
MOSS-TTS [SII-OpenMOSS, 2026]	1.20	0.788	1.85	0.734	-	-
Qwen3-TTS [Hu et al., 2026]	1.22	0.770	<b>1.23</b>	0.717	6.76	0.748
CosyVoice3.5	<b>0.87</b>	0.797	1.57	0.738	<b>5.71</b>	0.786
LongCat-AudioDiT-1B	1.18	<u>0.812</u>	1.78	0.762	6.33	<u>0.787</u>
LongCat-AudioDiT-3.5B	1.09	<b>0.818</b>	<u>1.50</u>	<u>0.786</u>	6.04	<b>0.797</b>

et al., 2024], we utilize a fine-tuned WavLM [Chen et al., 2022] (wavlm\_large\_finetune<sup>2</sup>) to extract the robust speaker embeddings.

- **UTMOS** [Saeki et al., 2022]: A highly correlated neural objective metric used to approximate human Mean Opinion Scores (MOS) regarding speech naturalness.
- **DNSMOS** [Reddy et al., 2021]: A widely adopted objective metric designed to evaluate the overall perceptual acoustic quality of the synthesized audio.

Note that a subset of these TTS metrics is also applied to evaluate the Wav-VAE reconstructions, allowing us to comparatively analyze the inherent gap between representation reconstruction (Wav-VAE) and generation (TTS).

Finally, we benchmark LongCat-AudioDiT against strong prior work, encompassing purely NAR diffusion models, AR models, and state-of-the-art hybrid TTS architectures.

## 5.2 Main Results

The evaluation results for both the full LongCat-AudioDiT pipeline and the standalone Wav-VAE are presented in Table 1 and Table 2, respectively.

**TTS Synthesis Performance** As demonstrated in Table 1, our proposed TTS model consistently outperforms the majority of prior art, achieving particularly remarkable gains in speaker similarity (SIM) over the highly competitive Seed-DiT architecture [Anastassiou et al., 2024]. Specifically, LongCat-AudioDiT establishes new state-of-the-art (SOTA) SIM scores on the demanding Seed-ZH and Seed-Hard benchmarks, while securing the second-best SIM score on Seed-EN. Most notably, our end-to-end framework decisively surpasses all previous diffusion-based paradigms—such as F5-TTS [Chen et al., 2024a]—that rely on intermediate mel-spectrograms

<sup>2</sup>[https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker\\_verification](https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification)

Table 2 | Objective evaluation results of the proposed Wav-VAE on the LibriTTS Zen et al. [2019] test-clean subset. **Bold** indicates the best score among continuous VAEs.  $N_q$  is the number of codebooks for discrete codecs. For codecs, frame per second (FPS) denotes the number of tokens per second.

Model	$N_q$	FPS	PESQ $\uparrow$	STOI $\uparrow$	UTMOS $\uparrow$
GT	–	–	4.644	1.0	4.056
<b>Discrete Codecs</b>					
DAC [Kumar et al., 2023]	9	900	3.908	0.970	3.910
Encodec [Défossez et al., 2022]	8	600	2.720	0.939	3.040
Vocos [Siuzdak, 2023]	8	600	2.807	0.943	3.695
WavTokenizer [Ji et al., 2024]	1	75	2.373	0.914	4.049
BigCodec [Xin et al., 2024]	1	80	2.697	0.939	4.097
<b>Continuous VAEs</b>					
VibeVoice [Peng et al., 2025]	1	7.50	3.068	0.828	<b>4.181</b>
Ours Wav-VAE	1	7.81	3.089	0.963	4.116
Ours Wav-VAE	1	11.72	<b>3.237</b>	<b>0.967</b>	4.013

as generation targets. This substantial margin strongly validates our core hypothesis: operating directly within the waveform latent space effectively circumvents compounding errors and yields superior voice cloning fidelity.

Regarding intelligibility (WER/CER), LongCat-AudioDiT achieves highly competitive performance relative to existing open-source baselines. While our error rates slightly trail heavily engineered proprietary systems like Qwen3-TTS [Hu et al., 2026] and CosyVoice3.5, it is crucial to emphasize that those models rely on complex multi-stage training pipelines and massive amounts of high-quality, human-annotated data. In contrast, LongCat-AudioDiT attains its performance with a remarkably simplified end-to-end architecture and a single training stage.

