DMOSpeech 2: Reinforcement Learning for Duration Prediction in Metric-Optimized Speech Synthesis

Yinghao Aaron Li[1*], **Xilin Jiang**[1*], **Fei Tao**[2†], **Cheng Niu**[2], **Kaifeng Xu**[2], **Juntong Song**[2], **Nima Mesgarani**[1]

[1]Columbia University
[2]NewsBreak
{yl4579, xj2289}@columbia.edu, fei.tao@newsbreak.com, nima@ee.columbia.edu

Abstract

Diffusion-based text-to-speech (TTS) have made remarkable progress in zero-shot speech synthesis, yet optimizing all components for perceptual metrics remains challenging. Prior work with DMOSpeech demonstrated direct metric optimization for speech generation components, but duration prediction remained unoptimized. This paper presents DMOSpeech 2, which extends metric optimization to the duration predictor through a reinforcement learning approach. The proposed system implements a novel duration policy framework using group relative preference optimization (GRPO) with speaker similarity and word error rate as reward signals. By optimizing this previously unoptimized component, DMOSpeech 2 creates a more complete metric-optimized synthesis pipeline. Additionally, this paper introduces teacher-guided sampling, a hybrid approach leveraging a teacher model for initial denoising steps before transitioning to the student model, significantly improving output diversity while maintaining efficiency. Comprehensive evaluations demonstrate superior performance across all metrics compared to previous systems, while reducing sampling steps by half without quality degradation. These advances represent a significant step toward speech synthesis systems with metric optimization across multiple components.

Audio Demo — https://dmospeech2.github.io

1 Introduction

Text-to-speech (TTS) synthesis has progressed dramatically in recent years, with state-of-the-art systems producing speech virtually indistinguishable from human recordings (Tan et al. 2024; Li et al. 2024a; Ju et al. 2024). Among the most significant advancements is zero-shot TTS, which is the ability to synthesize speech in the voice of an unseen speaker, given only a short audio sample without speaker-specific training. This capability has transformative potential across applications ranging from personalized digital assistants to accessibility tools and creative content production.

Despite impressive quality improvements, zero-shot TTS still faces a fundamental challenge: the lack of true end-to-end optimization for perceptual quality metrics. Current approaches struggle to directly optimize key metrics such as speaker similarity and intelligibility in an end-to-end manner, limiting their performance ceiling, especially for smaller and more efficient models. Reinforcement learning (RL) offers a potential indirect optimization approach (Chen et al. 2024a; Zhang et al. 2024; Gao et al. 2025; Tian et al. 2025; Hussain et al. 2025) but comes with significant limitations. The ceiling of RL-based improvement is essentially best-of-N sampling (Ichihara et al. 2025), making its effectiveness heavily dependent on the original model's output diversity, therefore limiting its applicability for smaller, more efficient models with limited output diversity. Additionally, traditional RL for TTS imposes substantial computational overhead, as each training step requires generating complete speech samples, often through hundreds of sampling steps, making large-scale training prohibitively expensive.

As the field has evolved, researchers have pursued two fundamentally different approaches to generating speech. Autoregressive models (Wang et al. 2023a; Peng et al. 2024; Chen et al. 2024c; Wang et al. 2024; Du et al. 2024a, 2025, 2024b; Zhu, Tian, and Xie 2024; Wang et al. 2025; Song et al. 2024; Ye et al. 2025) generate speech step-by-step, similar to how large language models produce text, naturally determining the duration of speech during generation but struggling with direct optimization due to the computational expense of backpropagating through their long generation sequences. Meanwhile, diffusion-based systems (Le et al. 2024; Shen et al. 2023; Li et al. 2024b; Eskimez et al. 2024; Yang et al. 2024; Lee et al. 2024; Chen et al. 2024d) treat speech synthesis as an inpainting task that requires knowing the total speech duration in advance, creating a division in the pipeline: first predicting how long the speech should be, then generating the actual audio content. Without a differentiable pathway in between, traditional optimization techniques cannot flow through the entire system. Research has demonstrated that input durations significantly impact key metrics like speaker similarity (SIM) and word error rate (WER) (Eskimez et al. 2024), yet existing systems either train duration predictors separately from speech generation (Le et al. 2024; Lee et al. 2024) or use heuristic approaches based on prompt speaking rates (Chen et al. 2024d; Eskimez et al. 2024).

The original **D**irect **M**etric **O**ptimization **Speech** framework (Li, Kumar, and Jin 2024) made significant progress by enabling direct metric optimization for the speech generation component through diffusion model distillation. By reducing

* Equal contribution.

† Project lead.

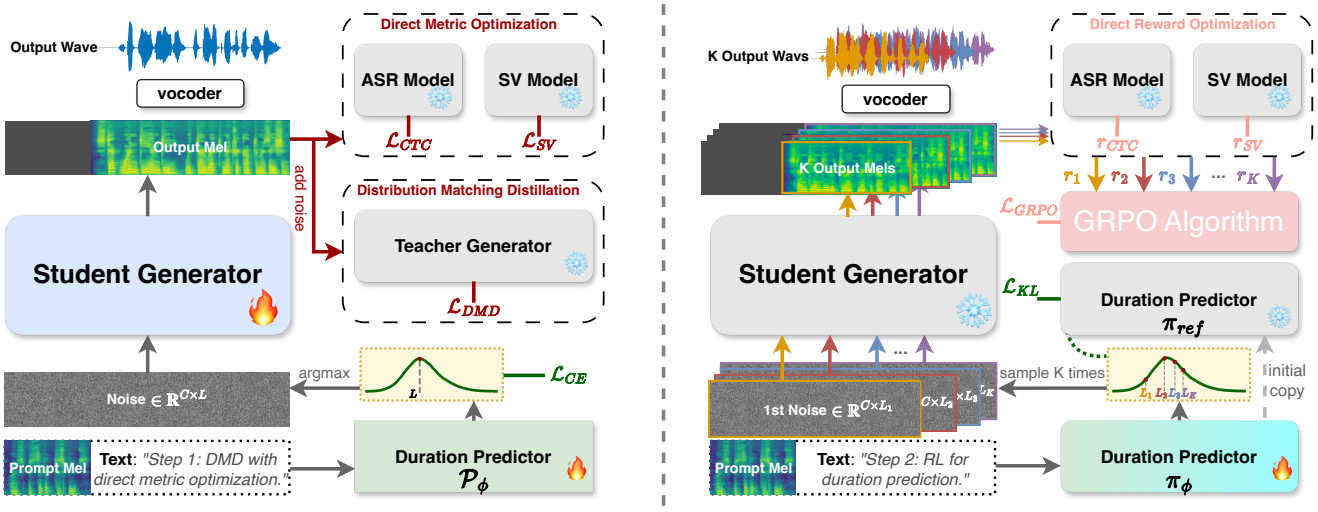

Figure 1: Overview of the DMOSpeech 2 framework. **(a) Left:** The original DMOSpeech architecture, where the duration predictor ($\mathcal{P}_\phi$) is trained self-supervisedly and separate from the TTS component, creating a disconnection that prevents end-to-end optimization. **(b) Right:** Our proposed DMOSpeech 2 framework, which employs Group Relative Policy Optimization (GRPO) to train the duration predictor with reinforcement learning (Algorithm 1), using speaker similarity and word error rate as reward signals, enabling optimization of the entire TTS pipeline.

