F5D-TTS: Aligning Flow-Matching Text-to-Speech Models with Direct Preference Optimization

Anonymous ACL submission

Abstract

Speech generation has achieved significant advances through flow-matching techniques. And reinforcement learning from human feedback shown the possibility of further improving the performance of speech generation models. In this work ,we propose F5D-TTS, a framework for integrating flow-matching text-to-speech models to human preferences by directly optimizing on human preference data pairs. Specifically, we start by using the base F5-TTS model to create a preference dataset with 10000 speech pairs (about 2 hours), with a winner speech and a loser speech generated in the same prompt. Then we align flow matching TTS model with Flow-DPO. Experiments show that F5D-TTS significantly outperforms both the base F5-TTS model and the supervised-finetuned F5-TTS model in speaker similarity (measured by SIM-O) while maintaining speech intelligence (measured by WER) and speech naturalness (measured by UTMOS). We also show Flow-DPO alignment is applicable to low-resource scenarios. Audio samples are available at https://demo-used.github.io/F5D-TTS

1 Introduction

Recent advancements in Text-to-Speech (TTS) systems have achieved remarkable progress in generating high-quality , natural and expressive speech. Current TTS model architectures bifurcate into two dominant paradigms: autoregressive (AR) (Wang et al., 2023; Zhang et al., 2023; Chen et al., 2024a; Song et al., 2025) and non-autoregressive (NAR) (Chen et al., 2024b; Wang et al., 2024; Ju et al., 2024; Lee et al., 2024; Jiang et al., 2025) modeling approaches. AR models typically employ sequential token prediction through speech codecs, leveraging language model architectures to achieve high-fidelity synthesis. NAR models utilize parallel generation techniques via denoising diffusion or flow matching, offering significantly faster inference while maintaining competitive audio quality.

The concurrent advancement of reinforcement learning (RL) techniques has opened new frontiers in generative model, which can enhance model performance by aligning with human preference rather than scaling up. Pioneering works in RL, exemplified by the DeepSeek series (Shao et al., 2024; Liu et al., 2024, 2025; Guo et al., 2025), demonstrates the efficacy of RL paradigms like Direct Preference Optimization (DPO) (Rafailov et al., 2023) and Group Relative Policy Optimization (GRPO) (Guo et al., 2025) for preference alignment. These methods have been successfully adapted to AR-TTS systems, where Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Direct Preference Optimization (DPO) optimize acoustic metrics such as speaker similarity (SIM-O) and word error rate (WER). However, integration of RL with NAR-TTS architectures remains conspicuously absent from the literature. This unresolved challenge presents a critical gap at the intersection of non-autoregressive speech synthesis and reinforcement learning research.

In this paper, we introduce F5D-TTS, which aligns flow-matching TTS models with Direct Preference Optimization (Flow-DPO) (Tian et al., 2025). The framework of F5D-TTS is shown in Fig.1. First, we generated an initial set of speech samples using the F5-TTS model and employed the SIM-O as the reward signal for preference selection. Through this process, we curated a total of 10,000 data pairs as our preference dataset. Subsequently, we fine-tuned the F5-TTS model with Flow-DPO. Experimental results demonstrate that significant improvement in speaker similarity can be achieved with low-resource scenarios (fine-tuning by only 1,000 data pairs and 1,000 optimization steps). The main

