

GLM-TTS Technical Report

Jiayan Cui^{*1} Zhihan Yang^{*1} Naihan Li¹ Jiankun Tian¹ Xingyu Ma¹
 Yi Zhang¹ Guangyu Chen¹ Runxuan Yang¹² Zijian Huang¹ Yuqing Cheng¹
 Yizhi Zhou¹ Guochen Yu^{†1} Xiaotao Gu¹ Jie Tang²

¹Zhipu AI ²Tsinghua University

*Equal contribution †Project leader

Code: github.com/zai-org/GLM-TTS
 Model: huggingface.co/zai-org/GLM-TTS
 Demo: audio.z.ai/

ABSTRACT

This work proposes GLM-TTS, a production-level TTS system designed for efficiency, controllability, and high-fidelity speech generation. GLM-TTS follows a two-stage architecture, consisting of a text-to-token autoregressive model and a token-to-waveform diffusion model. With only 100k hours of training data, GLM-TTS achieves state-of-the-art performance on multiple open-source benchmarks. To meet production requirements, GLM-TTS improves speech quality through an optimized speech tokenizer with fundamental frequency constraints and a GPO-based multi-reward reinforcement learning framework that jointly optimizes pronunciation, speaker similarity, and expressive prosody. In parallel, the system enables efficient and controllable deployment via parameter-efficient LoRA-based voice customization and a hybrid phoneme–text input scheme that provides precise pronunciation control. Our code is available at <https://github.com/zai-org/GLM-TTS>. Real-time speech synthesis demos are provided via Z.ai (audio.z.ai/), the Zhipu Qingyan app/web (chatglm.cn).

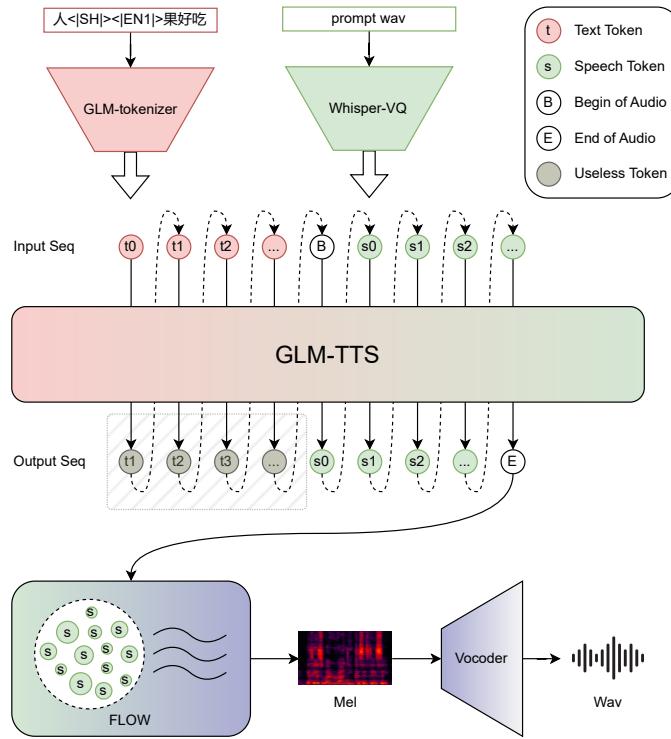


Figure 1: Overall Architecture of GLM-TTS

1 INTRODUCTION

Text-to-Speech (TTS) synthesis has evolved into a cornerstone technology powering human-computer interaction, content creation, and accessibility tools, from virtual assistants and podcast production to educational narration and dubbing. Over the past decade, TTS systems have evolved from early neural acoustic models that directly predict continuous speech representations—such as attention-based sequence-to-sequence models (e.g., Tacotron (Wang et al., 2017), Transformer-TTS (Li et al., 2019)) and non-autoregressive feed-forward architectures (e.g., Fastspeech (Ren et al., 2019) series) to Transformer-based (Vaswani et al., 2023) large language model (LLM)-driven paradigms that treat speech generation as discrete token modeling. This transition has enabled substantial progress in in-context learning (ICL) zero-shot voice cloning, multilingual synthesis, and prosodic naturalness.

Recent advanced TTS models can be broadly categorized into three paradigms: (1) auto-regressive (AR) zero-shot TTS based on neural codec such as SoundStream (Zeghidour et al., 2021) and En-Codec (Défossez et al., 2022) (e.g., VALL-E (Wang et al., 2023a), Spark-TTS (Wang et al., 2025b)); (2) flow-matching or diffusion-based non-autoregressive (NAR) paradigms (e.g., E3-TTS (Gao et al., 2023a), F5-TTS (Chen et al., 2025), F5R-TTS (Sun et al., 2025), et al.); (3) hybrid AR-NAR architectures (e.g., CosyVoice (Du et al., 2024a), FireRedTTS (Guo et al., 2025), Seed-TTS(Anastassiou et al., 2024), MiniMax-Speech (Zhang et al., 2025), and IndexTTS2 (Zhou et al., 2025a), et al.), which have collectively narrowed the gap between synthetic and human speech.

Despite these achievements, recent state-of-the-art (SOTA) TTS systems still face critical challenges that hinder production-level deployment. First, high-quality voice cloning typically requires large-scale training data and relatively long reference recordings, limiting applicability in low-resource scenarios. Second, emotional expressiveness remains constrained. Most models either fail to capture nuanced text-related emotions or rely on explicit emotion labels that complicate workflows and generalization. Third, the precision of pronunciation for polyphonic characters, rare words, and dialects remains suboptimal, especially in languages such as Chinese with rich phonetic variations. Fourth, reinforcement learning (RL), while promising for aligning speech outputs with human preferences, is underexplored in TTS due to difficulties in reward design and training stability. Finally, adapting premium or personalized voices often relies on costly full-model fine-tuning, making scalable customization impractical in production settings.

