Continuous-Token Diffusion for Speaker-Referenced TTS in Multimodal LLMs

Xinlu He[1], **Swayambhu Nath Ray**[2], **Harish Mallidi**[2], **Jia-Hong Huang**[2], **Ashwin Bellur**[2], **Chander Chandak**[2], **M. Maruf**[2], **Venkatesh Ravichandran**[2]
[1]Worcester Polytechnic Institute, USA [2]Amazon AGI, USA

Abstract

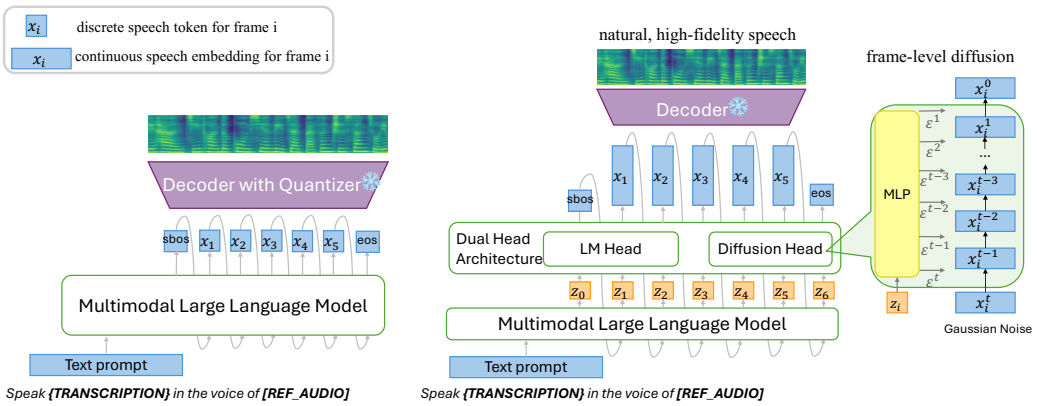
Unified architectures in multimodal large language models (MLLM) have shown promise in handling diverse tasks within a single framework. In the text-to-speech (TTS) task, current MLLM-based approaches rely on discrete token representations, which disregard the inherently continuous nature of speech and can lead to loss of fine-grained acoustic information. In this work, we investigate the TTS within the MLLM paradigm using continuous speech representations. We design a dual-head architecture and implement two complementary training strategies for a robust model. (1) A diffusion head generating continuous speech representations is added on the MLLM, which is on frame-level and strictly autoregressive. (2) The original language model head is retained to preserve multitask capability and to control the start and end of speech synthesis. (3) Masked training is employed to address exposure bias in autoregressive decoding. (4) To stabilize optimization, we propose a two-stage scheme where the LM is frozen in the second stage, ensuring the diffusion head learns from a fixed input distribution. Evaluations on LibriSpeech(PC) test-clean show that our approach achieves state-of-the-art autoregressive performance, with a WER of 1.95%, speaker similarity of 0.54, and UTMOS of 4.00. The two-stage training yields a 46% relative WER reduction over the one-stage training baseline. These results highlight the effectiveness of combining autoregressive modeling with continuous-token diffusion, supported by a two-stage training procedure.

1 Introduction

Recent advances in multimodal large language models (MLLMs) have enabled a single model to perform diverse tasks across modalities in an autoregressive manner [8, 33, 37]. In text-to-speech (TTS), the dominant approach converts speech into discrete tokens [11, 35], allowing TTS to be framed as a sequence prediction problem within the LLM framework. While effective, discrete quantization can discard fine-grained acoustic details, limiting naturalness and speech fidelity. In contrast, continuous speech representations—often learned by variational autoencoders (VAEs) or other self-supervised encoders [1, 6, 13]—better preserve the intrinsic properties of speech. Diffusion models, originally successful in high-fidelity image generation [12, 32, 39], have recently achieved state-of-the-art results for TTS by modeling such continuous representations in a non-autoregressive manner [14, 25, 28]. However, the integration of continuous-token diffusion into an autoregressive MLLM framework remains largely unexplored.

Building on this line, we propose to integrate diffusion directly into an autoregressive MLLM framework, as shown in Fig. 1. Prior autoregressive diffusion methods either rely on intermediate semantic tokens [34], or relax framewise causality in diffusion modules to predict multi-frame blocks per AR step [15, 24]. In contrast, our approach implements a strictly frame-by-frame autoregressive, continuous-representation diffusion head on top of an LLM backbone. This design enables the model

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Structured Probabilistic Inference & Generative Modeling (SPIGM).