sampling steps from 128 to 4 and establishing direct gradient pathways within the generation process, DMOSpeech enabled direct optimization for speaker similarity and intelligibility. However, a critical limitation remained: the duration predictor component was still outside the optimization loop, creating a bottleneck in overall system quality.

This paper introduces **DMOSpeech 2**, which addresses the duration prediction challenge through reinforcement learning. We propose modeling the duration predictor as a probabilistic policy and applying reinforcement learning with group relative policy optimization (GRPO), using speaker similarity and word error rate as reward signals. Importantly, by applying RL specifically to the duration predictor and operating on samples generated by our efficient 4-step student model, we dramatically reduce the computational overhead typically associated with RL for TTS. This targeted approach also mitigates the limitations of whole-system RL, as optimizing duration prediction is a much more constrained problem than optimizing speech generation directly. Additionally, to address the output diversity reduction observed in the original DMOSpeech as a consequence of distribution matching distillation (Yin et al. 2024b), we introduce teacher-guided sampling, a hybrid approach that leverages the teacher model for initial denoising steps before transitioning to the student model. This strategy restores diversity to near-teacher levels while still achieving a 2× reduction in sampling steps and maintaining the significant quality improvements enabled by our direct metric optimization approach. Using the flow-matching-based F5-TTS (Chen et al. 2024d) as our teacher model, our comprehensive evaluations demonstrate that DMOSpeech 2 significantly outperforms both the previous system and other recent baselines across all metrics. The reinforcement learning approach to duration prediction results in particularly notable improvements in speaker similarity and word error rate, precisely targeting the limitations identified in previous systems.

The contributions of this work are twofold: 1) we propose a computationally efficient reinforcement learning framework specifically for duration prediction in non-parallel TTS systems, enabling alignment with perceptual metrics without the overhead typically associated with RL approaches, and 2) we propose a teacher-guided sampling for diffusion model distillation, restoring output diversity while maintaining computational efficiency. We will also make the source code and pre-trained models publicly available for future research.

2 Related Works

Zero-Shot Text-to-Speech Synthesis Zero-shot TTS has evolved significantly over recent years, with approaches broadly categorized into two main paradigms. Early methods relied on speaker embeddings from pre-trained encoders (Casanova et al. 2022, 2021; Wu et al. 2022; Lee et al. 2022) or end-to-end speaker encoders (Li et al. 2024a; Min et al. 2021; Li, Han, and Mesgarani 2022; Choi et al. 2022), but struggled with generalization due to their dependence on extensive feature engineering and with direct metric optimization due to their non-differentiable components such as duration predictors. Recent advancements have primarily focused on prompt-based approaches, which can be divided into autoregressive and diffusion-based methods. Autoregressive models (Wang et al. 2023a; Peng et al. 2024; Chen et al. 2024c; Wang et al. 2024; Du et al. 2024a, 2025, 2024b; Zhu, Tian, and Xie 2024; Wang et al. 2025; Song et al. 2024; Ye et al. 2025) generate speech sequentially and naturally determine duration during generation, while diffusion-based approaches (Le et al. 2024; Shen et al. 2023; Li et al. 2024b; Eskimez et al. 2024; Yang et al. 2024; Lee et al. 2024; Chen et al. 2024d) require predetermined speech

duration. Although DMOSpeech (Li, Kumar, and Jin 2024) made progress by enabling direct optimization for the speech generation component, it still left the duration predictor outside the optimization loop. In DMOSpeech 2, we optimize the previously unoptimized duration predictor with reinforcement learning for perceptually relevant metrics.

Reinforcement Learning in Speech Synthesis Reinforcement learning (RL) has emerged as a promising approach for aligning speech synthesis systems with human perceptions, though its application to TTS presents unique challenges. Recent work has explored various RL techniques for improving TTS quality. SpeechAlign (Zhang et al. 2024) introduced an iterative self-improvement strategy for neural codec language models that constructs preference datasets and optimizes toward human preferences. Similarly, UNO (Chen et al. 2024a) proposed an uncertainty-aware optimization framework that integrates subjective human evaluation directly into the TTS training loop without requiring a separate reward model. Several approaches have focused on specific aspects of speech quality: Gao et al. (2025) developed Emo-DPO for controllable emotional speech synthesis, differentiating subtle emotional nuances through preference optimization, while Tian et al. (2025) demonstrated that direct preference optimization (DPO) consistently improves intelligibility and speaker similarity in LM-based TTS. Koel-TTS (Hussain et al. 2025) enhanced encoder-decoder TTS models through preference alignment guided by automatic speech recognition and speaker verification. For diffusion-based TTS specifically, Chen et al. (2024b) introduced diffusion model loss-guided RL policy optimization (DLPO) to improve naturalness and quality, and Sun et al. (2025) employed group relative policy optimization for flow-matching-based TTS models. However, most existing approaches apply RL to the entire TTS pipeline, incurring substantial computational overhead and facing effectiveness limitations dependent on the original model's output diversity. DMOSpeech 2 addresses these limitations by specifically targeting RL to the duration predictor component, dramatically reducing computational overhead by operating on samples generated through an efficient 4-step student model, while simultaneously addressing the critical optimization gap in current non-parallel zero-shot TTS.

3 Methods

3.1 DMOSpeech with Flow Matching

DMOSpeech (Li, Kumar, and Jin 2024) is a framework for efficient zero-shot TTS that combines distribution matching distillation (Yin et al. 2024b) with direct metric optimization. DMOSpeech 2 builds upon the original DMOSpeech framework while adopting F5-TTS (Chen et al. 2024d) as the teacher model. This section summarizes the key components of our approach, highlighting the adaptations made for flow matching-based models. Fig. 1a illustrates the DMOSpeech architecture with details in Appendix C.