To address these limitations, we present GLM-TTS, a production-level TTS system optimized for efficiency, controllability, and naturalness. As shown in Figure 1, built on a two-stage generation paradigm inspired by CosyVoice (Text-to-Token Autoregressive + Token-to-Wav Diffusion), GLM-TTS achieves SOTA performance on open-source benchmarks with only 100k hours of training data, which is far less than large-scale counterparts like CosyVoice 3 (1M hours) (Du et al., 2025) and FireRedTTS-2 (1.1M hours) (Xie et al., 2025).

Our contributions are structured around the practical requirements of industrial TTS deployment:

- **Speech Tokenizer:** Leveraging an optimized Whisper-VQ speech tokenizer with fundamental frequency constraints and expanded vocabulary (32k), GLM-TTS achieves high speaker similarity (SIM = 76.1) and low character error rate (CER = 1.03%) on Seed-TTS-eval zh test-set.
- **Multi-Reward Reinforcement Learning:** Adopting a GRPO-based RL framework, we fuse four critical rewards (CER for pronunciation accuracy, SIM for timbre fidelity, Emotion for expressive naturalness, and Laughter for paralinguistic realism) with dynamic sampling and gradient clipping. This resolves the reward hacking and training instability issues plaguing prior RL-based TTS, enabling GLM-TTS to outperform commercial models in nuanced emotion expression (e.g., happy, sadness, and anger) on the CV3-eval-emotion benchmark and to achieve superior WER and SIM metrics on the Seed-TTS-eval benchmark as well.
- **Low-Cost Premium Voice Customization:** Through optimized LoRA fine-tuning, GLM-TTS achieves full-model-level performance by adjusting only 15% of parameters—reducing data requirements to 1 hour of single-speaker audio and training costs by 80% compared to full fine-tuning.

- Precision Pronunciation Control: A “Hybrid Phoneme + Text” input scheme with a dynamically extensible dictionary addresses polyphonic and rare word ambiguities, a longstanding challenge in Chinese TTS, without sacrificing prosodic naturalness.
- Enhanced Waveform Reconstruction: A novel Vocos2D vocoder replaces 1D convolutions with 2D operations and DiT-style residual connections, improving frequency subband modeling. Mixed training with high-quality singing data expands the vocal range and adapts the model to complex acoustic conditions.

2 METHODOLOGY

2.1 DATA PROCESSING PIPELINE

Leveraging proprietary audio datasets, we have constructed a comprehensive and robust data processing pipeline to generate high-quality audio data for subsequent model training. The data pipeline consists of the following steps:

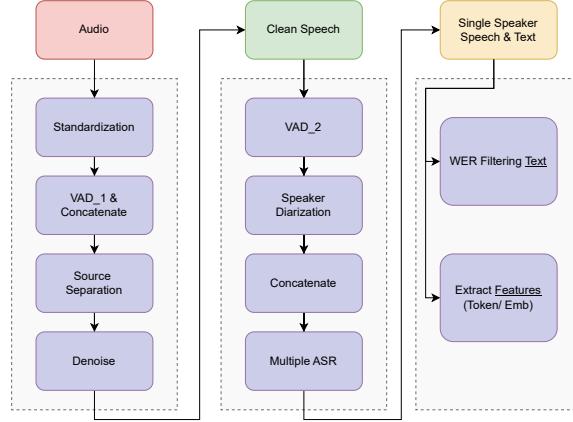


Figure 2: An overview of the data processing pipeline.

- **Speech Standardization and Coarse Segmentation.** First, we unify heterogeneous audio data into the WAV format to eliminate format-related inconsistencies. Then, Voice Activity Detection (VAD) technology (Bredin et al., 2019) is applied to segment valid speech fragments from the original audio, and these valid fragments are then concatenated into long audio clips of approximately 10 minutes for subsequent processing steps.
- **Source Separation and Denoising.** To obtain clean speech signals, we first adopt the Mel-Band Roformer model (an improved variant of the RoFormer model (Wang et al., 2023b)) to separate background sounds from the speech signals. After that, a self-developed denoising model is used to further suppress residual noise, ensuring the purity of the speech data.
- **Speaker Diarization and Concatenation.** We leverage the pyannote.audio model (Bredin et al., 2019) to implement multi-speaker separation, which can accurately distinguish speech fragments from different speakers. For the speech fragments of a single speaker, we perform amplitude normalization to ensure consistent volume levels, and then concatenate these fragments until the total length reaches the target duration (capping the sequence length at 40 seconds).
- **WER Filtering.** To select high-quality audio data, we conduct speech recognition and Word Error Rate (WER) calculation with a double-check mechanism:
 - For Chinese audio: Open-source Automatic Speech Recognition (ASR) models including Paraformer (Gao et al., 2023b) and SenseVoice (An et al., 2024) are used for transcription.

-
- For English audio: Open-source ASR models including Whisper (Radford et al., 2022) and Reverb (Bhandari et al., 2025) are adopted for transcription.

We calculate the WER for the transcribed text and retain only the audio data with a WER of less than 5% to ensure high accuracy of the speech-content correspondence.