(a) Discrete Speech Generation in MLLM

(b) Continuous Diffusion Speech Generation in MLLM

Figure 1: **Illustration of the proposed method**. Compared to discrete token-based generation in (a), our approach (b) adopts a dual-head MLLM with a diffusion head for frame-level autoregressive generation, enabling continuous speech synthesis with natural and high-fidelity quality.

to directly generate high-fidelity speech in the continuous speech representation space, avoiding the quantization bottleneck.

To maintain the MLLM's multi-task consistency, we designed a dual-head architecture. The diffusion head generates continuous speech embeddings each frame, which are then decoded to synthesize waveforms. The language model (LM) head retained from the original LLM predicts the start and end of speech tokens, enabling variable-length speech synthesis. This token-based control enables seamless integration of speech generation with multimodal generation. Unlike prior TTS methods [20, 26] introducing an external classifier module, our design preserves a single, unified framework within the MLLM.

Our investigation into continuous-token autoregressive TTS revealed two major training challenges: exposure bias and joint optimization instability. First, the mismatch between teacher-forced training and free-running inference causes small frame-level deviations to accumulate over long sequences. We mitigate this exposure bias [2] by introducing a masked training scheme inspired by audio generation [36], where a proportion of ground-truth frames is masked during training. This bridges the gap between training and inference conditions, improving robustness and temporal consistency in generated speech.

Then we observed that jointly optimizing the diffusion head with the LLM is unstable. The LLM's output distribution evolves during training, causing instability in the diffusion learning process. To address this, we employ a two-stage training strategy: in stage 1, train the LLM and diffusion head jointly, allowing the LLM to learn speech token prediction and the diffusion head to adapt to evolving inputs. In stage 2, we freeze the entire LLM side, including the backbone, the LM head, and the speech-projection, thereby fixing the input distribution to the diffusion head. We then train only the diffusion head. This allows the diffusion model to focus solely on refining the mapping from the LLM outputs to the target speech space. This separation stabilizes training and significantly improves generation quality.

Our main contributions are as follows:

- We introduce a frame-by-frame continuous-token diffusion head into an autoregressive MLLM for speaker-referenced TTS, distinguishing it from block-wise multi-frame designs.
- We propose a dual-head architecture where LM head supports variable-length speech and keeps unified multimodal framework.
- We mitigate autoregressive exposure bias via masked training, improving temporal consistency and model robustness.
- We stabilize training with a two-stage strategy, yielding large performance gains and cutting WER by 46%, reaching SOTA AR on LibriSpeech(PC) test-clean.[1]

[1]The models and results described in this paper are intended for research purposes only.

2 Related Work

Zero-shot TTS. Zero-shot text-to-speech (TTS) refers to synthesizing speech for previously unseen speakers by leveraging a short reference utterance as conditioning, thereby enabling speaker generalization without explicit speaker-specific training. Inspired by advances in large language models (LLMs) [4], zero-shot TTS is often formulated as a language modeling task [3, 40], where speech waveforms are transformed into sequences of tokens and synthesized via next-token prediction. Existing methods can be broadly categorized into multi-stage and single-stage pipelines. Multi-stage systems, such as VALL-E [35] and SALAD [34], autoregressively predict coarse units such as semantic [3] or codec tokens [38], which are then refined into waveforms. This decomposition improves stability but often discards fine-grained acoustic details. In contrast, single-stage approaches, exemplified by MegaTTS [16] and NaturalSpeech [17], directly generate high-information continuous representations, offering higher fidelity while facing greater challenges in robustness. Our method follows this single-stage paradigm.

Autoregressive Diffusion. Autoregressive language models were originally developed for discrete symbol sequences, whereas diffusion models are particularly effective for continuous data distributions [12]. Recent work has explored combining these paradigms for sequence generation. Several studies modify the diffusion process to behave autoregressively, for example by adjusting denoising schedules so earlier tokens are predicted before later ones [5, 9]. TransFusion [42] exemplifies this strategy with a shared transformer that applies causal attention to discrete tokens and bidirectional attention to continuous features, though it still struggles with strictly causal generation of continuous signals. Other efforts replace discrete codec units with continuous-valued tokens modeled directly through diffusion losses [21, 36]. By avoiding quantization, these approaches preserve fine-grained semantic and acoustic detail, positioning diffusion as a compelling alternative to conventional autoregressive modeling.