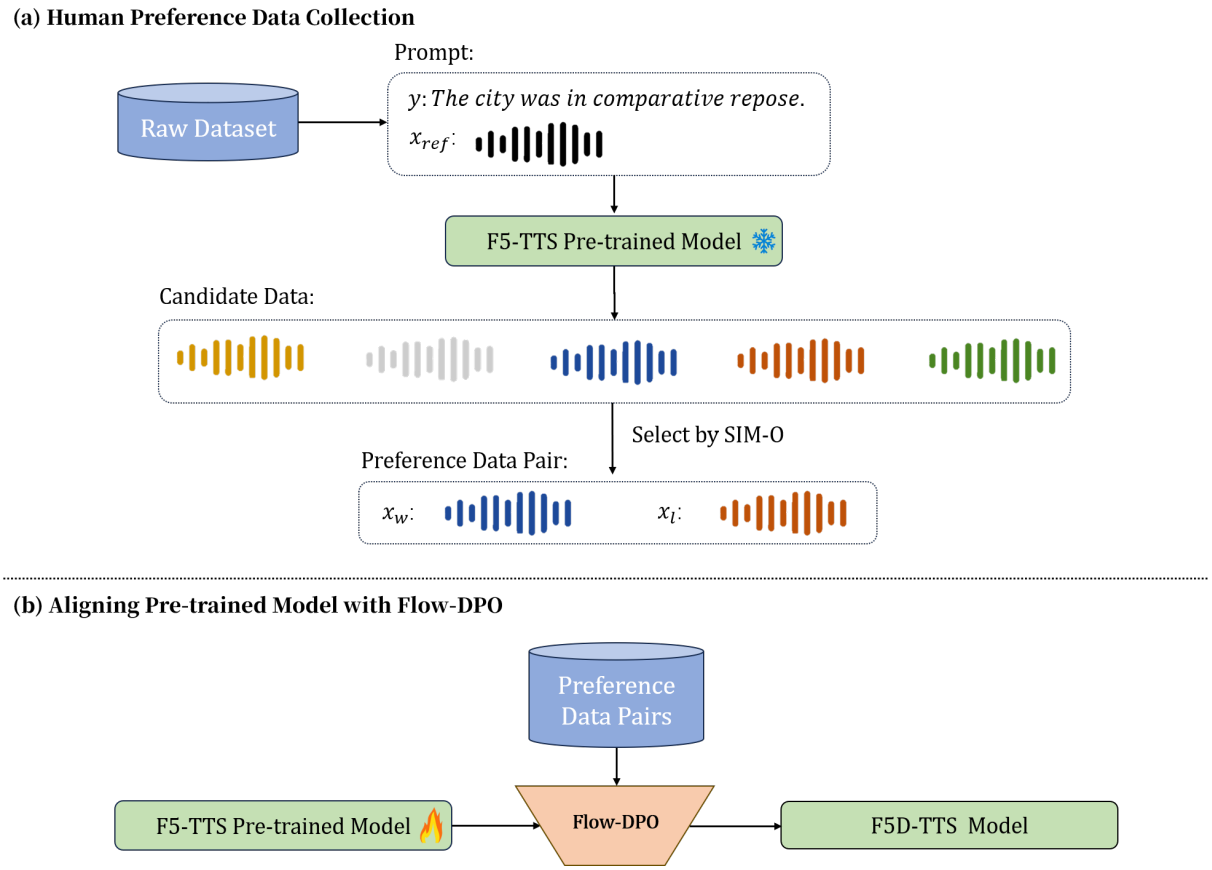

Figure 1: Framework of F5D-TTS. **(a) Human Preference Data Collection.** (Sec 3.2) We select a subset of LibriTTS dataset as prompts. Then we utilize the F5-TTS model to generate speech data for five times and select the winner speech and the loser speech using SIM-O as reward signals. We create 10,000 preference data pairs in total. **(B) Aligning Pre-trained Model with Flow-DPO.** (Sec 3.3) Once we get preference data pairs, we use them to fine-tune the F5-TTS model with Flow-DPO. In this work, we use 1,000 pairs (about 2 hours) of speech data for fine-tuning the F5-TTS pre-trained model.

contributions of this work are as follows:

- We introduce F5D-TTS, which aligns flow-matching TTS models with DPO. In this way, we can enhance model performance by aligning with human preference rather than simply scaling up.
- Experiments show that F5D-TTS can significantly improve speaker similarity while maintaining speech intellience and speech naturalness with low-resource scenarios.

2 Related Works

Aligning AR-TTS System. Reinforcement learning (RL) techniques have been applied in autoregressive text-to-speech (AR-TTS) systems to enhance performance through reward-driven optimization. In Seed-TTS (Anastassiou et al., 2024), reinforcement learning (RL) methods are employed to enhance system performance by fine-tuning the base model with tailored reward functions. Specifically, the REINFORCE (Ahmadian et al., 2024) algorithm is utilized to optimize two variants derived from the zero-shot in-context learning model: one incorporates SIM-O and word error rate (WER) metrics as rewards to improve speaker similarity and robustness, while the other leverages speech emotion recognition (SER) accuracy to enhance emotion controllability. Evaluations are conducted on both objective and subjective test sets, including a challenging textual dataset designed to stress-test autoregressive models, with results demonstrating the effectiveness of RL in refining speech attributes. This approach highlights the adaptability of reward-driven optimization in balancing attribute-specific control and implementation simplicity for TTS systems. Other works (Adler et al., 2024; Tian et al., 2025; Gao et al., 2025; Hussain et al., 2025) use RL techniques and achieve some improvement.