Unlike the original DMOSpeech which operated on latent representations from an audio autoencoder, DMOSpeech 2 directly generates mel-spectrograms, with waveforms synthesized using the pre-trained Vocos (Siuzdak 2023) vocoder. The framework consists of three training components. First, a student generator G_θ is trained through improved distribution matching distillation (DMD 2) (Yin et al. 2024a) to match a pre-trained teacher model in distribution. This allows the student to generate high-quality speech with significantly fewer sampling steps (4 steps). Second, multi-modal adversarial training with a discriminator improves the perceptual quality of the generated speech. Finally, the direct metric optimization component enables end-to-end optimization of word error rate and speaker similarity metrics with pre-trained automatic speech recognition (ASR) models and speaker verification (SV) models on mel-spectrograms.

During inference, DMOSpeech generates speech directly from noise in four denoising steps, conditioned on the input text and speaker prompt and the total duration of the target speech. The process begins with sampling Gaussian noise $\mathbf{z} \sim \mathcal{N}(0, I)$ at a predefined duration L, which is determined by a separate duration predictor. The student generator G_θ then transforms this noise into mel-spectrograms through four sequential steps using the sway sampling schedule (Chen et al. 2024d) with coefficient $u = -1$ at noise levels $t \in \{0.0000, 0.0761, 0.2929, 0.6173\}$ rather than uniform steps. The final spectrograms are converted to waveforms using the vocoder. While DMOSpeech enabled direct metric optimization for the generator, it still has a critical limitation: the duration predictor remained outside the optimization loop. DMOSpeech 2 addresses this limitation through reinforcement learning, as detailed in the following sections.

3.2 Speech Length Predictor with RL

As established in the previous section, while DMOSpeech enables direct optimization of the speech generator, a critical limitation remains: the duration predictor sits outside the optimization loop, creating a disconnection that prevents end-to-end optimization. This separation is particularly problematic because speech duration significantly impacts perceptual metrics like speaker similarity (SIM) and word error rate (WER) (Eskimez et al. 2024). To address this limitation, DMOSpeech 2 introduces a novel reinforcement learning approach specifically targeting the speech length predictor.

Duration Predictor Architecture We adopt an encoder-decoder transformer architecture similar to DiTTo-TTS (Lee et al. 2024) for our speech length predictor. Unlike conventional duration models that predict phoneme-level durations, our model is specifically designed to predict the total remaining length of speech to be generated. Formally, let $\mathbf{x}$ represent the input text sequence and $\mathbf{p}_t$ represent the speech prompt up to frame t. Our speech length predictor $\mathcal{P}_\phi$ with parameters ϕ models the probability distribution for L_t, the number of remaining frames needed to complete the utterance:

$$P(L_t = l|\mathbf{x}, \mathbf{p}_t) = [\mathcal{P}_\phi(\mathbf{x}, \mathbf{p}_t)]_l \tag{1}$$

where $[\cdot]_l$ denotes the model prediction for l frames remained. This formulation creates an autoregressive structure where the predicted remaining length decreases as the speech prompt extends. The architecture consists of a bidirectional text encoder that processes the input text to capture comprehensive contextual information. The decoder, equipped with causal masking to prevent future lookahead, takes the mel-spectrogram of the speech prompt as input. Cross-attention

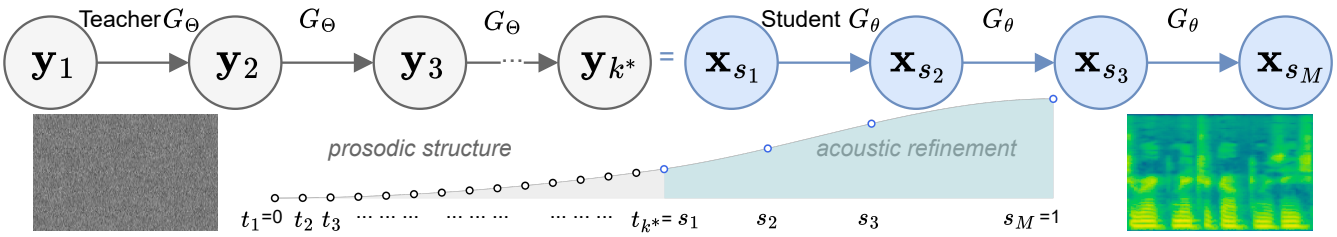

Figure 2: Illustration of teacher-guided sampling (Algorithm 2). The process begins with noise and uses the teacher model G_Θ for early denoising steps (gray circles) to establish prosodic structure up to a transition point t_{k^*}. Then, the student model G_θ (blue circles) takes over for the remaining steps to refine acoustic details in much fewer steps.

mechanisms integrate text features from the encoder, and the final layer applies softmax activation to predict a distribution over possible remaining lengths within a predefined maximum length. Our implementation uses a transformer with 4 encoder layers for text processing and 4 decoder layers with cross-attention mechanisms. The model employs 8 attention heads in each layer with a hidden dimension of 512. We set the maximum total duration to be 30 seconds binned into 300 possible duration classes, with increments of 100 ms. During training, the ground truth label for the remaining audio length decreases by one at each subsequent time step. For a batch of sequences with mel-spectrogram lengths $\{L_1, L_2, ..., L_B\}$, where B is the batch size, the target remaining length is a decreasing sequence $(L_i - 1, L_i - 2, ..., 1, 0)$ for each training example L_i. The predictor is initially trained separately from the flow-matching model using cross-entropy loss between the predicted distribution and the ground truth remaining lengths. In DMOSpeech 2, we extend this training process with reinforcement learning to directly optimize for perceptual quality metrics.

GRPO-based Duration Optimization To enable direct optimization for perceptual metrics, we formulate the speech length predictor as a stochastic policy in a reinforcement learning framework and apply group relative policy optimization (GRPO) (Shao et al. 2024), which allows us to optimize the length predictor directly for perceptual metrics without need of a differentiable pathway to the generator. The detailed algorithm is provided in Algorithm 1 in Appendix A.