- **Punctuation Optimization.** First, we perform text-speech forced alignment (Kürzinger et al., 2020) to obtain the pronunciation duration of each character in the transcribed text. Then, we calculate the pronunciation threshold as the sum of the mean and 2.6 times the variance of the character pronunciation durations. Finally, we optimize the punctuation based on the interval between adjacent characters: If the interval exceeds the threshold, we retain or add punctuation. Otherwise, punctuation is omitted.
- **Feature Extraction.** Based on the filtered single-speaker clean audio data, we extract audio speaker embeddings and speech tokens to support the training of subsequent models.
- **Overall Engineering Optimization.** To enable large-scale data processing efficiently, we adopt the gRPC framework (Google, 2015) and a server-worker architecture to accelerate each sub-module of the pipeline. Additionally, we rely on a distributed cluster to fully utilize the memory of multiple GPUs and the batch processing capability, significantly improving the overall processing efficiency of the pipeline.

2.2 TEXT TOKENIZER

Vocabulary Pruning for Alignment Stability. To alleviate the modeling burden of semantic-to-acoustic alignment, we prune the tokenizer’s vocabulary by removing tokens composed of more than two Chinese characters. Although we implement heuristic constraints during sampling to bound the speech-to-text length ratio (e.g., within [2, 20]), relying solely on such hard constraints is insufficient for optimal convergence. Long text tokens inherently exhibit high variance in acoustic duration and frequently stretch the upper bounds of the ratio, creating sparse and difficult-to-learn mappings. By enforcing a finer text granularity, we intrinsically normalize the information density and center the length ratio distribution. This structural optimization effectively alleviates the burden of learning extreme one-to-many text-to-acoustic alignments, ensuring robust autoregressive generation even at a high speech token rate of 25 Hz.

2.3 SPEECH TOKENIZER

GLM-TTS introduces a series of optimizations to the Whisper-VQ speech tokenizer (Radford et al., 2022) based on GLM-4-Voice (Zeng et al., 2024), aimed at improving pronunciation accuracy, naturalness, and expressiveness:

- **Increased Token Generation Rate.** The token rate is doubled from 12.5Hz to 25Hz, and the vocabulary size is expanded from 16k to 32k. This enhancement effectively reduces pronunciation glitches at high speaking speeds and improves the naturalness of paralinguistic features such as laughter and breathing sounds.
- **Introduction of Pitch Estimator (PE) Module.** A new pitch estimation module has been added to optimize pitch modeling accuracy, thereby improving the prosody alignment between the cloned TTS output and reference (prompt) audio.
- **Adoption of Non-Causal Architecture.** The original causal constraint has been lifted. The block attention structure was removed, and causal convolution has been replaced with standard convolution, removing sequential bottlenecks and improving the accuracy of both the ASR and PE modules.
- **Expanded Training Data Scope.** The scale and diversity of training data are significantly increased. Large-scale dialect datasets have been incorporated to strengthen dialect comprehension, and high-quality singing voice data have been added to enrich the model’s phonetic learning samples, further enhancing adaptability to diverse scenarios.

2.4 SPEECHLM RL-ALIGNMENT

Reinforcement learning has not yet been widely applied in speech synthesis, with major bottlenecks stemming from the complexity of reward mechanism design and the propensity for gradient vanish-

ing or performance degradation during training. GLM-TTS addresses these challenges by introducing the GRPO (Shao et al., 2024) reinforcement learning paradigm along with a series of strategies, significantly enhancing the core capabilities of both pre-trained and SFT models—including pronunciation accuracy, timbre similarity, and overall naturalness. Furthermore, GLM-TTS achieves superior human-like qualities, notably in emotional expressivity and the naturalness of paralinguistic features. Figure 3 illustrates the GLM-TTS-GRPO framework. Leveraging the GRPO algo-

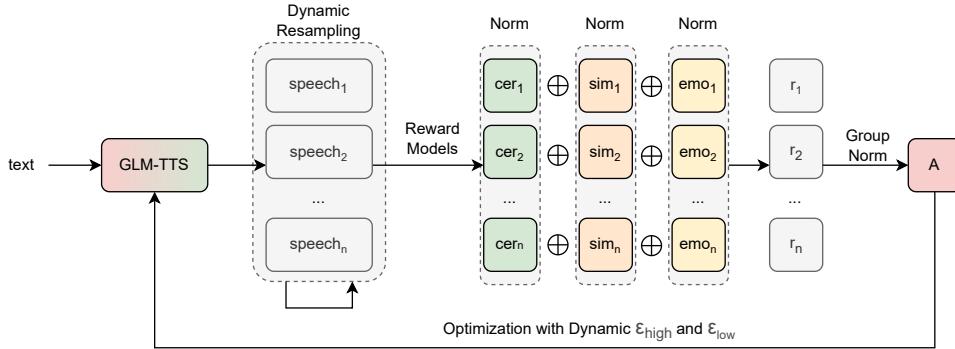


Figure 3: An overview of the GLM-TTS-GRPO framework.

rithmic framework, we achieve significant performance improvements through three major design advancements:

First, we introduce a multidimensional regularization reward mechanism, integrating four core rewards—CER (character error rate), SIM (similarity), Emotion, and Laughter. By employing a hierarchical processing strategy (“individual reward regularization → weighted fusion → overall regularization”), we effectively address the issue of different reward distributions.

Second, we implement a dynamic sampling strategy to mitigate potential gradient vanishing within a batch. This mechanism automatically triggers resampling (up to three times) when batch rewards become homogeneous, while limiting the total number of resampling to avoid negative optimization from poor-quality samples, thereby balancing training stability and efficiency.