3 Proposed Method

We aim to autoregressively generate speech in the space of continuous acoustic embeddings. Our framework builds upon a large language model backbone and introduces a dual-head architecture. The first part of the method addresses how continuous speech tokens are generated: Section 3.1 introduces continuous-token generation with a diffusion head, and Section 3.2 presents EOS control in the dual-head architecture, which together form the foundation for continuous speech generation within a multitask unified foundation model setting. The second part focuses on improving robustness and overall performance: Section 3.3 describes masked autoregressive learning, which exposes the model to imperfect histories, and Section 3.4 details a two-stage optimization scheme, which stabilizes training by mitigating distribution drift.

3.1 Continuous-token Generation with Diffusion Head

With the inherent continuous nature of speech, recent work has begun to adopt continuous representation for speech generation in TTS. Compared to discrete codebook tokens, continuous representations preserve fine-grained characteristics, while decreasing the potential information loss [22]. Inspired by image [21] and audio [36] generation, we introduce a lightweight diffusion head on top of a causal foundation model to generate high-fidelity speech from continuous embeddings.

Formally, the target is a sequence of continuous speech embeddings at frame level $x = \{x_1, \ldots, x_N\}$, where $x_i \in \mathbb{R}^d$ denotes the speech embedding in frame i. An off-the-shelf variational autoencoder $V = \{V_E, V_D\}$ provides the waveform-embedding mapping: the encoder V_E extracts embeddings from waveform w, and the decoder V_D reconstructs audio from predicted embedding $\hat{x}$.

Inference. Fig. 1 (b) illustrates the framework. Given a prompt p with transcription and reference audio, multi-modal causal LLM $\mathcal{C}_\theta$ takes p and past predictions $\hat{x}_{<i}$ to autoregressively produce a hidden state as condition $z_i = \mathcal{C}_\theta(p, \hat{x}_{<i})$. This vector z_i conditions a diffusion head with MLP denoiser M_ϕ, which starts from Gaussian noise $x_i^t \sim \mathcal{N}(0, I)$ and iteratively denoises to produce the next embedding $\hat{x}_i = x_i^0$.

Training. During training, we sample a total timestep $t \sim U\{1, \ldots, T\}$ for adding noise and noise $\varepsilon \sim \mathcal{N}(0, I)$, and form a noised target $x_i^t = \sqrt{\bar{\alpha}_t} x_i + \sqrt{1 - \bar{\alpha}_t}\varepsilon$, where $\bar{\alpha}_t$ defines a noise schedule

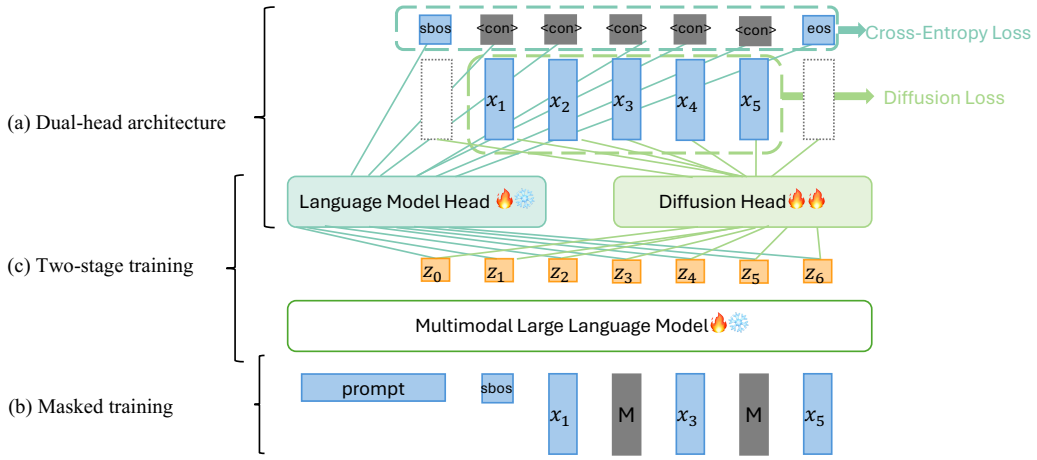


Figure 2: **Training and design details of our method.** (a) Dual-head architecture: The LM head predicts speech boundaries (<s-bos> and <eos>) and the special token <continuous_speech_gen>, while the diffusion head generates continuous frame-level speech embeddings. (b) Masked training strategy: A portion of speech inputs is masked to bridge the gap between teacher-forcing training and auto-regressive inference. (c) Two-stage training: In the first stage, all components are trained jointly; in the second stage, only the Diffusion head is further optimized.