For each input text $\mathbf{x}$ and prompt $\mathbf{p}$, we model the policy for choosing the total speech length (L) with our probabilistic duration predictitor, $\pi_\phi(L|\mathbf{x}, \mathbf{p}) := \mathcal{P}_\phi(\mathbf{x}, \mathbf{p})$. During training, we sample K different durations for each input, where K is the group size:

$$L_k \sim \pi_\phi(L|\mathbf{x}, \mathbf{p}), \quad k = 1, 2, ..., K, \tag{2}$$

For each sampled duration, we generate speech using our efficient 4-step student model:

$$\mathbf{y}_k = G_\theta(\mathbf{z}, \mathbf{x}, \mathbf{p}, L_k), \quad z \sim \mathcal{N}(0, I), \tag{3}$$

where G_θ is our student generator and $\mathbf{z}$ is the noise. We then compute rewards for each sample using a combination of speaker similarity and speech recognition metrics:

$$r_k = \log P_{CTC}(\mathbf{x}|C(\mathbf{y}_k)) + \lambda_{\text{SIM}} \cdot \frac{\mathbf{e}_\text{p} \cdot \mathbf{e}_{y_k}}{\|\mathbf{e}_\text{p}\| \, \|\mathbf{e}_{y_k}\|}, \tag{4}$$

where $C(\cdot)$ is a pre-trained CTC-based ASR model operating on mel-spectrograms, $\mathbf{e}_\text{p} = S(\mathbf{p})$ and $\mathbf{e}_{y_k} = S(\mathbf{y}_k)$ are the speaker embeddings of the prompt and student-generated speech, and λ_{SIM} is the weighting factor. We chose $\lambda_{\text{SIM}} = 3$ to balance the contributions from the embedding similarity and word error rate (see Appendix B.2 for details).

We normalize the reward to compute the advantage:

$$A_k = \frac{r_k - \mu_r}{\sigma_r}, \tag{5}$$

where μ_r and σ_r are the mean and standard deviation of rewards within the group. In GRPO, we maintain three distinct policies. The current policy π_ϕ is the speech length predictor being actively trained. The old policy π_{old} is the version from which the current batch of samples was generated. In practice, this is typically the policy from several optimization steps ago. The reference policy π_{ref} is a frozen copy of the initially supervised model created at the beginning of RL training and kept constant throughout the process to serve as an anchor for regularization. We define the ratio R_k as :

$$R_k = \frac{\pi_\phi(L_k|\mathbf{x}, \mathbf{p})}{\pi_{\text{old}}(L_k|\mathbf{x}, \mathbf{p})} \tag{6}$$

The GRPO loss for a single sample is:

$$\mathcal{L}_k = \min\left(A_k \cdot R_k, A_k \cdot \text{clip}\left(R_k, 1 \pm \varepsilon\right)\right) - \beta \cdot \text{KL} \tag{7}$$

where $\varepsilon = 0.2$ is the clipping parameter for the policy update magnitude, $\beta = 0.04$ controls the strength of KL regularization, and $\text{KL} = \mathbb{D}_{\text{KL}}[\pi_\phi || \pi_{\text{ref}}]$ is the KL divergence between the current policy and the reference policy, preventing the trained policy from deviating too far from the initial model.

The full GRPO loss is thus defined as:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{\mathbf{x},\mathbf{p}}\left[\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_k\right] \tag{8}$$

We used temperature-based sampling to encourage exploration of diverse length predictions and implemented quality control that skips batches with insufficient reward diversity $(\max(r) - \min(r) < 0.01)$, ensuring that the model only learns from batches where meaningful distinctions between good and bad duration predictions can be made.

3.3 Teacher-Guided Sampling

Mode Shrinkage in Distribution Matching Distillation One key limitation of distribution matching distillation observed in the original DMOSpeech is a phenomenon known

as *mode shrinkage*. When student models are trained to generate speech in significantly fewer steps than their teacher, they tend to focus on high-probability regions of the data distribution, reducing diversity of the generated samples. While the student model exhibits similar mode coverage in sound quality compared to the teacher as indicated by UTMOS (Saeki et al. 2022), it demonstrates less diversity in prosodic features such as intonation patterns, rhythm variations, and speech cadences (Figure 3). This suggests that diversity reduction primarily occurs in the temporal and structural dimensions of speech rather than in its spectral characteristics.

The root cause of this diversity reduction can be traced to the diffusion process dynamics. In diffusion-based speech synthesis, different noise levels correspond to distinct aspects of the speech generation process. At high noise levels (early denoising steps), the model primarily establishes prosodic elements, phoneme durations, pauses, pitch contours, and text-speech alignments, essentially the semantic and structural framework of the utterance. In contrast, at low noise levels (later denoising steps), the model refines acoustic details such as voice quality, speaker identity, and spectral characteristics. When the student model is constrained to generate speech in just a few steps, it necessarily compresses this hierarchical generation process. Our empirical observations suggest that this compression disproportionately affects the diversity of prosodic and structural elements in the early denoising phase.

Hybrid Sampling Strategy To address the mode shrinkage problem, we introduce *teacher-guided sampling*, a hybrid approach that leverages the teacher model's diversity while preserving the student model's efficiency and improved speaker similarity from direct metric optimization. The core insight of our approach is to exploit the natural division in the diffusion process: use the teacher model for early denoising steps on prosodic structure and the student model for acoustic refinement of later steps. Specifically, we employ the teacher model to perform the initial denoising steps up to a predefined noise level t_{switch}, which establihes diverse prosodic patterns and text-speech duration alignments. Then, we switch to the student model, which completes the remaining denoising process from t_{switch} to 1 in just a few efficient steps. This hybrid approach preserves the diversity benefits of the teacher model while still achieving significant computational savings.

Figure 2 and Algorithm 2 in Appendix A outline our teacher-guided sampling procedure. The process begins with random Gaussian noise $\mathbf{z}$ and progressively denoises it through a sequence of steps. The first K steps are performed by the teacher model using a flow matching formulation with the sway sampling schedule (Chen et al. 2024d), which allocates more samples to early time steps where most of the semantic structure is established. Once the noise level reaches t_{switch}, the algorithm transitions to the student model, which completes the remaining denoising in just M steps (typically 2-3). A key advantage of our approach is that it achieves a more favorable trade-off between computational efficiency and output diversity. By delegating the labor-intensive task of establishing prosodic structure to the teacher model and the refinement of acoustic details to the student model, we leverage the strengths of both approaches. The teacher model is employed for fewer steps than its typical full inference (approximately 6-14 steps instead of 32), while the student model still performs only a small number of denoising steps (2-3 instead of 4). Our empirical evaluation (Table 1) confirms that teacher-guided sampling successfully mitigates the mode shrinkage problem, restoring the diversity of the generated speech to levels comparable to the teacher model, particularly in terms of pitch variation and cadence diversity. Notably, this improvement comes with only a modest increase in computational cost compared to the pure student model but still $1.8\times$ faster than the full teacher model. Additionally, similar to the student model, our hybrid approach produces samples with better SIM and WER than the teacher-only samples, benefiting from the direct metric optimization of the DMOSpeech framework. The parameters K, t_{switch}, and M offer flexible control over the trade-off between computational efficiency and output diversity. For applications where diversity is critical, such as creative content production, a higher t_{switch} value (around 0.4-0.5) can be used, allocating more steps to the teacher model. Conversely, for applications where efficiency is paramount, such as real-time systems, a lower t_{switch} value (around 0.1-0.2) can be employed with minimal degradation in perceptual quality.

