BRIDGING THE GAP BETWEEN TRAINING AND INFERENCE IN LM-BASED TTS MODELS

Ruonan Zhang¹, Lingzhou Mu¹, Xixin Wu², and Kai Zhang^{1,*}

¹Tsinghua University, ²The Chinese University of Hong Kong

ABSTRACT

Recent advancements in text-to-speech (TTS) have shown that language model (LM) based systems offer competitive performance compared to traditional approaches. However, in training, TTS models use ground-truth (GT) tokens as prefixes to predict the next token, while in inference these tokens are not available, a gap between training and inference that is often neglected. In this study, we propose a promptguided hybrid training scheme to mitigate exposure bias in popular LM-based TTS systems. Our core idea is to adopt a hybrid training paradigm that combines teacher forcing with free running, thereby introducing self-generated tokens into the training process. This makes the training mode more consistent with inference, reducing the training-inference gap. In addition, we incorporate an EOS prediction mechanism during training to detect incorrect sequence termination and adaptively control the free running process. Experimental results provide a comprehensive evaluation of the impact of exposure bias on LM-based TTS, and demonstrate that our method effectively narrows the training-inference gap, thereby improving the quality of synthesized long-form speech.

Index Terms— Exposure bias, TTS, training - inference gap

1. INTRODUCTION

Recent years have witnessed rapid advancements in TTS synthesis models, enabling them to generate highly natural and expressive speech. [1,2]. These models commonly discretize speech signals into token sequences and adopt autoregressive next - token prediction, leveraging the generative capabilities of pre-trained Large Language Models (LLMs) for sequence modeling [3]. Most LM-based TTS models [4–6] suffer from a fundamental limitation, the training - inference gap, commonly known as exposure bias [7,8]. While training relies on teacher forcing with GT tokens as input, inference requires the model to generate tokens autoregressively based on its own previous predictions [3,9]. This discrepancy between training and inference often leads to cascading prediction errors.

Exposure bias is a common challenge in autoregressive

sequence modeling tasks [10-12]. In text generation tasks such as neural machine translation [10], exposure bias may lead to semantic drift. To alleviate exposure bias, scheduled sampling [7, 8, 13] is proposed to mitigate the traininginference gap. However, its impact is magnified in speech synthesis due to the high density of acoustic tokens. Unlike text, token-based speech synthesis operates at a much higher temporal resolution, where minor errors are not as easily tolerated. This high resolution typically requires 50-100 discrete tokens per second to represent continuous speech. A single syllable or phoneme is often mapped to several tokens. Consequently, even minor prediction errors can spread rapidly and negatively impact the listener's experience. Moreover, speech exhibits additional attributes such as speaking rate and prosody, which are absent in textual representations but play a important role in evaluation in TTS. This effect amplifies the impact of exposure bias, particularly for long-form speech synthesis. [14, 15].

The manifestations of exposure bias are diverse. First, TTS models tend to generate speech at an accelerated speed during long utterance synthesis. The second issue is EOS misprediction, which results in early stopping or repeating parts of the speech. Additionally, the synthesized speech exhibits excessive prosodic flatness relative to ground truth. Thus, training-inference gap is obviously a significant challenge for LM-based TTS models.

To mitigate exposure bias, we propose a prompt-guided hybrid training scheme that gradually transitions from fully guided supervision to self-conditioned generation. As illustrated in Figure 1, our approach operates iteratively within each training step. First, it replaces a portion of the GT input tokens with self-generated tokens in a previous pass. Concurrently, we use a prompt protection strategy to ensure a specific part of the GT tokens is always preserved. This controlled exposure not only maintains training stability but also enhances generalization. We further introduce an adaptive free running strategy guided by End-of-Sequence (EOS) prediction. If the EOS is mispredicted, we end free running early by masking subsequent steps and proceed directly to gradient updates. Our experimental results demonstrate that the iterative hybrid training strategy enhances the model's dependency on its token history, compensating for the absence of GT tokens during inference. This approach also narrows the gap between

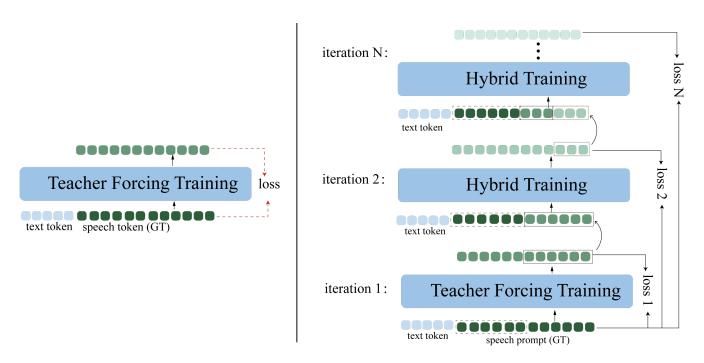


Fig. 1. The proposed Hybrid framework.

training and inference modes, leading to more stable and coherent sequence modeling.