[12, 27]. A small MLP denoiser M_ϕ predicts the noise $\hat{\varepsilon} = M_\phi(x_i^t,\ t,\ z_i)$, where x_i^t, t, z_i denotes the current state, the timestep, and the LLM condition z_i. We minimize the standard noise-prediction loss

$$\mathcal{L}_{\text{diff}}(\theta, \phi) \ = \ \mathbb{E}_t \left[\|\varepsilon - \hat{\varepsilon}\|^2 \right],$$

which backpropagates through z_i. This diffusion loss can be jointly optimized with the objective of the language model head. Further details are provided in the following sections.

3.2 EOS Control in Dual-Head Architecture

Prior speech generation approaches [35] typically rely on an auxiliary classifier or a fixed-length constraint to determine the endpoint of speech output. In contrast, our framework delegates boundary control to the LM head, enabling seamless integration into a unified multi-task backbone.

Inference. Generation proceeds under the control of the LM head. The model begins in the textual phase until the LM head emits a special token <*speech_bos*>, which triggers the speech generation phase. At each subsequent step, the LM head produces a control token. If it emits <*cont_speech_gen*>, this token is not added to the output sequence; instead, it signals the diffusion head to generate the next speech embeddings. Otherwise when LM head predicts <*eos*>, the speech generation phase terminates. This token-based mechanism provides a unified and modality-agnostic interface for switching between text and speech without additional architectural components.

Training. As shown in Fig 2, in order to supervise this control process, we extend the vocabulary with a special token <*cont_speech_gen*> in addition to <*speech_bos*> and <*eos*>. Although <*cont_speech_gen*> is not emitted during inference, its explicit supervision during training provides dense learning signals throughout the speech segment. This design reduces the risk of the LM head prematurely predicting <*eos*> in variable-length sequences, compared with supervising only the boundary tokens. The LM head is trained with the standard cross-entropy $\mathcal{L}_{\text{LM}}$ over the control tokens, while the diffusion head is trained with the noise-prediction loss $\mathcal{L}_{\text{diff}}$ (Sec. 3.1). The overall objective is the sum of the LM cross-entropy loss and the diffusion loss:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{diff}}$$

3.3 Masked Autoregressive Learning

Autoregressive generation is affected by exposure bias [2], where models are trained on ground-truth histories but must rely on its own potentially erroneous predictions during inference, resulting in

error accumulation. To mitigate this issue, various strategies such as masking have been explored in text generation [10] and audio generation [36]. Motivated by these advances, we adapt masking to continuous frame-level speech generation.

During training, before feeding the acoustic embedding sequence $x = \{x^1, \ldots, x^N\}$ into the causal predictor C_θ, we apply *zero embedding masking* to simulate imperfect histories. We define a binary mask $v = \{v^1, \ldots, v^N\},\ v^t \in \{0, 1\}$, where each entry is sampled independently as $v^t \sim \mathrm{Bernoulli}(1 - p_{\mathrm{mask}})$. Thus, with probability p_{mask}, the corresponding frame is masked and replaced by the zero vector. The corrupted sequence is then $\tilde{x} = x \odot v$, where $\odot$ denotes element-wise multiplication, so masked positions are replaced by the zero vector. This zero embedding masking strategy can be viewed as input-level masking, where the masking ratio p_{mask} explicitly controls the level of corruption in the autoregressive history. The training masking is shown in Fig. 2 (b). At inference time, no masking is applied. The model operates autoregressively on its own predictions. By training under such corrupted contexts, the model is encouraged to handle imperfect histories more robustly, thereby mitigating exposure bias and improving stability in long-form speech synthesis.

3.4 Two-Stage Scheme

When training the MLLM and diffusion head jointly in an end-to-end manner, we observed instability caused by distribution drift. The MLLM output distribution $p_\theta(z \mid p)$ evolves as the parameters θ are updated. The diffusion head $\mathcal{M}_\phi$ is expected to learn a mapping from z to continuous speech embeddings x. However, since the source distribution $p_\theta(z)$ is non-stationary during training, $\mathcal{M}_\phi$ must adapt to a shifting input space, making convergence unreliable and degrading generation quality. We hypothesize that freezing θ to fix $p_\theta(z)$ yields a stationary input distribution, allowing the diffusion head $\mathcal{M}_\phi$ to focus on modeling a stable transformation, thereby improving optimization stability and synthesis fidelity.