Third, we adopt an adaptive gradient clipping scheme, setting ϵ_{high} and ϵ_{low} as dynamic values that adjust according to training steps. In the early stages, tighter clipping prevents the model from quickly exploiting reward hacking shortcuts; in the later stages, the constraints are gradually relaxed to allow for broader exploration. Reasonable parameter ranges ensure the number of clipped tokens remains stable, preventing ineffective clipping or excessive restriction. Additionally, setting $\epsilon_{high} > \epsilon_{low}$ encourages the generation of low-probability tokens, significantly enhancing the human-likeness of synthesized speech.

2.5 LORA FOR PREMIUM VOICE CUSTOMIZATION

Full parameter fine-tuning in large speech models is often costly and unstable due to uneven data quality. To address this, we optimize the LoRA (Low-Rank Adaptation) fine-tuning paradigm for “Premium Voice Customization”. This approach aims to achieve stable, high-quality voice customization with low resource and cost overhead.

- **Efficiency and Efficacy.** By fine-tuning only about 15% of the core backbone parameters for approximately 100 epochs, we achieve voice similarity and naturalness comparable to full parameter fine-tuning. This is a significant improvement over initial explorations where tuning only 0.3% – 5% of parameters resulted in limited improvement in voice style and emotion.
- **Cost Efficiency and Data Robustness.** Customization requires only about 1 hour of high-quality single-speaker audio, significantly lowering development costs and barriers. This streamlined process eliminates the need for large-batch data testing and complex data

matching and quality filtering, which are often problematic due to the high inconsistency in data distribution and quality in SFT.

- **Stability.** Controlling the ratio of fine-tuned parameters (e.g., above 15%) enhances generalization and stability across different scenarios, ensuring production-level reliability. The improved stability is crucial as complex requirements for small-batch, high-demand voice customization are difficult to implement effectively using full parameter fine-tuning.

2.6 PHONEME-IN

In professional speech synthesis scenarios, such as education and standardized testing, there is an exceptionally high demand for pronunciation accuracy. Traditional large-scale TTS models typically rely on automatic sampling or default probabilities when handling complex linguistic features like polyphones (characters with multiple pronunciations) and rare characters. This lack of explicit control mechanisms often results in uncontrollable pronunciation and higher error rates. To address this, we introduce **Phoneme-in**, an enhancement capability that utilizes phoneme-level input to achieve precise, controllable pronunciation.

Vocabulary Construction and Regularization. We construct dedicated vocabularies for polyphones and rare characters to aggregate terms requiring precise control in key application scenarios. These vocabularies support the downstream logic for targeted phoneme replacement and allow for dynamic maintenance and expansion based on specific business requirements.

Hybrid Training Paradigm. To equip the model with the ability to understand and adapt to phoneme inputs, we employ a mixed-modality training strategy:

- For standard characters (excluding defined polyphones and rare characters), we employ a **two-stage probabilistic replacement strategy**. During training, the replacement process is triggered with a specific probability (e.g., $p = 0.2$). Once triggered, a random subset of characters is converted into phonemes, where the replacement ratio is uniformly sampled between 0 and a maximum threshold (e.g., 0.5). This dynamic “Hybrid Phoneme + Text” augmentation significantly enhances the model’s robustness to mixed-modality inputs.
- Characters belonging to the polyphone or rare character vocabularies are preserved as original text without conversion during training, ensuring the model retains semantic context for these complex cases.

Fine-grained Control at Inference. The inference process is designed to maximize precision:

1. The system first processes the entire input sentence through a Grapheme-to-Phoneme (G2P) module to generate a complete phoneme sequence (phoneme_list).
2. It iterates through the original text list; if a polyphone or rare character is encountered, the text is replaced with its corresponding phoneme from the generated list.
3. The final input to the model takes a “hybrid phoneme+text” format. This approach ensures precise pronunciation control for ambiguous characters while preserving the natural prosody associated with the standard text.

2.7 VOCOS2D

The original Vocos (Siuzdak, 2024), a GAN-based vocoder, uses 1D convolutions in its generator to process entire frames across all frequencies. Drawing inspiration from sub-band processing and image-processing techniques, we redesign the generator to incorporate 2D convolutions, enabling more focused handling of specific frequency subbands. Figure 4 illustrates the Vocos2D generator architecture.

For the input X_{in} , an initial point-wise convolution (implemented as a fully connected linear layer) is followed by learned per-frequency embeddings to facilitate inter-frequency communication, as translation invariance does not apply across frequency bins.

The Vocos2D backbone block adapts the original ConvNeXt design, augmented with additional shortcut connections from the input Mel spectrogram X_{in} . These are regressed through a linear layer

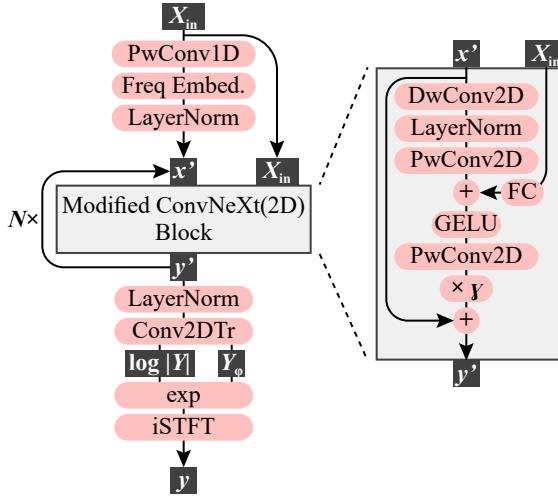


Figure 4: Diagram of the Vocos2D generator architecture. FC denotes a fully connected (linear) layer, PwConv denotes point-wise convolution, and DwConv denotes depth-wise convolution.