**Wav-VAE Reconstruction Quality** The intrinsic reconstruction capabilities of our Wav-VAE are detailed in Table 2. Operating at a comparable frame rate (FPS), our Wav-VAE exhibits superior overall reconstruction fidelity compared to the baseline Wav-VAE introduced in VibeVoice [Peng et al., 2025]. Furthermore, when juxtaposed with SOTA discrete audio codecs, our continuous Wav-VAE not only outperforms most of them in acoustic quality but does so while operating at a drastically reduced sequence length (fewer frames per second). This stark contrast strongly underscores the inherent capacity advantages and expressive efficiency of modeling continuous latent representations over discrete tokens.

### 5.3 Ablation Studies

To systematically validate our architectural choices and the proposed techniques, we conduct comprehensive ablation experiments. Specifically, our investigations are guided by the following three core research questions (RQs):

- **RQ1:** As a modeling target-TTS, does the waveform latent (Wav-VAE) outperform intermediate representations like the mel-spectrogram latent (Mel-VAE)?
- **RQ2:** What is the intrinsic relationship between VAE reconstruction fidelity and the downstream TTS synthesis quality? Does a superior VAE guarantee a better generative TTS model?
- **RQ3:** How effectively do our inference techniques, i.e., solving training-inference mismatch and APG, contribute to the overall generation quality?

#### 5.3.1 RQ1: Wav-VAE vs. Mel-VAE-TTS Generation

The central hypothesis underpinning LongCat-AudioDiT is that modeling directly within the waveform latent space is superior to utilizing intermediate representations, primarily due to the mitigation of compounding errors. Since recent work like DiTTo-TTS [Lee et al., 2024] has already established that Mel-VAE outperforms raw mel-spectrograms in diffusion-based TTS, we restrict our comparison directly to Wav-VAE versus Mel-VAE.

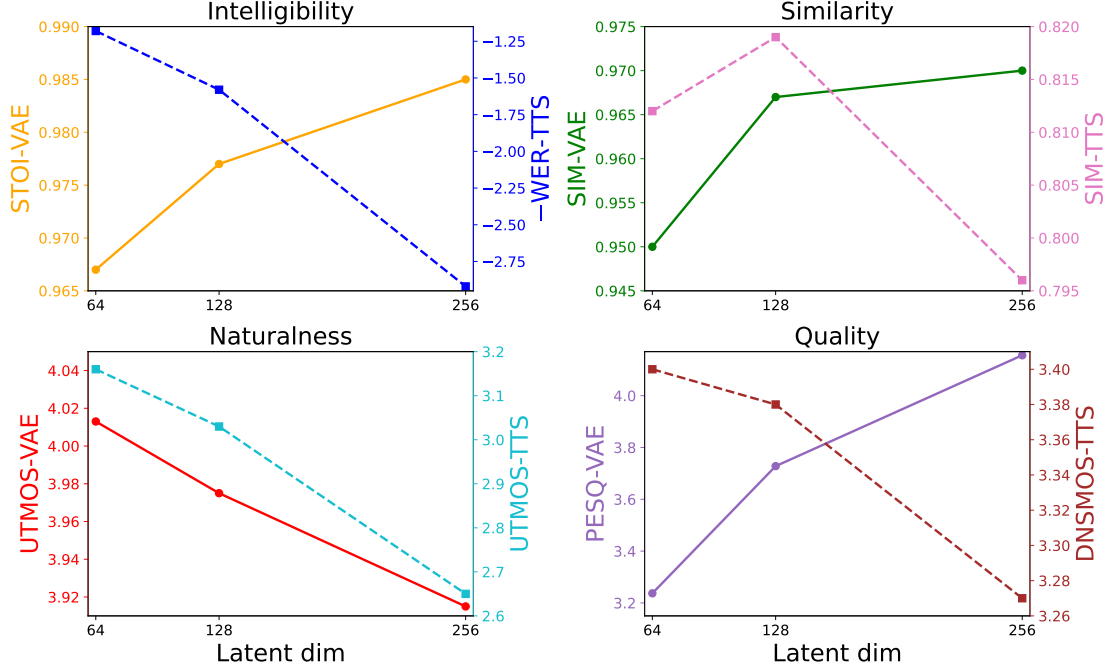


Figure 3 | Objective evaluation results for both Wav-VAE reconstruction and TTS synthesis under varying *latent dimensions*. For ease of reading, we negate WER-TTS.