To implement this idea, we adopt a two-stage training strategy, as shown in Fig. 2(c). In **Stage 1**, we jointly train the causal LM $\mathcal{C}\theta$ and the diffusion head $\mathcal{M}\phi$ in an end-to-end manner by minimizing the sum of the cross-entropy loss and the diffusion loss. This stage enables the model to align the LM outputs with the target distribution and to produce coarse speech embeddings. However, we observe that although the overall training objective consistently decreases, autoregressive evaluation metrics exhibit a non-monotonic trend, improving initially but then deteriorating. This is consistent with our hypothesis that distribution drift hinders stable refinement.

In **Stage 2**, we freeze the MLLM and LM head parameters θ and train only the diffusion$\mathcal{M}\phi$. With $p\theta(z)$ fixed, the input distribution to the diffusion head remains stationary, allowing it to focus on refining LM outputs into high-fidelity acoustic frames. This stage stabilizes optimization, mitigates the instability observed in joint training, and leads to improved synthesis quality.

4 Experiments

To evaluate our dual-head continuous speech generation framework, we conduct TTS experiments under different settings, comparing intelligibility, speaker identity preservation, and speech naturalness. We first introduce the datasets and evaluation protocols in Sec. 4.1, then present implementation details in Sec. 4.2, and finally describe the baselines in Sec. 4.3.

4.1 Dataset and Metrics

Datasets We adopt the LibriVox corpus as the training source. Specifically, we use a 50k-hour subset (derived from the Libri-Light collection [18]), which consists of read English audiobooks from thousands of speakers. For evaluation, we use the Librispeech (PC) test-clean dataset. Following the protocol of NaturalSpeech [17], we randomly select one utterance from each of 40 speakers and use an additional 3-second clip as the speaker reference.

Evaluation metrics

We evaluate our system on three aspects: intelligibility, speaker identity preservation, and speech quality. (1) Word error rate (WER) measures intelligibility, about how accurately the synthesized speech conveys the reference text. The generated speech is transcribed by Whisper-Large [30], and WER is calculated from insertions, substitutions, and deletions. (2) Speaker similarity is

measured as cosine similarity between embeddings extracted by ECAPA-TDNN [7]. We report SIM-R (to reference prompt) and SIM-G (to ground-truth speech). (3) Speech quality is estimated with UTMOS [31], an objective MOS predictor trained on large-scale human ratings.

4.2 Implementation Details

Speech Representation. Our system relies on two types of speech representations: one for speaker reference prompting and another for generation targets. For speaker reference prompting, we extract a 768-dimensional embedding from a three-second reference clip using LAM [29]. This embedding encodes speaker identity and is projected into the LLM input space as conditioning information.

For generation targets, we adopt a pretrained VAE-based vocoder that encodes speech into 64-dimensional embeddings at 25 frames per second. Given a stereo waveform sampled at 48 kHz, the encoder applies a pre-convolutional projection to 128 channels, followed by five downsampling ResNet blocks, producing frame-level features at 25 Hz. A post-convolutional projection and VAE bottleneck is then applied to yield frame-level 64-dimensional latent vectors.

Model Details Our architecture builds on an autoregressive LLM backbone and extends it with projection modules for multi modality and a diffusion head for speech generation. We adopt OPT-125M [41] as the LLM backbone. For multi-modality, we add two projectors: one maps the 768-dimensional reference embedding extracted from the speaker prompt into the LLM input space, and the other maps the 64-dimensional continuous acoustic tokens into the LLM input space. With these projections, the backbone functions as a multimodal language model (MLLM) rather than a purely text-based LLM.

We use the final hidden state from the LLM decoder as the input to the diffusion module. Before entering the diffusion process, this hidden representation is projected through the third linear layer to 768 dimensions as the diffusion condition. The diffusion head is implemented as a stack of MLP layers and operates with a DDPM-based denoising process, similar to [21, 36].

Diffusion Hyperparameters During training, we use a diffusion process with $T = 1000$ steps and adopt the cosine noise schedule [27],where derives β_t implicitly from the cumulative product $\bar{\alpha}_t$. The diffusion head is an MLP with residual blocks; we experimented with 3, 6, and 12 layers, and report 12-layer results unless otherwise noted. Each block consists of layer normalization, linear layers, and SiLU activation with adaptive layer normalization modulation, with no dropout. During inference, we reduce the denoising process to 100 steps and apply a sampling temperature of 0.9. CFG is set to 1.