4 Experiments

4.1 Experimental Setup

Datasets Following F5-TTS (Chen et al. 2024d), we utilize the in-the-wild multilingual speech dataset Emilia (He et al. 2024) to train our models. After filtering out transcription failures and misclassified language speech, we retain approximately 95k hours of English and Chinese data. For evaluation, we adopt three test sets: Seed-TTS (Anastassiou et al. 2024) *test-en* with 1088 samples from CommonVoice (Ardila et al. 2019), and Seed-TTS *test-zh* with 2020 samples from DiDiSpeech(Guo et al. 2021).

Training For our teacher model, we adopt F5-TTS (Chen et al. 2024d) with approximately 300M parameters, trained for 2M steps on the Emilia dataset. We maintain the same hyperparameter configuration as in the original F5-TTS, with a batch size of 307,200 audio frames (0.91 hours), using the AdamW optimizer (Loshchilov and Hutter 2018) with a peak learning rate of 7.5e-5, linear warmup for 20K updates, and linear decay afterwards. For the student model training in DMOSpeech 2, we follow the approach in Li, Kumar, and Jin (2024) but use half the batch size of the teacher model training. The learning rate for the student model resumes from the final learning rate of the teacher model training (around 6e-5) and continues for an additional 200K steps on the Emilia dataset. The duration predictor is initially trained on the Emilia dataset for 85K steps with a learning rate of 1e-4 and the same batch size as the F5-TTS teacher training. We use the AdamW optimizer with default parameters of Pytorch. After this initial training, we further fine-tune the duration predictor using GRPO (Sun et al. 2025) for an additional 1.5K steps with a group size of 16. All experiments were conducted on 8 NVIDIA H100 GPUs.

Baselines We compare several configurations of our models with both subjective and objective evaluations: (1) The

Model	*Seed-TTS-en*		*Seed-TTS-zh*		English		Chinese		CV_{f_0} ↑	RTF↓
	WER↓	SIM↑	CER↓	SIM↑	CMOS-N	CMOS-S	CMOS-N	CMOS-S		
Ground Truth	2.143	0.734	1.254	0.755	0.03	-0.13^{*}	0.02	-0.06	—	—
F5-TTS Teacher (32 steps)	1.947	0.662	1.695	0.750	-0.12^{*}	-0.04	-0.09	-0.11^{*}	**0.6659**	0.1671
DMOSpeech 2 (4 steps)	<u>1.752</u>	<u>0.698</u>	<u>1.527</u>	**0.760**	0.0	0.0	0.0	0.0	0.4640	**0.0316**
w/o duration predictor RL	3.750	0.672	2.000	0.750	-0.43^{**}	-0.48^{**}	-0.26^{*}	-0.31^{*}	S/A	S/A
Teacher-Guided (16 steps)	**1.738**	**0.699**	**1.468**	**0.760**	0.01	-0.03	0.45^{**}	0.3^{*}	<u>0.5932</u>	<u>0.0941</u>

Table 1: Objective and subjective evaluation results on *Seed-TTS-en* and *Seed-TTS-zh* evaluation sets. CMOS-S and CMOS-N refer to CMOS for similarity and naturalness, respectively, with DMOSpeech 2 (our system with 4 sampling steps) as the anchor (negative means DMOSpeech 2 is better). The best values for objective evaluations are shown in bold and the second-best values are underlined where S/A stands for the same as above. For subjective evaluations, the statistically significant results are marked by one asterisk if $p < 0.05$ and two asterisks if $p < 0.01$. CV_{f_0} is computed with the DDPM sampler for fairness.

ground truth recordings, (2) F5-TTS teacher without a duration predictor using 32 sampling steps, (3) DMOSpeech 2 with the RL-optimized duration predictor using 4 sampling steps, (4) student with the duration predictor before RL using 4 sampling steps, and (5) a teacher-guided sampling approach where the teacher model handles initial denoising steps before transitioning to the student model ($t_{switch} = 0.25$, with teacher handling 14 steps and student handling 2 steps, for a total of 16 steps). We use the pretrained Vocos vocoder (Siuzdak 2023) to convert generated mel-spectrograms to audio signals. We also compare our DMOSpeech 2 with several state-of-the-art TTS systems on objecetive metrics: CosyVoice 2 (Du et al. 2024b), Spark-TTS (Wang et al. 2025), LLaSA-8B (Ye et al. 2025), MaskGCT (Wang et al. 2024), and our F5-TTS teacher model (32 steps) (Chen et al. 2024d). All samples were resampled to 24 kHz for a fair comparison.

4.2 Evaluation Metrics

We evaluate models on the cross-sentence task (Le et al. 2024), where the model synthesizes speech with a prompt speaker's voice characteristics. For objective evaluation, we report: word error rate (WER) using Whisper-large-v3 (Radford et al. 2023) for English and Paraformer-zh (Gao et al. 2023) for Chinese (Anastassiou et al. 2024); SIM-o using WavLM-large-based (Chen et al. 2022) speaker verification to calculate cosine similarity between synthesized and ground truth speeches; RTF (real-time factor) on a single H100 GPU; and coefficient of variation of pitch (CV_{f_0}) from 50 samples per text-prompt pair across 20 pairs to measure diversity. For the teacher, we used DDPM (Ho, Jain, and Abbeel 2020) modified for flow-matching (Gao et al. 2024) to ensure fair comparison with students having additional noise injections (Algorithm 3). For subjective evaluation, we conduct CMOS tests for naturalness and similarity. Evaluators compare randomly ordered synthesized speech from test models against our anchor (DMOSpeech 2 with RL-optimized duration predictor, 4 sampling steps), rating which sounds more human-like and similar to the prompt (1). We report averages from 320 samples in both languages (details in Appendix D).