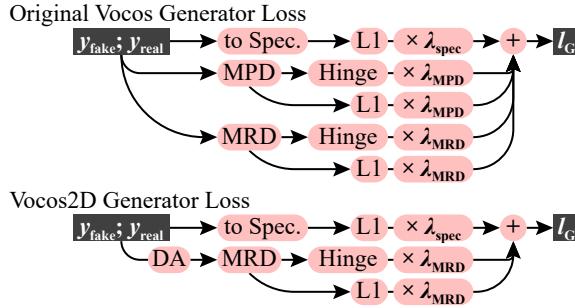


Figure 5: Comparison of the original Vocos generator loss (upper) and the proposed Vocos2D generator loss (lower). DA denotes discriminator augmentation, Hinge denotes hinge loss, MPD denotes multi-period discriminator, and MRD denotes multi-resolution discriminator.

and added to the inverted bottleneck stage, allowing direct incorporation of the input spectrogram condition.

Generator Loss. Figure 5 compares the generator losses of the original Vocos and Vocos2D. We make two key changes: (1) removing the multi-period discriminator, as it degraded performance on input linear spectrograms with more frequency bins; and (2) adding discriminator augmentation (DA) (Zhao et al., 2020; Karras et al., 2020) before feeding real and fake waveforms into the multi-resolution discriminator (Jang et al., 2021). DA improves training stability by applying differentiable transformations only to the discriminator, allowing gradients to flow back to the generator without forcing it to model augmentations. We use three transformations: random loudness adjustments within ± 6 dB, random sample shifts, and random phase rotations.

Discriminator Training. Apart from the removal of the multi-period discriminator, Vocos2D’s discriminator training mirrors the original Vocos, using hinge loss (Zhang et al., 2019; Pan et al., 2023) and the multi-resolution discriminator (Jang et al., 2021).

Dataset. To support 32 kHz high-quality wideband speech synthesis, we augment our proprietary speech dataset with high-quality open-source singing voice data, expanding pitch range coverage and enhancing overall sound quality and adaptability to varied vocalization techniques.

3 EXPERIMENTS AND RESULTS

3.1 SPEECH TOKENIZER EVALUATION

We evaluate GLM-TTS-tokenizer across two English test sets and five Chinese dialect and accented Mandarin test sets, comparing with GLM4-Voice-tokenizer. As shown in Table 1, the GLM-TTS-

tokenizer significantly outperforms the GLM-4-Voice tokenizer in Chinese recognition, and shows minor improvements in English recognition.

Further, we implement TTS systems on both tokenizers and evaluate TTS metrics. The results in Table 2 show that GLM-TTS outperforms the variant built on GLM4-Voice-tokenizer in terms of Speaker Similarity(SIM) and CER.

Table 1: Performance of TTS systems built on GLM4-Voice-tokenizer and GLM-TTS-tokenizer on ASR tasks. Values represent WER/CER, the lower, the better.

Tokenizer	Libri Other	Libri Clean	Sichuan dialect	Jiao-Liao Mandarin	Taiwan Mandarin	Cantonese	Shanghai dialect
GLM4-Voice-tokenizer	4.90	2.10	54.11	14.04	49.09	46.81	72.06
GLM-TTS-tokenizer	4.51	2.12	24.40	9.11	16.92	7.27	19.15

Table 2: Performance of GLM4-Voice-tokenizer and GLM-TTS-tokenizer on seed_test_zh.

Tokenizer	SIM↑	CER↓
GLM4-Voice-tokenizer	75.2	1.44
GLM-TTS-tokenizer	76.1	1.03

3.2 VOICE CLONING RESULTS ON SEED-TTS-EVAL

Table 3 reports results on the Seed-TTS-eval benchmark using standard TTS metrics: Character Error Rate (CER) and Word Error Rate (WER) for pronunciation accuracy (lower is better), and Speaker Similarity (SIM, higher is better) measured by calculating the cosine similarity between speaker embeddings extracted using fine-tuned WavLM-large(Chen et al., 2022). We compare GLM-TTS (1.5B parameters, open-source) with 19 state-of-the-art baselines, including both closed-source systems (e.g., MiniMax-Speech, Seed-TTS, et al.) and open-source models (e.g., IndexTTS2, FireRedTTS-2, VibeVoice, et al.).

On the test-zh set, closed-source models deliver leading performance: MiniMax-Speech achieves a state-of-the-art CER of 0.83%, while Seed-TTS attains a standout SIM of 79.6. GLM-TTS attains a CER of 1.03% and SIM of 76.1, delivering results that are generally comparable to the open-source SOTA TTS model, such as VoxCPM and IndexTTS2. After applying our multi-reward GRPO reinforcement learning, GLM-TTS_RL further improves to CER=0.89% and SIM=76.4, substantially narrowing the gap with the leading closed-source systems.

Notably, due to practical industrial constraints, GLM-TTS is trained with a limited amount of English data (less than half of the Chinese training data). Despite this limitation, on the test-en set, GLM-TTS still achieves reasonable WER (2.23%) and SIM (67.2) performance on the English test set, suggesting potential for further improvement with additional English data.