Table 3 | Objective evaluation results of TTS models based on Wav-VAE and Mel-VAE on the Seed benchmark [Anastassiou et al., 2024]. **Bold** indicates the best score.

TTS Latent Model	ZH		EN		ZH-Hard	
	CER (%) ↓	SIM ↑	WER (%) ↓	SIM ↑	CER (%) ↓	SIM ↑
Mel-VAE	1.29	0.706	2.20	0.714	7.70	0.696
Wav-VAE	<b>1.18</b>	<b>0.812</b>	<b>1.78</b>	<b>0.762</b>	<b>6.33</b>	<b>0.787</b>

For this experiment, we adopt the open-source Mel-VAE introduced in ACE-Step [Gong et al., 2025]. Although originally designed for music generation, we empirically verify that this Mel-VAE yields high-fidelity speech reconstruction at a similar frame rate to our proposed Wav-VAE. We train a baseline 1B parameter TTS model using this Mel-VAE as the modeling target. During inference, the generated latents are decoded into mel-spectrograms, which are subsequently inverted into time-domain waveforms using the officially provided high-quality vocoder<sup>3</sup>.

The comparative evaluation results are presented in Table 3. As observed, the LongCat-AudioDiT model built upon the Wav-VAE consistently and significantly outperforms the Mel-VAE-based baseline across all metrics, validating our core assumption. Remarkably, while improvements in intelligibility (WER/CER) are solid, the Wav-VAE yields a drastic boost in the speaker similarity (SIM) metric. This targeted improvement elegantly corroborates our hypothesis: fine-grained, high-frequency acoustic details—which are essential for zero-shot voice cloning—are intrinsically fragile and easily lost during the cascading conversions (latent → mel-spectrogram → waveform) inherent to the Mel-VAE pipeline.

### 5.3.2 RQ2: The Interplay Between Wav-VAE Reconstruction and TTS Generation

We investigate the intrinsic relationship between the reconstruction fidelity of the Wav-VAE and the generation quality of the downstream TTS model. A naive assumption is that a superior Wav-VAE guarantees better TTS performance, given that the VAE’s reconstruction fidelity inherently defines the upper bound for the generative model. To test this hypothesis, we train multiple Wav-VAEs with varying latent dimensionalities and temporal frame rates (FPS), subsequently training a corresponding TTS backbone for each VAE variant. Specifically,