4.3 Results

Main Results Table 1 demonstrates DMOSpeech 2 with RL-optimized duration predictor significantly outperforms both teacher and student without optimization. For English: WER improves to 1.752 (vs. 1.947 F5-TTS, 3.750 DMOSpeech w/o RL) and SIM reaches 0.698 (vs. 0.662 for F5-TTS and 0.672 for DMOSpeech w/o RL). For Chinese, CER achieves 1.527 and SIM 0.760 (vs. $1.695/0.750$ for F5-TTS and $2.000/0.750$for DMOSpeech w/o RL). DMOSpeech 2 maintains an RTF of 0.0316, $5\times$ faster than the teacher's 0.1671. Teacher-guided sampling achieves WER of 1.738 and CER of 1.468 with silghtly increased computation. CMOS evaluation further confirms our approach's effectiveness. DMOSpeech 2 significantly outperforms DMOSpeech without RL optimization across both languages. For English, we observe strong improvements in naturalness (CMOS-N$= -0.43$) and similarity (CMOS-S$= -0.48$), both highly significant at $p < 0.01$. Chinese shows similar trends with naturalness at -0.26 and similarity at -0.31, significant at $p < 0.05$. When compared to F5-TTS, DMOSpeech 2 also demonstrates advantages, achieving significantly better English naturalness (CMOS-N$= -0.12$) and Chinese similarity (CMOS-S $= -0.11$), both at $p < 0.05$. The teacher-guided sampling variant particularly excels for Chinese, achieving remarkable scores of $+0.45$ for naturalness ($p < 0.01$) and $+0.3$ for similarity ($p < 0.05$). Most impressively, DMOSpeech 2 produces speech that is statistically indistinguishable from ground truth recordings in terms of naturalness for both English and Chinese. For English similarity, it even slightly surpasses ground truth with a CMOS-S score of $-0.13 (p < 0.05)$, demonstrating the effectiveness of our approach in producing human-quality speech.

Comparison with State-of-the-Art Models Table 2 shows DMOSpeech 2 outperforms previous state-of-the-art models. Student-only DMOSpeech 2 achieves English WER of 1.752 and Chinese CER of 1.527, surpassing all similar-sized models, including teacher F5-TTS (WER$= 1.947$, CER$= 1.695$) while being $5.3\times$ faster. Teacher-guided variant improves WER to 1.738 and CER to 1.468, maintaining $1.8\times$ speed advantage, although with doubled parameters (0.6B vs 0.3B) due to the need to store the teacher's parameters. Speaker similarity scores (0.698 to 0.699 English, 0.760 Chinese) exceed most baselines except MaskGCT, which trades superior similarity for worse intelligibility and $75\times$ slower inference with an RTF of 2.397. DMOSpeech 2 with only 0.3B parameters outperforms LLaSA-8B across all metrics, demonstrating

Model	#Params	Dataset (# Hours)	*Seed-TTS-en*		*Seed-TTS-zh*		RTF↓
			WER↓	SIM↑	CER↓	SIM↑	
Ground Truth	–	–	2.143	0.734	1.254	0.755	–
F5-TTS (32 steps) (Chen et al. 2024d)	0.3B	Emilia (He et al. 2024) (95k hrs)	1.947	0.662	1.695	0.750	0.167
CosyVoice 2 (Du et al. 2024b)	0.5B	Proprietary (200k hrs)	3.358	0.641	1.582	0.754	0.527
Spark-TTS (Wang et al. 2025)	0.5B	VoxBox (Wang et al. 2025) (100k hrs)	2.308	0.572	1.717	0.657	1.784
MaskGCT (Wang et al. 2024)	0.7B	Emilia (He et al. 2024) (95k hrs)	2.622	**0.713**	2.395	**0.772**	2.397
LLaSA-8B (Ye et al. 2025)	8B	Proprietary (200k hrs)	3.994	0.594	4.214	0.671	1.374
DMOSpeech 2 (Student-Only, 4 steps)	0.3B	Emilia (He et al. 2024) (95k hrs)	1.752	0.698	1.527	0.760	**0.032**
DMOSpeech 2 (Teacher-Guided, 16 steps)	0.6B	Emilia (He et al. 2024) (95k hrs)	**1.738**	0.699	**1.468**	0.760	0.094

Table 2: Comparison with state-of-the-art models on *Seed-TTS-en* and *Seed-TTS-zh* evaluation sets. The best values in each column are shown in bold and the second-best values are underlined. All samples from baseline models were synthesized using the official checkpoints released by the authors.

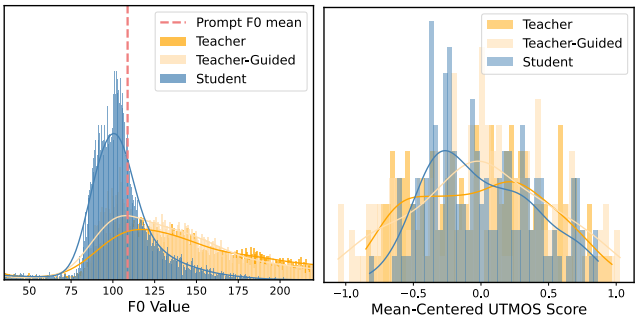

Figure 3: Comparison of diversity across sampling methods. (a) F0 value distributions (shown as histograms and kernel density estimates). The student model (light blue) exhibits a much narrower distribution compared to the teacher model (yellow), indicating mode shrinkage in prosodic patterns. The teacher-guided approach (orange) successfully recovers much of this diversity. (b) Mean-centered UTMOS score distributions. Acoustic quality remains consistent across all models despite differences in prosodic diversity, supporting our hypothesis that diversity reduction primarily affects prosodic and temporal aspects rather than acoustics.

targeted optimization's superiority over pure scaling. With an RTF of 0.032, DMOSpeech 2 is $5.2\times$ faster than F5-TTS, $16.5\times$ faster than CosyVoice 2, $55.8\times$ faster than Spark-TTS, and $42.9\times$ faster than LLaSA-8B.