Aligning NAR-TTS System. Aligning NAR-TTS system to human preferences has so far been much less explored than AR-TTS system. F5R-TTS (Sun et al., 2025) applys the GRPO method to NAR-TTS models, using WER and SIM as reward signals, and significantly improve the WER and SIM. The main method of this work is transforming the outputs of flow-matching TTS models into probabilistic representations, which cannot align flow-matching TTS models with human preference. The fine-tuning process of Group Relative Policy Optimization (GRPO) exhibits some constraints: Although GRPO reduces memory overhead compared to PPO, it still incurs substantial computational costs, rendering the GRPO methodology inapplicable in low-resource scenarios. Furthermore, compared to DPO, GRPO demonstrates suboptimal stability, which introduces additional implementation risks and constraints in practical applications.

3 Methodology

The core idea of F5D-TTS is aligning flow-matching TTS models with Direct Preference Optimization. To accomplish this, we first select some speech-text pairs as prompt and synthesize some candidate speech. Then we use SIM-O as reward signals to collect preference data pairs. And we use Flow-DPO, a method to align flow-based model with human preference to imporve the model performance in speaker similarity.

In the following sections, we first briefly review the F5-TTS model. Then, we introduce the method we used to construct the preference dataset. Finally, we present the method of aligning flow-matching TTS models with DPO. It is important to note that though we use the F5-TTS to conduct F5D-TTS, this method is applicable to any flow-matching TTS models.

3.1 Overview of F5-TTS

F5-TTS is a novel non-autoregressive text-to-speech synthesis system based on flow-matching and Diffusion Transformer, designed to achieve efficient and high-fidelity speech generation through a simplified architecture. Unlike traditional methods that rely on complex components (e.g., phoneme alignment, duration models, or pre-trained language models), F5-TTS adopts a minimalist end-to-end paradigm, directly concatenating padded text sequences with speech inputs and implicitly modeling text-speech semantic alignment via contextual learning. F5-TTS introduces ConvNeXt modules to optimize text representations, significantly enhancing the fidelity and naturalness of synthesized speech in zero-shot scenarios. Additionally, it proposes a dynamic inference strategy named Sway Sampling, which adjusts the sampling distribution of flow steps to optimize generation quality and efficiency without increasing training costs. The model demonstrates rapid convergence during training and achieves near-real-time synthesis speed during inference.

3.2 Human Preference Data Collection

To collect human preference data, we should generate speech samples from various prompts with the pre-trained F5-TTS model. This work conclude three stages as follows:

***Stage 1*: Create speech-text pairs for generation.** In this stage, we utilize LibriTTS (Zen et al., 2019), a multi-speaker English corpus of approximately 585 hours of read English speech at 24kHz sampling rate. In the initial experimental setup, we first curated a subset of the LibriTTS dataset, specifically, we selected speech samples with 3-15 second durations as prompts (denoted by x_{ref}), and randomly shuffled their corresponding transcripts as target texts (denoted by y). Through this process, we created 50,000 speech-text pairs for subsequent data generation.

***Stage 2*: Multiple inferences from the same prompt.** Following the creation of speech-text paired datasets, we employ the F5-TTS model to synthesize speech outputs. For each input prompt, defined as a tuple comprising textual content y and a reference speech x_{ref}, we perform five parallel synthesis iterations under identical hyperparameter configurations (e.g., fixed temperature, deterministic sampling seeds). This procedure yields a set of candidate speech samples (y, x_1, x_2, x_3, x_4, x_5) per prompt. By systematically aggregating these multi-output generations across the dataset, we construct a diversified pool of candidate preference pairs.