<sup>3</sup><https://github.com/ace-step/ACE-Step>

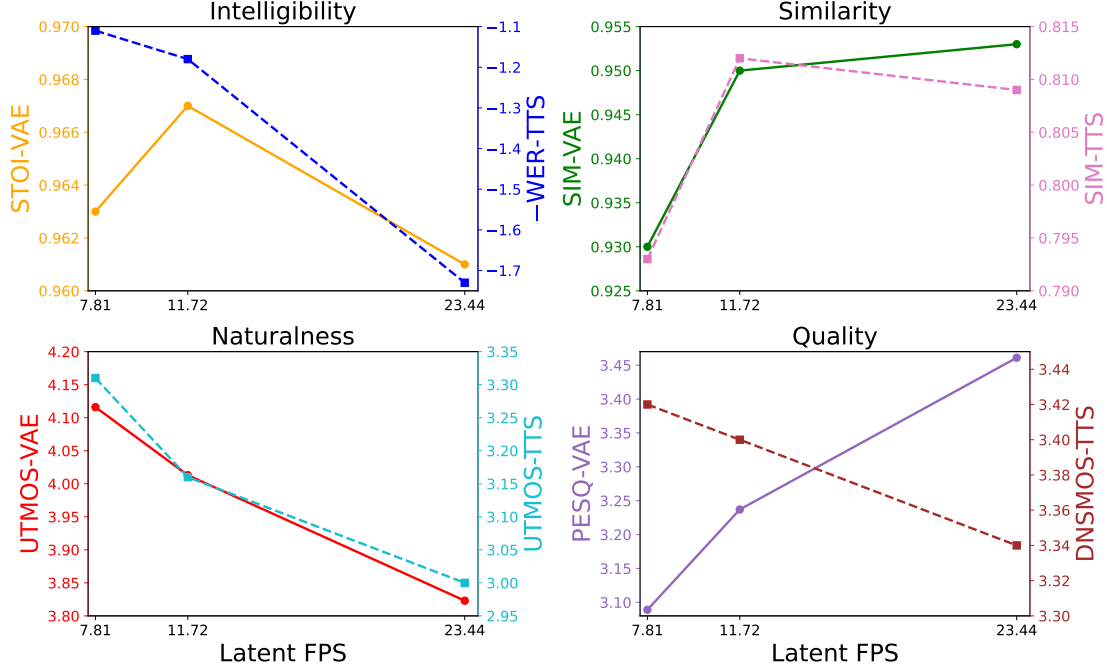


Figure 4 | Objective evaluation results for both Wav-VAE reconstruction and TTS synthesis across varying latent frame rates (FPS). For ease of reading, we negate WER-TTS.

we select latent dimensions from the set  $\{64, 128, 256\}$  and frame rates from  $\{7.81, 11.72, 23.44\}$ , yielding a total of 6 unique Wav-VAE models and 6 paired TTS models. For the dimension ablation (3 models), we fix the frame rate at 20 Hz; conversely, for the frame rate ablation (3 models), we fix the latent dimension at 64. All TTS models in this ablation are trained using the exact configurations as the LongCat-AudioDiT-1B baseline.

The comprehensive evaluation results are visualized in Fig. 3 and Fig. 4. To facilitate a clear comparison across domains, we categorize the metrics into four analogous groups: intelligibility (STOI-VAE & WER-TTS), speaker similarity (SIM-VAE & SIM-TTS), naturalness (UTMOS-VAE & UTMOS-TTS), and overall acoustic quality (PESQ-VAE & DNSMOS-TTS). Note that the VAE similarity (SIM-VAE) is calculated by comparing the ground truth (GT) utterance against its direct reconstruction.

**Observation 1: The Dimension-Capacity Trade-off.** Under a fixed TTS parameter budget, increasing the latent dimension consistently improves the Wav-VAE’s reconstruction fidelity but simultaneously degrades the TTS generation quality (see Fig. 3). This finding directly contradicts the naive assumption. We initially hypothesized that increasing the TTS model capacity might resolve this mismatch; thus, we scaled up the TTS backbone to 3.5B parameters, conditioned on the 128-dimensional Wav-VAE. However, while this larger variant achieved a marginal gain in SIM score, its overall performance remained inferior to the 3.5B model conditioned on the 64-dimensional Wav-VAE (as reported in Table 1). This suggests that excessively high-dimensional continuous latents impose a severe modeling burden on the diffusion backbone that cannot be easily overcome merely by scaling up parameters.

**Observation 2: The Frame Rate Sweet Spot.** There exists an optimal temporal frame rate (FPS) that balances VAE and TTS performance, though this sweet spot is not necessarily identical for both tasks (see Fig. 4). For the Wav-VAE, a lower FPS surprisingly yields better intelligibility and naturalness, but penalizes similarity and overall acoustic quality. This behavior is intuitive: an aggressively downsampled (lower FPS) latent forces the autoencoder to discard fine-grained, high-frequency acoustic details (hurting SIM and PESQ) while preserving global phonetic structures (aiding STOI). Conversely, for the generative TTS model, a lower FPS substantially boosts the overall synthesis quality. We observe that the diffusion backbone struggles to accurately model the complex, highly correlated temporal dynamics of high-FPS latents, leading to unstable generation.

Synthesizing these two critical observations, we empirically identify the 64-dimensional, 11.72-Hz Wav-VAE as the optimal representation target, and adopt it as the default configuration for all LongCat-AudioDiT models.

Table 4 | Objective evaluation results of the ablation studies on noise-prompt dual masking and APG on the Seed-ZH benchmark [Anastassiou et al., 2024]. **Bold** indicates the best score.