Overall, GLM-TTS achieves top-tier performance among 1.5B-scale open-source models on the Seed-TTS-eval test set, with a negligible performance gap relative to state-of-the-art closed-source counterparts. These results demonstrate the effectiveness of the proposed overall framework, and further validate the effectiveness of our core design choices—multi-reward GRPO reinforcement learning for consistently improving pronunciation accuracy and timbre fidelity.

3.3 SPEECH-LM RL-ALIGNMENT

We adopt two techniques from DAPO (Yu et al., 2025): **Clip-Higher** and **Dynamic Sampling**. First, we set ϵ_{high} to 0.3 and ϵ_{low} to 0.2. For Dynamic Sampling, to facilitate training, we resample batches with zero advantages up to three times. As shown in Table 4, Clip-Higher yields fully positive gains. Dynamic Sampling improves both SIM and CER, yet leads to a decline in EMO. This is because repeated sampling can easily induce variance in SIM or CER across different samples, while the distribution of EMO is more extreme, approximating a bimodal distribution concentrated near 0 and 1. Consequently, we no longer consider the variance introduced by SIM in the resampling process.

Table 3: Performance comparison on different test sets. **test-zh**: Chinese standard test set; **test-en**: English test set; **test-hard**: Hard case test set (polyphones, rare words). CER/WER are lower-is-better (\downarrow), SIM is higher-is-better (\uparrow).

Model	Params	Open source	<i>test-zh</i>		<i>test-en</i>	
			CER \downarrow	SIM \uparrow	WER \downarrow	SIM \uparrow
MegaTTS3 (Jiang et al., 2025)	0.5B	No	1.52	79.0	2.79	77.1
DiTAR (Jia et al., 2025)	0.6B	No	1.02	75.3	1.69	73.5
CosyVoice3 (Du et al., 2025)	1.5B	No	1.12	78.1	2.22	72.0
Seed-TTS (Anastassiou et al., 2024)	-	No	1.12	79.6	2.25	76.2
MiniMax-Speech (Zhang et al., 2025)	-	No	0.83	78.3	1.65	69.2
F5-TTS (Chen et al., 2025)	0.3B	Yes	1.53	76.0	2.00	67.0
MaskGCT (Wang et al., 2024)	-	Yes	2.27	77.4	2.62	71.7
CosyVoice (Du et al., 2024a)	0.3B	Yes	3.63	72.3	4.29	60.9
CosyVoice2 (Du et al., 2024b)	0.5B	Yes	1.38	75.7	3.09	65.9
CosyVoice3 (Du et al., 2025)	0.5B	Yes	1.16	78.0	2.02	71.8
SparkTTS (Wang et al., 2025a)	0.5B	Yes	1.54	66.0	3.14	57.3
FireRedTTS (Guo et al., 2025)	0.5B	Yes	1.51	63.5	3.82	46.0
FireRedTTS-2 (Xie et al., 2025)	-	Yes	1.14	73.6	1.95	66.5
Qwen2.5-Omni (Xu et al., 2025)	7B	Yes	1.70	75.2	2.72	63.2
OpenAudio-s1-mini (OpenAudio, 2024)	0.5B	Yes	1.18	68.5	1.94	55.0
IndexTTS 2 (Zhou et al., 2025a)	1.5B	Yes	1.03	76.5	2.23	70.6
VibeVoice (Peng et al., 2025)	1.5B	Yes	1.16	74.4	3.04	68.9
HiggsAudio-v2 (BosonAI, 2025)	3B	Yes	1.50	74.0	2.44	67.7
VoxCPM (Zhou et al., 2025b)	0.5B	Yes	0.93	77.2	1.85	72.9
GLM-TTS (Ours)	1.5B	Yes	1.03	76.1	2.23	67.2
GLM-TTS_RL (Ours)	1.5B	Yes	0.89	76.4	1.91	68.1

Table 4: Performance of Pretrain-GRPO with Clip-Higher and Dynamic Sampling on an internal emotion test set. c represents Clip-Higher and d represents Dynamic Sampling.

Model	CER \downarrow	SIM \uparrow	EMO \uparrow
Pretrain-base	2.05	80.0	0.525
Pretrain-GRPO	1.99	80.3	0.565
Pretrain-GRPO_c	1.93	80.4	0.660
Pretrain-GRPO_d	1.91	80.8	0.440

Since Clip-Higher has demonstrated improvement, we further explored dynamically adjusting parameters such as ϵ_h during training to provide the model with sufficient exploration space while maintaining training stability. Specifically, we selected three parameters, ϵ_h , ϵ_l , and T with initial values of 0.3, 0.2, and 1, respectively, and allowed them to linearly increase as training progressed until training completion. As shown in Table 5, we observed that more aggressive parameter settings tend to lead to less stable training: while emotional expressiveness is enhanced, pronunciation becomes less clear. Moreover, excessively loose constraints are also prone to resulting in reward hacking.

To further improve laughter modeling, we introduced a laughter reward, which can be summarized as follows: if the text contains two or more consecutive laughter words (such as “ha” and “hey”), and the laughter detection model identifies a laughter segment, then (1) if the ASR system transcribes the segment as a “deletion”(empty string), the reward is set to 1; (2) if the ASR system transcribes the corresponding text, the reward is set to 0.

We experimented with different weights λ_{laugh} for the laughter reward and the results are shown in Table 6. Increasing the laughter reward leads to decreases in CER and similarity scores, as laughter segments cannot be recognized by the ASR model as textual content, and the timbre of laughter

Table 5: Performance of SFT-GRPO with dynamic ϵ_h , ϵ_l , and T on an internal emotion test set. * means freezing parameters during training process.