Training Configuration. All experiments are conducted on NVIDIA A100 GPUs with a global batch size of 2048. We use the Adam optimizer without weight decay and employ FP16 mixed precision for efficiency. In stage 1, the learning rate is linearly warmed up from 3×10^{-5} to 3×10^{-4} over the first steps and then decayed to zero using a cosine schedule, for a total of 300k steps. In stage 2, the model is further trained for 300k steps with a constant learning rate of 2×10^{-4}.

4.3 Baseline

We use the model from the first-stage joint training of all components, including the MLLM backbone, LM head and diffusion head, as our baseline. During training, the loss decreases monotonically, but the evaluation WER first decreases and then increases because of the dynamic condition for the diffusion head. We therefore apply the early stopping and select the checkpoint with the lowest validation WER as the baseline model, from which the stage-2 training is initialized.

5 Results

We first present the main results by comparing our model with representative baselines. We then provide ablations and analyses of key design choices and inference settings to understand their impact on intelligibility, speaker similarity, and naturalness.

5.1 Main Results

Table 1 reports the comparison between our model and representative hybrid autoregressive baselines. VALL-E, which relies on discrete tokens, yields a WER of 6.11% and a speaker similarity of 0.47,

Table 1: Objective evaluation on LibriSpeech(PC) test-clean. Results are reported for WER (%), speaker similarity (cosine SIM), and UTMOS. † denotes results reproduced by NaturalSpeech3. Our method outperforms larger hybrid AR+NAR baselines while using fewer parameters.

Method	**Modeling**	**Token**	**# Params**	**WER(%)↓**	**SIM↑**	**UTMOS↑**
Ground Truth	-	-	-	2.84	0.69	4.16
Vocoder	-	-	-	2.56	0.61	3.82
VALL-E† ([35])	AR+NAR	Discrete	400M	6.11	0.47	3.68
Mega TTS† ([16])	AR+NAR	Continuous	500M	2.32	0.53	**4.02**
Voicebox† ([19])	NAR	Continuous	400M	2.14	0.48	3.73
StyleTTS2† ([23])	NAR	Continuous	700M	2.49	0.38	3.94
Stage-1 Baseline	AR	Continuous	160M	3.61	0.49	3.21
Proposed Method	AR	Continuous	160M	**1.95**	**0.54**	4.00

illustrating the limitations of quantization in preserving fine-grained acoustic details. MegaTTS, a continuous-token model with 500M parameters, achieves stronger results with a WER of 2.32% and a similarity of 0.53. By contrast, our model attains a WER of 1.95% and a similarity of 0.54, while maintaining competitive perceptual quality (UTMOS 4.00 compared with 4.02 for MegaTTS). Despite using only 160M parameters, our system consistently outperforms larger models, demonstrating the efficiency of combining an autoregressive backbone with a continuous diffusion head. For additional context, we also include ground-truth speech and vocoder reconstructions in the table for reference.

We further analyze the effect of our two-stage training strategy. The Stage-1 baseline achieves a WER of 3.61%, a speaker similarity of 0.49, and a UTMOS of 3.21, indicating that the diffusion head initially suffers from unstable input distributions. After Stage-2 training, the WER is reduced by 46% relative (from 3.61% to 1.95%), while speaker similarity increases from 0.49 to 0.54 and UTMOS rises from 3.21 to 4.00. These improvements show that stabilizing the diffusion head's input distribution in Stage-2 not only reduces recognition errors but also enhances both speaker consistency and perceived naturalness.

5.2 Ablation and Analysis

To better understand the contributions of different components and design choices, we conduct a series of ablation and analysis experiments. We first investigate the effect of masked training, which aims to mitigate exposure bias during autoregressive decoding. We then examine the role of diffusion head capacity and the impact of our two-stage training strategy. Finally, we analyze the influence of stopping criteria and inference hyperparameters, highlighting how these factors jointly affect intelligibility, speaker similarity, and naturalness.

Masked Ratio. Table 2 shows the impact of different masking rates in the masked training scheme. Without masking (0%), the model suffers from severe exposure bias, leading to a WER of 15.06%. Introducing moderate masking improves robustness, with the best performance at 30% masking (WER 6.17%, UTMOS 3.21). However, excessive masking (50%) degrades both intelligibility and naturalness, as too much corruption disrupts semantic alignment. This confirms that moderate masking helps bridge the gap between training and inference conditions.

Table 2: Performance with different masking rates, using a 3-layer MLP diffusion head.