Experiment	CER (%) ↓	SIM ↑	UTMOS ↑	DNSMOS ↑
LongCat-AudioDiT-1B	<b>1.18</b>	<b>0.812</b>	<b>3.16</b>	<b>3.40</b>
training-inference mismatch	1.21	0.769	2.83	3.34
w/o APG	<b>1.18</b>	<b>0.812</b>	3.06	3.38

### 5.3.3 RQ3: Effectiveness of the Proposed Techniques for Inference

Finally, we address RQ3 by evaluating the individual contributions of solving the training-inference mismatch and APG. To this end, we conduct two targeted ablation experiments on the LongCat-AudioDiT-1B backbone. In the first configuration (*training-inference mismatch*), we keep  $z_t^{ctx}$  as the model prediction and do not overwrite it with the GT noisy latent for inference. We also retain  $z_t^{ctx}$  to compute the unconditional velocity. In the second configuration (*w/o APG*), we replace the APG inference algorithm with standard CFG (Eq. 8). The comparative results are summarized in Table 4.

- **Impact of the training-inference mismatch:** The overall performance of the utterances synthesized by LongCat-AudioDiT-1B consistently and significantly outperforms those synthesized without solving the training-inference mismatch problem. This clear performance degradation validates the existence of the recognized problem and the effectiveness of our method to mitigate it.
- **Impact of APG:** While the baseline model employing standard CFG achieves comparable intelligibility (CER) and speaker similarity (SIM) scores, the integration of APG yields superior UTMOS and DNSMOS scores. This demonstrates that APG effectively mitigates the oversaturation artifacts inherent to high-scale CFG, thereby elevating the perceptual naturalness and overall acoustic quality of the synthesized speech.

## 6 Conclusion and Future Work

In this paper, we present LongCat-AudioDiT, a state-of-the-art non-autoregressive diffusion-based TTS model. The core advancement of LongCat-AudioDiT lies in modeling the generative process directly within the waveform latent space, bypassing intermediate acoustic representations such as mel-spectrograms widely adopted in prior literature. This unified design not only drastically simplifies the overall TTS pipeline but also fundamentally eliminates the compounding errors inherently caused by two-stage acoustic-to-waveform conversions. Furthermore, we introduce two critical improvements to the inference process: first, we identify and rectify a long-standing training-inference mismatch; second, we replace traditional CFG with APG to elevate generation quality.

Extensive experimental results demonstrate that LongCat-AudioDiT achieves new SOTA zero-shot speaker similarity on the rigorous Seed benchmark while maintaining competitive intelligibility. Notably, this is accomplished through an end-to-end approach, without relying on sophisticated multi-stage training pipelines or expensive high-quality human annotations. By outperforming previous diffusion-based baselines by a considerable margin, our work robustly validates the superiority of waveform-level latent modeling over traditional intermediate representations.

Finally, through comprehensive ablation studies, we systematically dissect the individual contributions of our proposed components. Most importantly, our deep dive into the interplay between the Wav-VAE’s reconstruction fidelity (e.g., varying dimensions and frame rates) and the downstream TTS generation quality reveals non-trivial trade-offs. We believe these empirical insights advance the understanding of the synergy between representation learning and generative modeling, shedding light on the future design of audio foundation models.

**Future Work** Promising directions for future research include pushing the performance ceiling via alignment-free reinforcement learning (RLHF for audio), and accelerating the inference speed through knowledge distillation techniques for real-time deployment.



## 7 Contributor

### Core Contributors

Detai Xin, Shujie Hu, Chengzuo Yang

### Tech Leads

Chen Huang, Guoqiao Yu, Guanglu Wan, Xunliang Cai

### Contributors

(Sorted in alphabetical order)

Disong Wang, Fengjiao Chen, Fengyu Yang, Hui Yang, Jiamu Li, Jun Wang, Qi Li, Qian Yang, Quanxiu Wang, Rumei Li, Shuaiqi Chen, Xu Xiang, Xuezhi Cao, Yi Chen, Yuchen Sun, Zheng Zhang, Zhiqing Hong, Ziwen Wang

