GLASS: GRPO-Trained LoRA for Acoustic Style Steering in Zero-Shot Text-to-Speech

Anonymous ACL submission

Abstract

We propose **GLASS**, a framework for composable acoustic style control in zero-shot autoregressive text-to-speech (TTS) that learns controls from post-generation rewards rather than style labels. For each control axis, GLASS freezes the TTS backbone and trains one lightweight LoRA adapter with Group Relative Policy Optimization (GRPO), using speech-token length and mean F_0 as style rewards and WER as an intelligibility anchor. Because each control is represented as a LoRA weight update, independently trained adapters can be swapped, interpolated, and composed through linear LoRA arithmetic. Experiments on speaking rate and pitch control show targeted style shifts while preserving naturalness, speaker similarity, and intelligibility.

1 Introduction

Modern zero-shot text-to-speech (TTS) models (Du et al., 2024; Anastassiou et al., 2024; Chen et al., 2024; Wang et al., 2023) can synthesize natural speech that closely matches the prompted speaker from a short reference prompt. However, the same prompt also implicitly specifies prosody and acoustic style attributes such as speaking rate and pitch. As a result, speaker identity and style are not independently addressable; changing the style of a speaker typically requires another prompt from the same speaker in the desired style. Furthermore, prompt selection offers only example-based control and provides no explicit mechanism for continuous interpolation or multi-axis composition.

A complementary line of controllable TTS methods conditions generation on style labels, acoustic attributes, or natural-language descriptions (Lou et al., 2024; Guo et al., 2023; Lyth and King, 2024). These controls are effective when the corresponding supervision is available, but adding new axes still requires annotations, reference examples, or curated data pipelines. Moreover, such controls are often tied to predefined categories rather than continuous, speaker-preserving manipulation.

In this paper, we propose **GLASS**, a reward-guided training framework for composable acoustic style control in zero-shot TTS without requiring style annotations. Given a fixed speaker prompt, GLASS aims to modify acoustic attributes while preserving speaker identity, intelligibility, and naturalness. Instead of relying on target-style examples or style labels, GLASS defines each control axis through post-generation rewards. Specifically, GLASS applies Group Relative Policy Optimization (GRPO) (Shao et al., 2024; DeepSeek-AI, 2025): given multiple samples from the same text and speaker prompt, GLASS scores speech-token length for speaking rate, mean F_0 for pitch, and WER for intelligibility, then updates the model using relative within-group rewards.

To make these learned controls reusable and composable, GLASS stores each control axis as a lightweight LoRA adapter on a frozen backbone. This keeps a shared backbone and one adapter per style direction; we show that independently trained controls can be swapped, interpolated, and composed at inference time via LoRA arithmetic (Ilharco et al., 2023; Huang et al., 2023; Shah et al., 2023). Training with multi-speaker prompts further encourages speaker-agnostic style directions that transfer to unseen prompts.

Our contributions are as follows:

- We propose **GLASS**, a framework for composable acoustic style control in zero-shot TTS that modifies style while keeping the speaker prompt fixed, without requiring style labels.
- We introduce a GRPO-based objective that converts post-generation acoustic and ASR measurements into speaker-preserving TTS style directions.
- We show that the resulting independently trained LoRA adapters can be swapped,