Masking Rate(%)	**WER (%)↓**	**SIM-R↑**	**SIM-G↑**	**UTMOS ↑**
0	15.06	0.45	0.42	2.00
15	12.65	0.45	0.42	1.39
30	6.17	0.46	0.43	3.21
50	8.13	0.46	0.43	2.84

Diffusion Head Depth. Table 3 compares different diffusion head depths with and without the proposed two-stage training. Increasing the depth from 3 to 12 layers progressively improves WER

and speaker similarity, demonstrating the benefit of a stronger decoder. More importantly, enabling two-stage training further reduces WER to 1.95% and boosts similarity and naturalness, highlighting the effectiveness of stabilizing the diffusion head with a fixed input distribution.

Table 3: Comparison of different numbers of MLP layers in the diffusion head, with dropout rate set to 30%. Stage-2 FT indicates whether two-stage fine-tuning is applied.

# MLP	Stage-2 FT	# Params	WER (%)↓	SIM-R↑	SIM-G↑	UTMOS ↑
3	w/o	148.7M	6.17	0.46	0.43	3.10
6	w/o	164.4M	5.12	0.50	0.46	3.10
12	w/o	159.9M	3.61	0.49	0.46	3.21
12	w	159.9M	1.95	0.54	0.50	4.00

Stopping Criteria. Table 4 compares different stopping strategies. Using ground-truth durations causes unstable outputs (WER 29.36%), while oracle endpoint supervision achieves low WER but requires non-causal labels. Our EOS-token design achieves comparable WER and UTMOS without relying on oracle information, while maintaining stable generation speed, making it a practical choice for unified MLLM-based TTS.

Table 4: Performance with different stopping criteria. GT-Dur.: using oracle duration; GT-EP.: using oracle end-of-speech (oracle stopping point); EOS Token: our method, where the LM head predicts an end-of-sequence token during inference.

Stopping Criteria	WER (%)↓	SIM-R↑	SIM-G↑	UTMOS↑
GT-Dur.	29.36	0.48	0.43	2.55
GT-EP.	3.46	0.49	0.46	3.21
EOS Token	3.61	0.49	0.46	3.21

Inference Hyperparameters. Table 5 analyzes the influence of diffusion inference parameters. Lower temperatures tend to produce cleaner but truncated outputs, leading to higher WER and lower similarity. Conversely, higher temperatures improve diversity but may reduce naturalness. We find that a temperature of 0.9 with 100 denoising steps achieves the best trade-off, yielding the lowest WER (1.95%), highest similarity (0.54), and best UTMOS (4.00).

Table 5: Performance under different inference parameters.

Temperature	Inference Steps	WER(%)↓	SIM-R↑	SIM-G↑	UTMOS↑
1	200	15.06	0.47	0.44	2.40
1	100	7.53	0.48	0.44	3.27
0.9	100	1.95	0.54	0.50	4.00
0.8	100	16.11	0.45	0.41	3.01
0.8	80	19.88	0.44	0.39	4.07

6 Conclusion

In this work, we present a dual-head multimodal language model that integrates a frame-level continuous-token diffusion head with an autoregressive LLM backbone for speaker-referenced TTS. By combining continuous speech representations, our approach avoids the quantization bottleneck and achieves high-fidelity and natural speech. To overcome exposure bias and improve training performance, we adapt masked training and a two-stage optimization scheme, which together substantially improve robustness and quality. Evaluations on LibriSpeech(PC) demonstrate significant gains, including a 46% relative WER reduction over our baseline, along with higher speaker similarity and audio quality. These results highlight the effectiveness of bridging autoregressive modeling with diffusion-based refinement for continuous speech generation. Looking ahead, this framework provides a path toward unified foundation models that can support multiple speech and multimodal tasks within a single framework.

References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477.

[2] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks, 2015. URL https://arxiv.org/abs/1506.03099.

[3] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour. Audiolm: a language modeling approach to audio generation, 2023. URL https://arxiv.org/abs/2209.03143.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

[5] T. Chen, R. Zhang, and G. Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning, 2023. URL https://arxiv.org/abs/2208.04202.

[6] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2041–2053, 2019. doi: 10.1109/TASLP.2019.2938863.

[7] B. Desplanques, J. Thienpondt, and K. Demuynck. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pages 3830–3834, 2020.

[8] J.-B. A. et al. Flamingo: a visual language model for few-shot learning. 2022.

[9] Y. Fang, J. Bai, J. Wang, and X. Zhang. Vector quantized diffusion model based speech bandwidth extension, 2024. URL https://arxiv.org/abs/2409.05784.

[10] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models, 2019. URL https://arxiv.org/abs/1904.09324.

[11] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan. Promptttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096285.