***Stage 3*: Ranking and preference data selection.** In stage 3, we need to select the winner speech x_w and the loser speech x_l from data pool created in stage 2. We choose SIM-O as reward signals to replace human feedback. We calculated SIM-O of all samples in data pool. For each prompt, we ranked the five generated speech samples by their SIM-O scores, designating the highest-scoring and lowest-

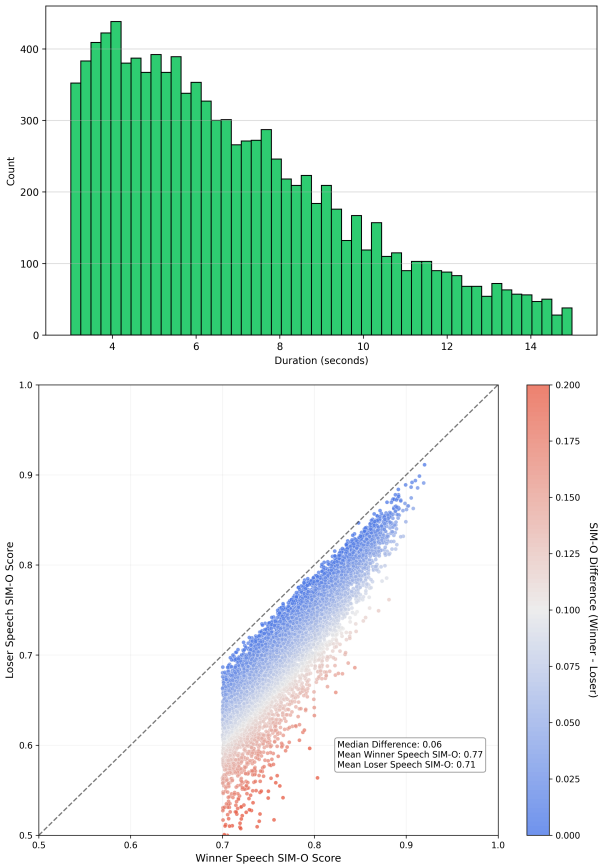

Figure 2: Statistics of data. These two figures show the distribution of speech durations and SIM-O of the winner speech and the loser speech.

scoring samples as the winner speech and loser speech respectively. This process yielded 50,000 ranked pairs (y, x_w, x_l) based on SIM-O metrics. Subsequently, we computed UTMOS scores for all winner speech and ultimately selected 10,000 high-quality pairs for Flow-DPO fine-tuning and data scale ablation studies. The distribution characteristics of these paired samples are illustrated in Fig.2.

3.3 Aligning Pre-trained Model with Flow-DPO

Reinforcement Learning from Human Feedback. Reinforcement Learning from Human Feedback (RLHF) is applied to learn a speech generation policy $\pi_\theta(x_0|y)$ that aligns with human preference. Given a dataset $\mathcal{D} = \{y, x_w, x_l\}$, where each sample includes prompt y and two outputs x_w (preferred) and x_l (dispreferred) synthesized by a reference model $\pi_{\text{ref}}(x|y)$, RLHF optimizes the policy to maximize a reward model $r(x, y)$. To prevent excessive divergence from the reference model, a KL-divergence regularization term weighted by β is incorporated. The optimization objective is formulated as Eq.1:

$$\begin{aligned}\max_{\pi_\theta} \mathbb{E}_{y\sim\mathcal{D},\, x_0\sim\pi_\theta(x_0|y)}\left[r(x_0,y)\right] \\ -\beta\, \mathbb{D}_{\text{KL}}\left[\pi_\theta(x_0|y) \parallel \pi_{\text{ref}}(x_0|y)\right]\end{aligned} \tag{1}$$

Flow-DPO. Despite the success of RLHF, this approach still faces challenges such as high resource consumption and the difficulty in acquiring reliable reward models. Motivated by these challenges. DPO introduces a simple approach for policy optimization using human preferences directly. Compared with RLHF, DPO demonstrate a more pratical method for alignment with human preference, without explicit reward modeling. DPO objective allows the model to learn from both desirable (winner) data and undesirable (loser) data. Recent research shows that DPO has demonstrated success in AR models. Diffusion-DPO (Wallace et al., 2024) further demonstrated that aligning DPO with NAR models is also effective. The objective of Diffusion-DPO $\mathcal{L}_{\text{Diffusion-DPO}}(\theta)$ is defined by Eq.2