Effect of Teacher-Guided Sampling on Diversity As shown in Table 1, teacher-guided sampling successfully addresses diversity limitations in our distilled student model. The coefficient of variation of pitch (CV_{f_0}) reveals the teacher model's superior diversity (0.6659) compared to the student model's reduced variation (0.4640, a 30.3% decrease), indicating the student model suffers from mode shrinkage. Our teacher-guided approach recovers much of this diversity (0.5932, 89.1% of teacher's diversity) while maintaining superior WER and speaker similarity from the student model with direct metric optimization. Figure 3a illustrates this effect through F0 distributions. The student model shows a narrower, more peaked distribution than the teacher model, demonstrating mode shrinkage from aggressive step reduction. The teacher-guided approach successfully broadens this distribution. In Figure 3b, we plot the mean-centered UTMOS score distributions since different models demonstrate significant differences in their mean UTMOS scores. Despite this, the mean-centered distributions after remain consistent across all models, indicating diversity reduction occurs primarily in prosodic aspects rather than spectral characteristics. This hybrid approach achieves a favorable trade-off between efficiency (RTF = 0.0941) and output diversity by leveraging the teacher model for establishing prosodic structure and the student model for efficient acoustic refinement.

5 Conclusion

This paper introduces DMOSpeech 2, which addresses two critical limitations in end-to-end diffusion-based TTS systems: optimizing the duration predictor for perceptual metrics and mitigating diversity reduction in distilled models. Through GRPO-based reinforcement learning, we directly optimize the duration predictor for speaker similarity and intelligibility, while teacher-guided sampling restores prosodic diversity. DMOSpeech 2 significantly outperforms state-of-the-art models across all metrics while maintaining exceptional computational efficiency. Optimizing the previously isolated duration predictor marks significant progress in pipeline-level metric optimization for TTS. Future work could apply our targeted RL approach to other difficult-to-optimize generative pipeline components, such as the teacher model in hybrid sampling, and explore additional rewards beyond WER and SIM for better human perception alignment.

DMOSpeech 2 presents important societal considerations. While offering significant benefits for accessibility, personalized assistants, and content creation through improved speaker similarity and intelligibility, it also poses risks for voice spoofing and deepfakes. Our approach's computational efficiency democratizes access, amplifying both benefits and risks. Addressing these concerns requires developing robust synthetic speech detection methods and establishing appropriate governance frameworks. We will release our source code and pre-trained models publicly, believing open-source development will accelerate progress in addressing both technical challenges and ethical considerations associated with advanced TTS systems.

References

Anastassiou, P.; Chen, J.; Chen, J.; and et al. 2024. SeedTTS: A Family of High-Quality Versatile Speech Generation Models. *ArXiv*, abs/2406.02430.

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Casanova, E.; Shulby, C.; Gölge, E.; Müller, N. M.; De Oliveira, F. S.; Junior, A. C.; Soares, A. d. S.; Aluisio, S. M.; and Ponti, M. A. 2021. SC-GlowTTS: An efficient zero-shot multi-speaker text-to-speech model. *arXiv preprint arXiv:2104.05557*.

Casanova, E.; Weber, J.; Shulby, C. D.; Junior, A. C.; Gölge, E.; and Ponti, M. A. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, 2709–2720. PMLR.

Chen, C.; Hu, Y.; Wu, W.; Wang, H.; Chng, E. S.; and Zhang, C. 2024a. Enhancing zero-shot text-to-speech synthesis with human feedback. *arXiv preprint arXiv:2406.00654*.

Chen, J.; Byun, J.-S.; Elsner, M.; and Perrault, A. 2024b. Reinforcement learning for fine-tuning text-to-speech diffusion models. *arXiv preprint arXiv:2405.14632*.

Chen, S.; Liu, S.; Zhou, L.; Liu, Y.; Tan, X.; Li, J.; Zhao, S.; Qian, Y.; and Wei, F. 2024c. VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2406.05370*.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

Chen, Y.; Niu, Z.; Ma, Z.; Deng, K.; Wang, C.; Zhao, J.; Yu, K.; and Chen, X. 2024d. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Choi, H.-S.; Yang, J.; Lee, J.; and Kim, H. 2022. NANSY++: Unified voice synthesis with neural analysis and synthesis. *arXiv preprint arXiv:2211.09407*.

Du, C.; Guo, Y.; Wang, H.; Yang, Y.; Niu, Z.; Wang, S.; Zhang, H.; Chen, X.; and Yu, K. 2025. Vall-t: Decoder-only generative transducer for robust and decoding-controllable text-to-speech. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Du, Z.; Chen, Q.; Zhang, S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; Zheng, S.; Gu, Y.; Ma, Z.; et al. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Du, Z.; Wang, Y.; Chen, Q.; Shi, X.; Lv, X.; Zhao, T.; Gao, Z.; Yang, Y.; Gao, C.; Wang, H.; et al. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Eskimez, S. E.; Wang, X.; Thakker, M.; Li, C.; Tsai, C.-H.; Xiao, Z.; Yang, H.; Zhu, Z.; Tang, M.; Tan, X.; et al. 2024. E2 TTS: Embarrassingly Easy Fully Non-Autoregressive Zero-Shot TTS. *arXiv preprint arXiv:2406.18009*.

Gao, R.; Hoogeboom, E.; Heek, J.; Bortoli, V. D.; Murphy, K. P.; and Salimans, T. 2024. Diffusion Meets Flow Matching: Two Sides of the Same Coin.

Gao, X.; Zhang, C.; Chen, Y.; Zhang, H.; and Chen, N. F. 2025. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Gao, Z.; Li, Z.; Wang, J.; Luo, H.; Shi, X.; Chen, M.; Li, Y.; Zuo, L.; Du, Z.; Xiao, Z.; et al. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*.

Guo, T.; Wen, C.; Jiang, D.; Luo, N.; Zhang, R.; Zhao, S.; Li, W.; Gong, C.; Zou, W.; Han, K.; et al. 2021. Didispeech: A large scale mandarin speech corpus. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6968–6972. IEEE.

He, H.; Shang, Z.; Wang, C.; Li, X.; Gu, Y.; Hua, H.; Liu, L.; Yang, C.; Li, J.; Shi, P.; Wang, Y.; Chen, K.; Zhang, P.; and Wu, Z. 2024. Emilia: An Extensive, Multilingual, and Diverse Speech Dataset For Large-Scale Speech Generation. *2024 IEEE Spoken Language Technology Workshop (SLT)*, 885–890.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hussain, S.; Neekhara, P.; Yang, X.; Casanova, E.; Ghosh, S.; Desta, M. T.; Fejgin, R.; Valle, R.; and Li, J. 2025. Koel-TTS: Enhancing LLM based Speech Generation with Preference Alignment and Classifier Free Guidance. *arXiv preprint arXiv:2502.05236*.