### References

- Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.
- Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36, 2024.
- Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.
- Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.
- Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.
- Bowen Zhang, Congchao Guo, Geng Yang, Hang Yu, Haozhe Zhang, Heidi Lei, Jialong Mai, Junjie Yan, Kaiyue Yang, Mingqi Yang, et al. Minimax-speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder. *arXiv preprint arXiv:2505.07916*, 2025.
- Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, robust and controllable text to speech. *Proc. NeurIPS*, 32, 2019.
- Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE spoken language technology workshop (SLT)*, pages 682–689. IEEE, 2024a.
- Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024a.
- Keon Lee, Dong Won Kim, Jaehyeon Kim, Seungjun Chung, and Jaewoong Cho. Ditto-tts: Diffusion transformers for scalable text-to-speech without domain-specific factors. *arXiv preprint arXiv:2406.11427*, 2024.
- James Betker. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*, 2023.
- Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024a.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.
- Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409*, 2021.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.
- Ricky T. Q. Chen. torchdiffeq, 2018. URL <https://github.com/rtqichen/torchdiffeq>.
- Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11341–11345. IEEE, 2024.
- Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689. IEEE, 2024b.
- Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022a.
- Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.
- Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. *arXiv preprint arXiv:2506.13053*, 2025.
- Yuan Gao, Nobuyuki Morioka, Yu Zhang, and Nanxin Chen. E3 tts: Easy end-to-end diffusion-based text to speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023a.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- Yanqing Liu, Ruiqing Xue, Lei He, Xu Tan, and Sheng Zhao. Delightfultts 2: End-to-end speech synthesis with adversarial vector-quantized auto-encoders. In *Proc. Interspeech*, 2022b.
- Yongjoon Lee and Chanwoo Kim. Wave-u-mamba: an end-to-end framework for high-quality and efficient speech super resolution. In *Proc. ICASSP*, 2025.
- Chunyu Qiang, Hao Li, Yixin Tian, Yi Zhao, et al. High-fidelity speech synthesis with minimal supervision: All using diffusion models. In *Proc. ICASSP*, 2024.
- Zhikang Niu, Shujie Hu, Jeongsoo Choi, Yushen Chen, Peining Chen, Pengcheng Zhu, Yunting Yang, Bowen Zhang, Jian Zhao, Chunhui Wang, et al. Semantic-vae: Semantic-alignment latent representation for better speech synthesis. *arXiv preprint arXiv:2509.22167*, 2025.
- Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Forty-first International Conference on Machine Learning*, 2024.

- Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, 33:1583–1594, 2020.
- Chun Yat Wu, Jiajun Deng, Guinan Li, Qiuqiang Kong, and Simon Lui. Clear: Continuous latent autoregressive modeling for high-quality and low-latency speech synthesis. *arXiv preprint arXiv:2508.19098*, 2025.
- Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507, 2021.
- Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993, 2023.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc., 2020.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
- Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253, 2020.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6441–6451, 2024b.
- 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. *arXiv preprint arXiv:2410.06940*, 2024.
- Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. mhubert-147: A compact multilingual hubert model. *arXiv preprint arXiv:2406.06371*, 2024.
- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.
- Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*, 2023.
- Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023.
- Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.
- Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. ICLR*, 2018.
- Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *Proc. Interspeech*, 2019.

- Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Proc. ICASSP*, volume 2, pages 749–752. IEEE, 2001.
- 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.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
- Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. Funasr: A fundamental end-to-end speech recognition toolkit. In *Interspeech 2023*, pages 1593–1597, 2023b. doi:[10.21437/Interspeech.2023-1428](https://doi.org/10.21437/Interspeech.2023-1428).
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.
- Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.
- Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE, 2021.
- Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*, 2024.
- Xiaohui Sun, Ruitong Xiao, Jianye Mo, Bowen Wu, Qun Yu, and Baoxun Wang. F5r-tts: Improving flow-matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*, 2025.
- Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024b.
- Hao-Han Guo, Yao Hu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, and Kun Xie. Fireredtts-1s: An upgraded streamable foundation text-to-speech system. *arXiv preprint arXiv:2503.20499*, 2025.
- Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*, 2025a.
- Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, et al. Ditar: Diffusion transformer autoregressive modeling for speech generation. *arXiv preprint arXiv:2502.03930*, 2025.
- Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, et al. Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning. *arXiv preprint arXiv:2509.24650*, 2025b.
- SII-OpenMOSS. Moss-tts technical report. *arXiv preprint arXiv:2603.18090*, 2026.
- Hangrui Hu, Xinfu Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.
- Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.
- Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *arXiv preprint arXiv:2306.00814*, 2023.

- Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.
- Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. Bigcodec: Pushing the limits of low-bitrate neural speech codec. *arXiv preprint arXiv:2409.05377*, 2024.
- Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, et al. Vibevoice technical report. *arXiv preprint arXiv:2508.19205*, 2025.
- Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. Ace-step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*, 2025.