$$\begin{aligned}-\mathbb{E}\Big[\log\sigma\Big(-\frac{\beta}{2}\big(\|\epsilon^w-\epsilon_\theta(x_t^w,t)\|^2-\|\epsilon^w-\epsilon_{\text{ref}}(x_t^w,t)\|^2 \\ -\big(\|\epsilon^l-\epsilon_\theta(x_t^l,t)\|^2-\|\epsilon^l-\epsilon_{\text{ref}}(x_t^l,t)\|^2\big)\big)\Big)\Big]\end{aligned} \tag{2}$$

where $x_t^* = (1-t)x_0^* + t\epsilon, \epsilon \sim \mathcal{N}(0, I)$, The "*" indicates either "w" or "l". The expectation is computed over samples $\{x_0^w, x_0^l\} \sim D$ and the noise schedule parameter t.

Based on Diffusion-DPO, Flow-DPO (Liu et al., 2025) proved that in rectified flow, we can relate the noise vector ϵ^* to a velocity field v^* by Eq.3

$$\|\epsilon^* - \epsilon_{\text{pred}}(x_t^*, t)\|^2 = (1-t)^2\|v^* - v_{\text{pred}}(x_t^*, t)\|^2 \tag{3}$$

And the objective of Flow-DPO $\mathcal{L}_{\text{Flow-DPO}}(\theta)$ is given by Eq.4

$$\begin{aligned}-\mathbb{E}\Big[\log\sigma\Big(-\frac{\beta_t}{2}\big(\|v^w-v_\theta(x_t^w,t)\|^2-\|v^w-v_{\text{ref}}(x_t^w,t)\|^2 \\ -\big(\|v^l-v_\theta(x_t^l,t)\|^2-\|v^l-v_{\text{ref}}(x_t^l,t)\|^2\big)\big)\Big)\Big]\end{aligned} \tag{4}$$

Here, $\beta_t = \beta(1-t)^2$, and the expectation is computed over samples $\{x_0^w, x_0^l\} \sim D$ and the noise schedule parameter t. Intuitively, it is noteworthy that many experiments indicate better results when using a constant β_t. Therefore, in our experiments,

Table 1: Results on three test sets, LibriSpeech-PC *test-clean*, Seed-TTS *test-en* and Seed-TTS *test-zh*. The boldface indicates the best result.

Model	SIM-O↑	WER% ↓	UTMOS↑
LibriSpeech-PC *test-clean*			
F5-TTS	0.639	1.69	3.85
F5-TTS + SFT	0.650	**1.65**	**3.92**
F5D-TTS	**0.667**	1.75	3.80
Seed-TTS *test-en*			
F5-TTS	0.648	1.50	3.75
F5-TTS + SFT	0.661	**1.45**	**3.82**
F5D-TTS	**0.690**	1.62	3.70
Seed-TTS *test-zh*			
F5-TTS	0.747	1.79	2.96
F5-TTS + SFT	0.753	1.82	**2.99**
F5D-TTS	**0.760**	**1.75**	2.84

β_t is set as a constant, and its specific value will be discussed in Sec. 4.3.

Flow-DPO on TTS. As Flow-DPO aligns flow-matching models with preference by solving the RLHF objective (Eq. 1) analytically, we can optimize policy alignment with human preference via supervised training. We use the preference dataset collected in Sec. 3.2, minimizing $\mathcal{L}_{\text{Flow-DPO}}(\theta)$ means that the predict velocity field v_θ gets closer to the target velocity field of desirable speech v_w. By aligning flow-matching TTS models with preference (In this work, we use SIM-O as a reward signal), we can significantly improve the performance of base model.

4 Experiments

4.1 Experimental Setup

Datasets and Model. For datasets, as mentioned in Sec. 3.2, we use the F5-TTS model to generate 10,000 pairs, 20 hours of preference speech data, and we randomly select 1,000 pairs for model fine-tuning. We utilize three datasets for evaluation: LibriSpeech-PC (Meister et al., 2023), in this work, we use the 4-to-10-second sample test set based on LibriSpeech-PC, with 1,127 samples in the subset. Seed-TTS (Anastassiou et al., 2024) test-en with 1,088 samples from Common Voice (Ardila et al., 2019) and Seed-TTS test-zh with 2,020 samples from DiDiSpeech (Guo et al., 2021). For model, we use the F5-TTS model as base model.