Model	params			metrics		
	T	ϵ_h	ϵ_l	CER \downarrow	SIM \uparrow	EMO \uparrow
SFT-base	-	-	-	2.13	76.1	0.695
SFT-GRPO*	-	-	-	2.21	76.3	0.720
SFT-GRPO	1.5	0.5	0.4	2.09	76.7	0.705
SFT-GRPO	2	1	0.4	2.16	75.5	0.790
SFT-GRPO	3	1	0.4	2.21	78.1	0.885

differs from the speaker’s regular speaking voice. Nevertheless, enhancing laughter synthesis can also improve the model’s emotional expressiveness.

Table 6: Performance of SFT-GRPO with different $\lambda_{laughter}$ on an internal emotion test set.

Model	params		metrics		
	$\lambda_{laughter}$	CER \downarrow	SIM \uparrow	EMO \uparrow	
SFT-base	-	3.11	76.3	0.44	
SFT-GRPO	2	2.86	74.6	0.64	
SFT-GRPO	5	3.18	74.8	0.66	
SFT-GRPO	10	3.06	74.8	0.72	

3.4 EFFECTIVENESS OF PHONEME-IN

To rigorously evaluate the impact of the **Phoneme-in** mechanism on pronunciation accuracy, we conducted a targeted ablation study using a proprietary internal dataset. Unlike standard public benchmarks, this dataset is specifically constructed to simulate challenging industrial scenarios, characterized by a high density of polyphonic characters, low-frequency words, and ambiguous contexts that typically confuse end-to-end TTS systems.

We compared the performance of GLM-TTS with and without the Phoneme-in module enabled. As illustrated in Table 7, the baseline model, which relies solely on text input and implicit grapheme-to-phoneme prediction, yields a Phoneme Error Rate (PER) of 13.23%. This relatively high error rate indicates the inherent difficulty of the test set. However, when the Phoneme-in mechanism is activated—allowing for fine-grained, phoneme-level intervention—the PER dramatically drops to 5.14%.

Table 7: Ablation study of the Phoneme-in mechanism on the internal hard-case dataset. The use of hybrid phoneme input significantly reduces pronunciation errors.

Model Settings	Input Modality	PER (\downarrow)
GLM-TTS (w/o Phoneme-in)	Text Only	13.23
GLM-TTS (w/ Phoneme-in)	Hybrid (Text + Phoneme)	5.14

3.5 VOCOS2D VOCODER

Table 8: Performance comparison between Vocos and Vocos2D.

	NISQA \uparrow	UTMOS \uparrow	Ab. Aes.-PQ \uparrow	MOS \uparrow
GT	3.47	2.11	7.68	4.77
Vocos	3.16	1.87	7.56	3.58
Vocos2D	3.40	1.91	7.64	4.16

We evaluate Vocos2D on an internal test set of randomly selected audio samples. Objective metrics include UTMOS (Saeki et al., 2022), NISQA (Mittag et al., 2021), and the production quality (PQ) score from Meta AudioBox Aesthetics (Tjandra et al., 2025), complemented by subjective MOS evaluations. As shown in Table 8, Vocos2D consistently outperforms the original Vocos across all metrics, demonstrating the effectiveness of the proposed architectural and training improvements.

4 CONCLUSION

In this technical report, we introduce **GLM-TTS**, a production-level text-to-speech system that systematically addresses several long-standing challenges in modern TTS. With only 100k hours of training data, GLM-TTS achieves competitive pronunciation accuracy and speaker similarity. The proposed multi-reward GRPO-based reinforcement learning framework effectively aligns synthesized speech with human perceptual preferences, leading to consistent improvements in pronunciation accuracy, emotional expressiveness, and speaker similarity without sacrificing training stability.

Beyond core model performance, GLM-TTS emphasizes practical deployability. The optimized LoRA-based premium voice customization strategy significantly reduces both data and computational costs, enabling scalable personalization in production environments. The hybrid phoneme-text input mechanism provides precise and controllable pronunciation for polyphonic and rare words, addressing a critical requirement in professional and educational TTS applications, especially for Chinese. Overall, GLM-TTS provides a practical framework for efficient and controllable speech synthesis. We hope it can serve as a foundation for future research on expressive, customizable, and scalable speech generation.

REFERENCES

- Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, Jiaqi Shi, Xian Shi, Hao Wang, Wen Wang, Yuxuan Wang, Zhangyu Xiao, Zhiping Yan, Yexin Yang, Bin Zhang, Qinglin Zhang, Shiliang Zhang, Nan Zhao, and Siqi Zheng. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms, 2024. URL <https://arxiv.org/abs/2407.04051>.
- Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuan-hao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-TTS: A family of high-quality versatile speech generation models, 2024. URL <https://arxiv.org/abs/2406.02430>.
- Nishchal Bhandari, Danny Chen, Miguel Ángel del Río Fernández, Natalie Delworth, Jennifer Drexler Fox, Miguel Jetté, Quinten McNamara, Corey Miller, Ondřej Novotný, Ján Profant, Nan Qin, Martin Ratajczak, and Jean-Philippe Robichaud. Reverb: Open-source asr and diarization from rev, 2025. URL <https://arxiv.org/abs/2410.03930>.
- BosonAI. Higgs Audio v2: Redefining expressiveness in audio generation, 2025. URL <https://github.com/boson-ai/higgs-audio>.
- Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. pyannote.audio: neural building blocks for speaker diarization, 2019. URL <https://arxiv.org/abs/1911.01255>.
- 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, 2025. URL <https://arxiv.org/abs/2410.06885>.