Ichihara, Y.; Jinnai, Y.; Morimura, T.; Ariu, K.; Abe, K.; Sakamoto, M.; and Uchibe, E. 2025. Evaluation of Best-of-N Sampling Strategies for Language Model Alignment. *arXiv preprint arXiv:2502.12668*.

Ju, Z.; Wang, Y.; Shen, K.; Tan, X.; Xin, D.; Yang, D.; Liu, Y.; Leng, Y.; Song, K.; Tang, S.; et al. 2024. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Le, M.; Vyas, A.; Shi, B.; Karrer, B.; Sari, L.; Moritz, R.; Williamson, M.; Manohar, V.; Adi, Y.; Mahadeokar, J.; et al. 2024. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36.

Lee, K.; Kim, D. W.; Kim, J.; and Cho, J. 2024. DiTTo-TTS: Efficient and Scalable Zero-Shot Text-to-Speech with Diffusion Transformer. *arXiv preprint arXiv:2406.11427*.

Lee, S.-H.; Kim, S.-B.; Lee, J.-H.; Song, E.; Hwang, M.-J.; and Lee, S.-W. 2022. HierSpeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis. *Advances in Neural Information Processing Systems*, 35: 16624–16636.

Li, Y. A.; Han, C.; and Mesgarani, N. 2022. Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. *arXiv preprint arXiv:2205.15439*.

Li, Y. A.; Han, C.; Raghavan, V.; Mischler, G.; and Mesgarani, N. 2024a. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36.

Li, Y. A.; Jiang, X.; Han, C.; and Mesgarani, N. 2024b. StyleTTS-ZS: Efficient High-Quality Zero-Shot Text-to-Speech Synthesis with Distilled Time-Varying Style Diffusion. *arXiv preprint arXiv:2409.10058*.

Li, Y. A.; Kumar, R.; and Jin, Z. 2024. DMOSpeech: Direct Metric Optimization via Distilled Diffusion Model in Zero-Shot Speech Synthesis. *arXiv preprint arXiv:2410.11097*.

Liu, Z.; Wang, S.; Inoue, S.; Bai, Q.; and Li, H. 2024. Autoregressive diffusion transformer for text-to-speech synthesis. *arXiv preprint arXiv:2406.05551*.

Loshchilov, I.; and Hutter, F. 2018. Fixing Weight Decay Regularization in Adam.

Min, D.; Lee, D. B.; Yang, E.; and Hwang, S. J. 2021. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. In *International Conference on Machine Learning*, 7748–7759. PMLR.

Peng, P.; Huang, P.-Y.; Li, D.; Mohamed, A.; and Harwath, D. 2024. VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. *arXiv preprint arXiv:2403.16973*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Saeki, T.; Xin, D.; Nakata, W.; Koriyama, T.; Takamichi, S.; and Saruwatari, H. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, K.; Ju, Z.; Tan, X.; Liu, Y.; Leng, Y.; He, L.; Qin, T.; Zhao, S.; and Bian, J. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Siuzdak, H. 2023. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *arXiv preprint arXiv:2306.00814*.

Song, X.; Xing, M.; Ma, C.; Li, S.; Wu, D.; Zhang, B.; Pan, F.; Zhou, D.; Zhang, Y.; Lei, S.; et al. 2024. TouchTTS: An Embarrassingly Simple TTS Framework that Everyone Can Touch. *arXiv preprint arXiv:2412.08237*.

Sun, X.; Xiao, R.; Mo, J.; Wu, B.; Yu, Q.; and Wang, B. 2025. F5R-TTS: Improving Flow Matching based Text-to-Speech with Group Relative Policy Optimization. *arXiv preprint arXiv:2504.02407*.

Tan, X.; Chen, J.; Liu, H.; Cong, J.; Zhang, C.; Liu, Y.; Wang, X.; Leng, Y.; Yi, Y.; He, L.; et al. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tian, J.; Zhang, C.; Shi, J.; Zhang, H.; Yu, J.; Watanabe, S.; and Yu, D. 2025. Preference alignment improves language model-based tts. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Wang, H.; Liang, C.; Wang, S.; Chen, Z.; Zhang, B.; Xiang, X.; Deng, Y.; and Qian, Y. 2023b. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Wang, X.; Jiang, M.; Ma, Z.; Zhang, Z.; Liu, S.; Li, L.; Liang, Z.; Zheng, Q.; Wang, R.; Feng, X.; et al. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Wang, Y.; Zhan, H.; Liu, L.; Zeng, R.; Guo, H.; Zheng, J.; Zhang, Q.; Zhang, X.; Zhang, S.; and Wu, Z. 2024. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Wu, Y.; Tan, X.; Li, B.; He, L.; Zhao, S.; Song, R.; Qin, T.; and Liu, T.-Y. 2022. Adaspeech 4: Adaptive text to speech in zero-shot scenarios. *arXiv preprint arXiv:2204.00436*.

Yang, D.; Huang, R.; Wang, Y.; Guo, H.; Chong, D.; Liu, S.; Wu, X.; and Meng, H. 2024. SimpleSpeech 2: Towards Simple and Efficient Text-to-Speech with Flow-based Scalar Latent Transformer Diffusion Models. *arXiv preprint arXiv:2408.13893*.

Ye, Z.; Zhu, X.; Chan, C.-M.; Wang, X.; Tan, X.; Lei, J.; Peng, Y.; Liu, H.; Jin, Y.; DAI, Z.; et al. 2025. Llasa: Scaling Train-Time and Inference-Time Compute for Llama-based Speech Synthesis. *arXiv preprint arXiv:2502.04128*.

Yin, T.; Gharbi, M.; Park, T.; Zhang, R.; Shechtman, E.; Durand, F.; and Freeman, W. T. 2024a. Improved Distribution Matching Distillation for Fast Image Synthesis. *arXiv preprint arXiv:2405.14867*.

Yin, T.; Gharbi, M.; Zhang, R.; Shechtman, E.; Durand, F.; Freeman, W. T.; and Park, T. 2024b. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6613–6623.

Zhang, D.; Li, Z.; Li, S.; Zhang, X.; Wang, P.; Zhou, Y.; and Qiu, X. 2024. Speechalign: Aligning speech generation to human preferences. *arXiv preprint arXiv:2404.05600*.

Zhu, X.; Tian, W.; and Xie, L. 2024. Autoregressive Speech Synthesis with Next-Distribution Prediction. *arXiv preprint arXiv:2412.16846*.