2. PROPOSED APPROACH

2.1. Prompt-guided hybrid training scheme

Figure 1(right) illustrates our training framework in comparison with standard teacher forcing training (left). We propose a prompt-guided hybrid training scheme that integrates teacher forcing with free running. The basic idea of our training strategy is to iteratively put the self-generated tokens into the LLM being trained to simulate auto-regressive inference process and back propagate both teacher forcing loss and accumulated free running loss in a single training step. To further strengthen its prompt following ability, we randomly replace the starting speech tokens with GT tokens in later iterations. This design leverages the stability of teacher forcing while gradually introducing self-generated prefixes, thereby improving robustness against exposure bias.

Moreover, when prompt protection strategy is enabled and the first few speech tokens of input sequence is replaced by GT tokens, the replaced portion also provides an additional teacher forcing loss to further stabilize training process. As training progresses, we increase the number of replaced tokens to make a smooth shift from initial teacher forcing dominated training to self-generated token dominated training. This fosters a smooth transition, guiding the model toward robust and reliable self-generation.

2.2. Adaptive free running scheduling via EOS prediction

Earlier studies have found that teacher forcing demonstrates greater stability and faster convergence compared to free running. Therefore, too many iterations of free running may interfere with model convergence. To address this issue, adaptive free running scheduling via EOS prediction is introduced as a compensatory solution.

We observe that the predicted EOS token serves as a reliable indicator of exposure bias and training stability, allowing dynamic adjustment of free running iterations. When the EOS token is not correctly predicted, it suggests potential exposure bias or model degradation during training. In such cases, the model should increase its reliance on ground-truth supervision to correct its learning trajectory in subsequent steps. Conversely, when EOS token is successfully predicted in several consecutive training steps, the model increases the number of free running iterations to better align the inference procedure. This adaptive mechanism dynamically adjusts the number of free running iterations based on output quality, thereby improving both training efficiency and generation stability.

2.3. Training object

The overall training objective is influenced by two components: teacher forcing loss, which provides stable supervision using GT tokens, and a weighted sum of free running losses, which improves prediction quality under autoregressive conditions. As depicted in Figure 1(right), the first iteration of our training framework involves computing the teacher force-

ing loss \mathcal{L}_{TF} , identical to the cross-entropy loss in conventional LLM training. \mathcal{L}_{TF} can be described as follow:

$$\mathcal{L}_{TF} = -\sum_{t=1}^{T} \log P(y_t \mid y_{i < t}^{gt}, \mathbf{X})$$
 (1)

where, y_t denotes the $t_{\rm th}$ speech token generated from its GT prefixes $y_{i < t}^{\rm gt}$. X refers to the input prompt tokens, including SOS token, text embeddings, speaker embedding and other prompt tokens.

During the second to the last iteration, we replace the first T1 tokens of the output from the previous iterations with GT and then input this modified sequence into the model. We compute cross entropy loss between model output with GT tokens as follow:

$$\mathcal{L}_{FR} = -\sum_{t=1}^{T_1} \log P(y_t \mid y_{i < t}^{gt}, \mathbf{X}) - \sum_{t=T_1+1}^{T_2} \log P(y_t \mid y_{i < T_1}^{gt}, y_{T_1 < i < t}^{pred}, \mathbf{X})$$
 (2)

This free running loss $\mathcal{L}_{\mathrm{FR}}$ is a sum of two terms. The first term refers to the cross entropy between GT and the first T_1 speech tokens of model output. These tokens are generated based on GT prefixes $y_{i < t}^{\mathrm{gt}}$ and X. The second term refers to the cross entropy between GT and the rest of the speech tokens which is generated based on GT tokens $y_{i < T_1}^{\mathrm{gt}}$, predicted tokens $y_{T_1 < i < t}^{\mathrm{pred}}$, and X. Our overall training objective can be formulated as follow:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{TF}} + \sum_{n=1}^{N} w_n \mathcal{L}_{\text{FR}}^{(n)}$$
 (3)

where $\mathcal{L}_{\text{FR}}^{(n)}$ represents the free running loss from iteration n, and w_n is its associated weight. We use the weighted average of the loss from each iteration as the final loss to balance the effects across iterations.