-
- Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. Large-scale self-supervised speech representation learning for automatic speaker verification, 2022. URL <https://arxiv.org/abs/2110.05777>.
- Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens, 2024a. URL <https://arxiv.org/abs/2407.05407>.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. CosyVoice 2: Scalable streaming speech synthesis with large language models, 2024b. URL <https://arxiv.org/abs/2412.10117>.
- Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training, 2025. URL <https://arxiv.org/abs/2505.17589>.
- Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022. URL <https://arxiv.org/abs/2210.13438>.
- Yuan Gao, Nobuyuki Morioka, Yu Zhang, and Nanxin Chen. E3 tts: Easy end-to-end diffusion-based text to speech, 2023a. URL <https://arxiv.org/abs/2311.00945>.
- Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition, 2023b. URL <https://arxiv.org/abs/2206.08317>.
- Google. gRPC: A high-performance, open-source universal rpc framework. Official Website, 2015. URL <https://grpc.io/>.
- Hao-Han Guo, Yao Hu, Kun Liu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications, 2025. URL <https://arxiv.org/abs/2409.03283>.
- Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation, 2021. URL <https://arxiv.org/abs/2106.07889>.
- Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, and Yuxuan Wang. Ditar: Diffusion transformer autoregressive modeling for speech generation, 2025. URL <https://arxiv.org/abs/2502.03930>.
- Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jiong-hao, Xiaoda Yang, Jialong Zuo, Yu Zhang, Rui Liu, Xiang Yin, and Zhou Zhao. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis, 2025. URL <https://arxiv.org/abs/2502.18924>.
- Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data, 2020. URL <https://arxiv.org/abs/2006.06676>.
- Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. Ctc-segmentation of large corpora for german end-to-end speech recognition, 2020. ISSN 1611-3349. URL http://dx.doi.org/10.1007/978-3-030-60276-5_27.
- Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and Ming Zhou. Neural speech synthesis with transformer network, 2019. URL <https://arxiv.org/abs/1809.08895>.

-
- Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, interspeech_2021. ISCA, August 2021. doi: 10.21437/interspeech.2021-299. URL <http://dx.doi.org/10.21437/Interspeech.2021-299>.
- OpenAudio. OpenAudio s1: A cutting-edge text-to-speech model that performs like voice actors. Official Blog, 2024. URL <https://openaudio.com/blogs/s1>.
- Yudong Pan, Ning Li, Yangsong Zhang, Peng Xu, and Dezhong Yao. Short-length ssvep data extension by a novel generative adversarial networks based framework, 2023. URL <https://arxiv.org/abs/2301.05599>.
- Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, and Furu Wei. VibeVoice technical report, 2025. URL <https://arxiv.org/abs/2508.19205>.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL <https://arxiv.org/abs/2212.04356>.
- Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech, 2019. URL <https://arxiv.org/abs/1905.09263>.
- Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnouke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022, 2022. URL <https://arxiv.org/abs/2204.02152>.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
- Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis, 2024. URL <https://arxiv.org/abs/2306.00814>.
- 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, 2025. URL <https://arxiv.org/abs/2504.02407>.
- Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound, 2025. URL <https://arxiv.org/abs/2502.05139>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL <https://arxiv.org/abs/1706.03762>.
- Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023a.
- Ju-Chiang Wang, Wei-Tsung Lu, and Minz Won. Mel-band roformer for music source separation, 2023b. URL <https://arxiv.org/abs/2310.01809>.
- Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfu Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-TTS: An efficient lilm-based text-to-speech model with single-stream decoupled speech tokens, 2025a. URL <https://arxiv.org/abs/2503.01710>.

-
- Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfu Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens, 2025b. URL <https://arxiv.org/abs/2503.01710>.
- 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, 2024. URL <https://arxiv.org/abs/2409.00750>.
- Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis, 2017. URL <https://arxiv.org/abs/1703.10135>.
- Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. Fireredtts-2: Towards long conversational speech generation for podcast and chatbot, 2025. URL <https://arxiv.org/abs/2509.02020>.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report, 2025. URL <https://arxiv.org/abs/2503.20215>.
- Qiyng Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source llm reinforcement learning system at scale, 2025. URL <https://arxiv.org/abs/2503.14476>.
- Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Sound-Stream: An end-to-end neural audio codec, 2021. URL <https://arxiv.org/abs/2107.03312>.
- Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot, 2024. URL <https://arxiv.org/abs/2412.02612>.
- Bowen Zhang, Congchao Guo, Geng Yang, Hang Yu, Haozhe Zhang, Heidi Lei, Jialong Mai, Junjie Yan, Kaiyue Yang, Mingqi Yang, Peikai Huang, Ruiyang Jin, Sitan Jiang, Weihua Cheng, Yawei Li, Yichen Xiao, Yiying Zhou, Yongmao Zhang, Yuan Lu, and Yucen He. MiniMax-Speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder, 2025. URL <https://arxiv.org/abs/2505.07916>.
- Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks, 2019. URL <https://arxiv.org/abs/1805.08318>.
- Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training, 2020. URL <https://arxiv.org/abs/2006.10738>.
- 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, 2025a. URL <https://arxiv.org/abs/2506.21619>.
- Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, Zhiyong Wu, and Zhiyuan Liu. VoxCPM: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning, 2025b. URL <https://arxiv.org/abs/2509.24650>.