Training and Inference. We fine-tune F5-TTS model on 8 NVIDIA A100 80G GPUs, with a batch size of 2 (pairs of data) and gradient accumulation of 4 steps. We use AdamW (Loshchilov and Hutter, 2017) and a learing rate of 1×10^{-7} is used without warmup. For the divergence penalty parameter β, we find the model demonstrates superior performance when $\beta \in [500, 10000]$. In the series of experiments, we set $\beta = 500$. During the inference phase, our setup remains largely consistent with F5-TTS, with one modification that we don't use the Exponential Moving Averaged (EMA) (Karras et al., 2024) weights.

Metrics. We evaluate speaker similarity using SIM-O, speech intelligibility using the word error rate (WER) and speech naturalness using mean opinion score (MOS). For SIM-O, we utilize WavLM-large-based (Chen et al., 2022) speaker verification model to calculate the cosine similarity score between the generated speech and the prompt speech. For WER, we utilize Whisper-large-v3 (Radford et al., 2023) to transcribe English speech and Paraformer-zh (Gao et al., 2023) to transcribe Chinese speech. The WER is calculated by comparing the transcribed text with the target text. For MOS, we use UTokyo-SaruLab MOS Prediction System (UTMOS) (Saeki et al., 2022) to automatically calculate the MOS of the generated speech.

4.2 Results

The main results are shown in Tab.1. We can see that F5D-TTS achieves better SIM-O scores, comparable to the F5-TTS base model and SFT model. The improvement of SIM-O indicates that aligning the NAR-TTS model with Flow-DPO is effective, with only a small number of data pairs. Though

Table 2: Ablation studies on data scales. Experiments on Seed-TTS test-en demonstrate that the SIM-O stabilizes upon reaching a data scale of 1,000 pairs.

Model	Dataset	Data scale	SIM-O↑
F5-TTS	Emilia	-	0.639
F5D-TTS	LibriTTS	100 pairs	0.681
		500 pairs	0.685
		1,000 pairs	**0.690**
		2,000 pairs	0.689
		5,000 pairs	0.688
		10,000 pairs	0.689

we observed reward hacking phenomena in WER and UTMOS metrics, experiments demonstrate that such effects can be confined within tolerable bounds, inducing negligible impact on synthesized speech quality. The comprehensive experimental results demonstrate that F5D-TTS achieves substantial improvements in speaker similarity performance under low-resource conditions (requiring only minimal training data (e.g., 1,000 samples) and reduced computational budgets), while maintaining speech intelligence and speech naturalness.

4.3 Ablation Studies

Ablation on β_t. As discussed in Sec 3.3, the parameter β_t controls the strength of the KL divergence constraint. Ablation studies with varying β_t demonstrate comparable performance within the range $\beta_t \in [500,\ 10000]$. The main results are shown in Fig.3.

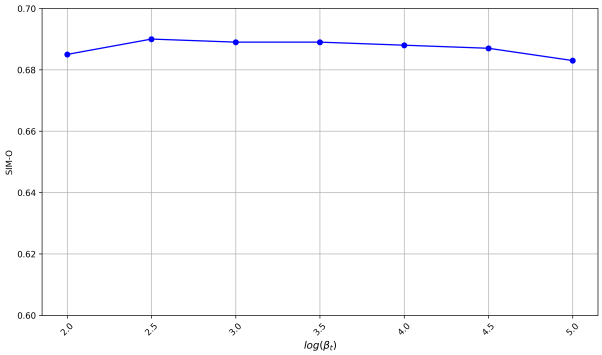

Figure 3: Ablation studies on β_t . Experiments on Seed-TTS test-en demonstrate that varying β_t demonstrate comparable performance within the range $\beta_t \in [500,\ 10000]$.

Ablation on data scale. To verify the impact of the data scale on SIM-O, we fine-tuned the model using randomly sampled subsets of varying sizes from the pre-constructed dataset. The results show that the model performance stabilizes when the data scale reaches 1,000 pairs. This experiment demonstrates that F5D-TTS significantly improves SIM-O performance with only a small amount of data. The main results are shown in Tab.2.

Ablation on training steps. To validate the optimal number of training steps, we fine-tuned the model using 1,000 data samples. The experimental results indicate the maximum performance of the model in 1 k training steps. Furthermore, the experiments reveal that reward hacking intensifies progressively with increasing training steps. The main results are shown in Fig.4.