3. EXPERIMENTS

3.1. Implementation details

The proposed prompt-guided hybrid training scheme is employed to fine-tuning the CosyVoice [16] and CosyVoice2 [17] on the LibriSpeech corpus [18], which contains approximately 40K hours of transcribed speech data. For evaluation, we adopt two test sets, LibriSpeech test-clean and SeedTTS [19], which includes 1,000 utterances sampled from the common voice corpus to reflect more diverse acoustic conditions. During training, we adopt the AdamW optimizer with a constant learning rate of 1×10^{-5} . Speech signals are discretized into token sequences using a vector quantizer with a single codebook of 4096 entries. We apply 10k warm-up steps before introducing the proposed hybrid training strategy.

3.2. Subjective evaluation

As shown in Table 1, we compare against the baselines IndexTTS [20], CosyVoice and CosyVoice2. CosyVoice-TF denotes fine-tuning the pretrained CosyVoice model using the standard teacher forcing scheme. CosyVoice2-TF is defined analogously. On CosyVoice2, our prompt-guided hybrid training strategy achieves the best overall performance. It yields superior Word Error Rate (WER) and speaker smilarity performance over baselines on LibriSpeech and relatively smaller improvement on Seed-TTS (where utterances are < 10s). The relatively small improvement on Seed-TTS dataset may be attributed to the shorter duration of its speech samples, as our method tends to perform better on longer speech samples by mitigating the accumulated error caused by exposure bias. It is also quite intuitive that longer sequence is easier to suffer from accumulated error in the process of autoregressive generation.

Table 1. Comparison results with existing methods.

System	LibriSpeech		Seed-TTS	
	WER↓	SIM↑	WER↓	SIM↑
GT	1.94	0.82	2.14	0.89
IndexTTS	8.30	0.68	6.53	0.84
CosyVoice - TF	8.17	0.67	6.54	0.78
CosyVoice2 - TF	6.23	0.74	4.83	0.83
CosyVoice - Ours	5.38	0.68	6.34	0.85
CosyVoice2 - Ours	4.21	0.78	4.64	0.84

3.3. Objective evaluation

For objective evaluation, we evaluate our method using both mean opinion score (MOS) and A/B preference tests with human listeners. Thirty samples are obtained from the LibriSpeech test-clean set, and the MOS results are shown in Figure 2. Our method outperforms teacher forcing training and yield comparable results with GT speech samples.

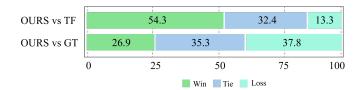


Fig. 2. Objective Evaluation

3.4. Visualization of exposure bias

To visualize the impact of exposure bias, we compute tokenlevel accuracy by comparing the predicted token index with the ground-truth token index at each position and calculating the proportion of exact matches. As shown in Figure 3, the blue line indicates the token prediction accuracy under teacher forcing, where GT tokens are used as context, while the yellow line corresponds to the accuracy when tokens are generated auto-regressively. We point out that speech token accuracy is inherently low, which is determined by the nature of speech token. To be specific, after speech is tokenized, similar sounds can be represented by different speech tokens, making sound-token mapping is not strictly one-to-one.

Despite the low accuracy, the relative difference between teacher forcing and free running inference remains informative for analyzing exposure bias. As demonstrated in Figure 3, a clear accuracy gap appears between teacher forcing (blue) and free-running inference (yellow). This discrepancy highlights the presence of exposure bias. In contrast, our hybrid-trained model not only boosts the absolute accuracy in both modes but also significantly narrows the gap between them.

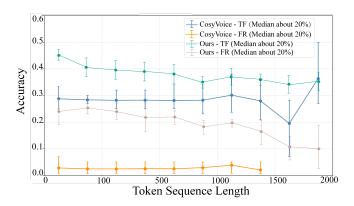


Fig. 3. Exposure bias illustrated by accuracy gap between teacher forcing and free running.

3.5. Validation of EOS-guided scheduling

To validate EOS guided adaptive iteration introduced in Section 2.2, we track the number of iterations before the model incorrectly predicts a premature EOS token. Experiments are conducted under different maximum iteration settings (2, 4, and 6). As shown in Figure 4, during the early phase of hybrid training, premature EOS often occurs within the first few iterations, suggesting that the model is not fully prepared for free running. As training goes on, the occurrence of premature EOS shifts to later steps, suggesting improved robustness under free running conditions. Based on this observation, we increase the number of free running iterations when premature EOS predictions are less frequent. Notably, despite the gradual increase in iterations during training, our method requires only 1.5× the computation of the baseline and converges in fewer steps. This training efficiency attributed to the fact that we only do backward pass once in a training step.