[12] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

[13] W.-N. Hsu, Y. Zhang, and J. Glass. Unsupervised learning of disentangled and interpretable representations from sequential data, 2017. URL https://arxiv.org/abs/1709.07902.

[14] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 2595–2605, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3547855. URL https://doi.org/10.1145/3503161.3547855.

[15] D. Jia, Z. Chen, J. Chen, C. Du, J. Wu, J. Cong, X. Zhuang, C. Li, Z. Wei, Y. Wang, and Y. Wang. Ditar: Diffusion transformer autoregressive modeling for speech generation, 02 2025.

[16] Z. Jiang, Y. Ren, Z. Ye, J. Liu, C. Zhang, Q. Yang, S. Ji, R. Huang, C. Wang, X. Yin, Z. Ma, and Z. Zhao. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias, 2023. URL https://arxiv.org/abs/2306.03509.

[17] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X.-Y. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and S. Zhao. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models, 2024. URL https://arxiv.org/abs/2403.03100.

[18] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673, 2020. https://github.com/facebookresearch/libri-light.

[19] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu. Voicebox: Text-guided multilingual universal speech generation at scale, 2023. URL https://arxiv.org/abs/2306.15687.

[20] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou. Neural speech synthesis with transformer network, 2019. URL https://arxiv.org/abs/1809.08895.

[21] T. Li, Y. Tian, H. Li, M. Deng, and K. He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024.

[22] Y. Li, R. Xie, X. Sun, Y. Cheng, and Z. Kang. Continuous speech tokenizer in text to speech, 2025. URL https://arxiv.org/abs/2410.17081.

[23] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models, 2023. URL https://arxiv.org/abs/2306.07691.

[24] Z. Liu, S. Wang, S. Inoue, Q. Bai, and H. Li. Autoregressive diffusion transformer for text-to-speech synthesis, 2024. URL https://arxiv.org/abs/2406.05551.

[25] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *Proc. ICASSP*, 2024.

[26] L. Meng, L. Zhou, S. Liu, S. Chen, B. Han, S. Hu, Y. Liu, J. Li, S. Zhao, X. Wu, H. Meng, and F. Wei. Autoregressive speech synthesis without vector quantization, 2025. URL https://arxiv.org/abs/2407.08551.

[27] A. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2102.09672.

[28] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:234483016.

[29] Prabash Reddy Male and Swayambhu Nath Ray and Harish Arsikere and Akshat Jaiswal and Prakhar Swarup and Prantik Sen and Debmalya Chakrabarty and K V Vijay Girish and Nikhil Bhave and Frederick Weber and Sambuddha Bhattacharya and Sri Garimella. DuRep: Dual-Mode Speech Representation Learning via ASR-Aware Distillation. In *Interspeech 2025*, pages 5808–5812, 2025. doi: {10.21437/Interspeech.2025-1242}.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

[31] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. pages 4521–4525, 09 2022. doi: 10.21437/Interspeech.2022-439.

[32] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

[33] V. A. Trinh, R. Southwell, Y. Guan, X. He, Z. Wang, and J. Whitehill. Discrete multimodal transformers with a pretrained large language model for mixed-supervision speech processing, 2024. URL https://arxiv.org/abs/2406.06582.

[34] A. Turetzky, N. Shabtay, S. Shechtman, H. Aronowitz, D. Haws, R. Hoory, and A. Dekel. Continuous speech synthesis using per-token latent diffusion, 2024. URL https://arxiv.org/abs/2410.16048.

[35] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023. URL https://arxiv.org/abs/2301.02111.

[36] S. wen Yang, B. Kim, K.-P. Huang, Q. Tang, H. Phan, B.-R. Lu, H. Sundar, S. Ghosh, H. yi Lee, C.-C. Kao, and C. Wang. Generative audio language modeling with continuous-valued tokens and masked next-token prediction, 2025. URL https://arxiv.org/abs/2507.09834.

[37] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. A survey on multimodal large language models. *National Science Review*, 11(12), Nov. 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403. URL http://dx.doi.org/10.1093/nsr/nwae403.

[38] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi. Soundstream: An end-to-end neural audio codec, 2021. URL https://arxiv.org/abs/2107.03312.

[39] C. Zhang, C. Zhang, M. Zhang, I. S. Kweon, and J. Kim. Text-to-image diffusion models in generative ai: A survey, 2024. URL https://arxiv.org/abs/2303.07909.

[40] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities, 2023. URL https://arxiv.org/abs/2305.11000.

[41] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

[42] C. Zhou, L. YU, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SI2hI0frk6.