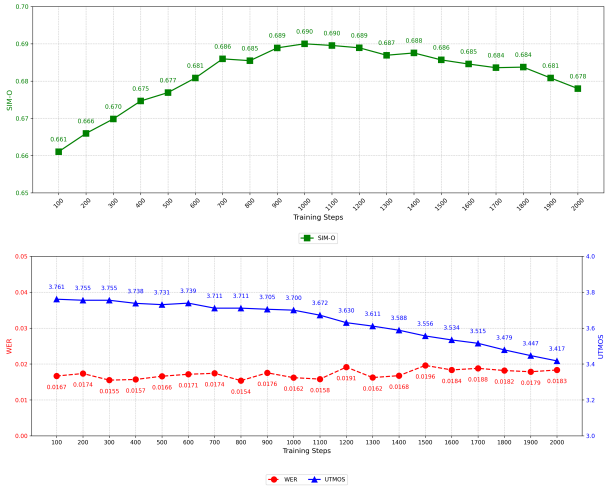

Figure 4: Ablation studies on training steps. The SIM-O metric attains its peak value of 0.690 at 1,000 training steps, followed by a progressive decline in subsequent iterations. Reward hacking intensifies with prolonged training, as evidenced by deteriorating UTMOS scores.

5 Conclusion

In this paper, we introduce F5D-TTS, a method to align flow-matching TTS models with human preferences by directly optimizing on human pref-

erence data. We utilized the F5-TTS model to construct 50,000 speech-text pairs from the LibriTTS dataset, and then generate speech five times with the same condition. Using the SIM-O metric for preference data construction, we ultimately selected 10,000 high-quality preference data pairs. Additionally, we use Flow-DPO to align flow-matching system to human preferences. Experiments show that F5D-TTS can significantly improve the speaker similarity while maintain the speech naturalness and speech intelligence using few-scale data and computationally efficient training.

Limitations

In this section, we discuss the limitations of this work and discuss potential directions for further research.

- **High-quality preference dataset required.** DPO relies on high-quality preference data, as low-quality preference data can negatively impact model performance (particularly in terms of speaker naturalness), which necessitates the screening of high-quality speech data during the data generation process.
- **Reward hacking.** Compared to F5R-TTS, though we get better speaker similarity (SIM-O) in a low-cost way, UTMOS score experiences a slight decline, making it challenging to balance all aspects of speech quality metrics simultaneously.

References

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, and 1 others. 2024. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024a. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024b. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. 2022. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151. IEEE.

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. 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)*, pages 1–5. IEEE.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and 1 others. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, and 1 others. 2021. Didispeech: A large scale mandarin speech corpus. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6968–6972. IEEE.

Shehzeen Hussain, Paarth Neekhara, Xuesong Yang, Edresson Casanova, Subhankar Ghosh, Mikyas T Desta, Roy Fejgin, Rafael Valle, and Jason Li. 2025. Koel-tts: Enhancing llm based speech generation with preference alignment and classifier free guidance. *arXiv preprint arXiv:2502.05236*.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, and 1 others. 2025. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng,

Kaitao Song, Siliang Tang, and 1 others. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. 2024. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184.

Keon Lee, Dong Won Kim, Jaehyeon Kim, and Jaewoong Cho. 2024. Ditto-tts: Efficient and scalable zero-shot text-to-speech with diffusion transformer. *arXiv preprint arXiv:2406.11427*.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, and 1 others. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, and 1 others. 2025. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. 2023. Librispeech-pc: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end asr models. In *2023 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 1–7. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Yakun Song, Zhuo Chen, Xiaofei Wang, Ziyang Ma, and Xie Chen. 2025. Ella-v: Stable neural codec language modeling with alignment-guided sequence reordering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25174–25182.

Xiaohui Sun, Ruitong Xiao, Jianye Mo, Bowen Wu, Qun Yu, and Baoxun Wang. 2025. F5r-tts: Improving flow matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*.

Jinchuan Tian, Chunlei Zhang, Jiatong Shi, Hao Zhang, Jianwei Yu, Shinji Watanabe, and Dong Yu. 2025. Preference alignment improves language model-based tts. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.