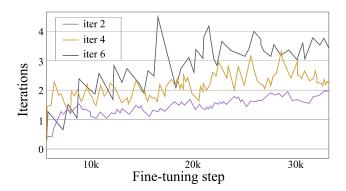


Fig. 4. Frequency of early termination vs. iteration number. The rising trend indicates improved EOS error detection and adaptive free running, with increased quality control at the cost of computation.

3.6. Ablation Study

As summarized in Table 2, we conduct an ablation study to validate the effectiveness. The results show that removing either prompt protection or EOS adaptive leads to higher WER and lower speaker similarity. The degradation is more pronounced when Prompt Protection is removed.

Table 2. Ablation Study. w refers to our training framework with all method enabled.

systems	WER↓	SIM ↑
w/o Prompt Protection	7.25	0.65
w/o EOS Adaptive	4.98	0.72
W	4.21	0.80

4. CONCLUSION

We propose a novel training framework for LM-based TTS to align the training process with autoregressive inference. By constructing a hybrid input that mixes GT tokens with the model's self-generated tokens, our method effectively mitigates exposure bias. This strategy effectively simulates the data distribution encountered during inference. The prompt protection mechanism guides the model to gradually shift from supervised guidance to self-sufficient autoregressive generation. Concurrently, it offers a reliable feedback EOS signal for dynamically tng the free-running scheduling strategy. In our experiments, we first visualize the significant exposure bias between the training and inference stages. Furthermore, the framework alleviates the problem of the model struggling to predict the EOS token during free running generation, a classic symptom of exposure bias. The results demonstrate that our proposed training framework significantly improves speech synthesis quality, especially for long-form speech synthesis.

5. REFERENCES

- [1] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 6706–6713.
- [2] Yakun Song, Zhuo Chen, Xiaofei Wang, Ziyang Ma, and Xie Chen, "Ella-v: Stable neural codec language modeling with alignment-guided sequence reordering," arXiv preprint arXiv:2401.07333, 2024.
- [3] Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, et al., "Autoregressive speech synthesis without vector quantization," *arXiv preprint arXiv:2407.08551*, 2024.
- [4] Chen Chen, Yuchen Hu, Wen Wu, Helin Wang, Eng Siong Chng, and Chao Zhang, "Enhancing zeroshot text-to-speech synthesis with human feedback," arXiv preprint arXiv:2406.00654, 2024.
- [5] Ruibo Fu, Xin Qi, Zhengqi Wen, Jianhua Tao, Tao Wang, Chunyu Qiang, Zhiyong Wang, Yi Lu, Xiaopeng Wang, Shuchen Shi, et al., "Asrrl-tts: Agile speaker representation reinforcement learning for text-to-speech speaker adaptation," arXiv preprint arXiv:2407.05421, 2024.
- [6] Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen, "Emo-dpo: Controllable emotional speech synthesis through direct preference optimization," in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.
- [7] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," *Advances in neural in*formation processing systems, vol. 28, 2015.
- [8] Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu, "Bridging the gap between training and inference for neural machine translation," *arXiv preprint arXiv:1906.02448*, 2019.
- [9] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen, "F5tts: A fairytaler that fakes fluent and faithful speech with flow matching," arXiv preprint arXiv:2410.06885, 2024.
- [10] Qingkai Fang and Yang Feng, "Understanding and bridging the modality gap for speech translation," 2023.
- [11] Chaojun Wang and Rico Sennrich, "On exposure bias, hallucination and domain shift in neural machine translation," 2020.

- [12] 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 textto-speech," 2025.
- [13] Daniel Duckworth, Arvind Neelakantan, Ben Goodrich, Lukasz Kaiser, and Samy Bengio, "Parallel scheduled sampling," *arXiv preprint arXiv:1906.04331*, 2019.
- [14] Jinchuan Tian, Chunlei Zhang, Jiatong Shi, Hao Zhang, Jianwei Yu, Shinji Watanabe, and Dong Yu, "Preference alignment improves language model-based tts," in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.
- [15] Yinghao Aaron Li, Cong Han, and Nima Mesgarani, "Styletts: A style-based generative model for natural and diverse text-to-speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, 2025.
- [16] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al., "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," arXiv preprint arXiv:2407.05407, 2024.
- [17] 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," 2024.
- [18] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.
- [19] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al., "Seed-tts: A family of high-quality versatile speech generation models," *arXiv* preprint arXiv:2406.02430, 2024.
- [20] Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang, "Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system," 2025.