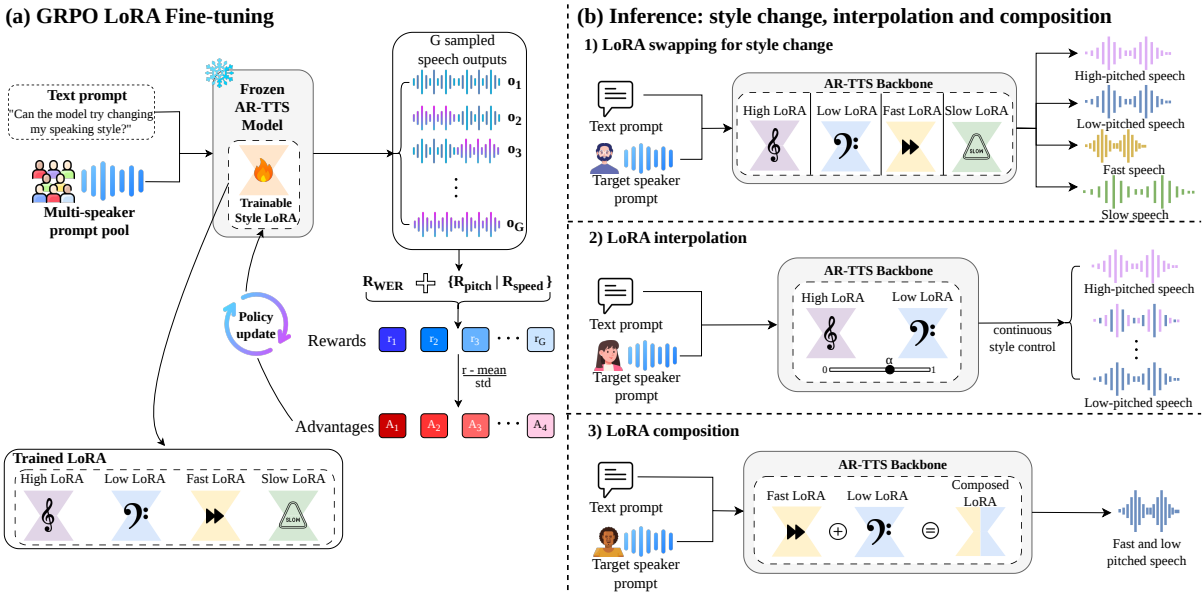

Figure 1: GLASS trains LoRA adapters with GRPO using WER and acoustic style rewards, then combines the learned adapter pool for swapping, interpolation, and composition while keeping the speaker prompt fixed.

smoothly interpolated, and composed across axes via LoRA arithmetic, while preserving intelligibility, speaker similarity, and naturalness better than the baseline.

2 Related Work

2.1 Style-controllable Text-to-Speech Models

Style-controllable TTS has been studied through attribute conditioning, latent or reference style representations, natural-language prompts, and zero-shot speech prompts (Ren et al., 2021; Łańcucki, 2021; Wang et al., 2018; Zhang et al., 2019; Guo et al., 2023; Yang et al., 2024; Wang et al., 2023; Du et al., 2024; Anastassiou et al., 2024; Chen et al., 2024). These approaches provide effective control, but generally require style evidence in the form of predefined attributes, textual descriptions, or reference speech. In contrast, GLASS learns style controls from measurable acoustic outcomes rather than from style annotations or reference examples.

2.2 Reinforcement Learning in Text-to-Speech Models

Preference optimization and reinforcement learning have recently been applied to TTS to optimize objectives beyond likelihood-based training. For example, Seed-TTS uses WER and speaker-similarity rewards with REINFORCE (Anastassiou et al., 2024), while subsequent GRPO-based TTS methods optimize ASR-, quality-, or prosody-related rewards (Sun et al., 2025; Liu et al., 2026; Zhong et al., 2025; Li et al., 2026). In parallel, Direct Preference Optimization (DPO) (Rafailov et al., 2023) has been used with pairwise preferences to improve emotional control or prosodic naturalness when reliable automatic prosody rewards are unavailable (Gao et al., 2025; Shin et al., 2026). These works primarily optimize the underlying generator or policy; GLASS is complementary, using GRPO to train lightweight LoRA adapters as modular style controls for swapping, interpolation, and composition.

Due to space constraints, we defer additional related work to Appendix E.

3 Method

3.1 Backbone and Adapters

We model autoregressive (AR) TTS as a speech-token policy $\pi(\mathbf{y} \mid \mathbf{x})$, where $\mathbf{x}$ contains text and a speaker prompt, and $\mathbf{y}$ is a sequence of discrete speech tokens. GLASS freezes the base TTS model and learns one LoRA adapter per style direction k.

3.2 GRPO Training

For each input $\mathbf{x}$, we sample a group of G completions from the same text and speaker prompt, and assign a reward r to each sample using an acoustic reward model R_k:

$$\mathbf{y}_1, \ldots, \mathbf{y}_G \sim \pi_{\theta_k}(\cdot \mid \mathbf{x}), \quad r_i = R_k(\mathbf{y}_i, \mathbf{x}). \quad (1)$$

Table 1: Individual style control on Seed-TTS-eval `test_en`. Syllables-per-second (SPS) and F_0 measure target styles; WER, SpkSim, and UTMOS, S-MOS, and N-MOS measure quality. Bold highlights GLASS results in pairwise comparisons.

	Style Metrics			Quality				
Method	SPS	F_0 M	F_0 F	WER↓	SpkSim↑	UTMOS↑	S-MOS↑	N-MOS↑
Baseline (CosyVoice2-0.5B)	3.65	120.4	192.2	2.81	0.655	3.28	–	–
Speed-axis control								
DSP speed-up	5.48	121.2	195.7	3.50	0.475	1.56	3.08	2.76
Fast LoRA (ours)	**5.59**	120.4	191.0	3.49	**0.617**	**3.30**	**4.72**	**4.68**
DSP slow-down	2.19	121.7	193.7	2.67	0.500	1.45	2.76	2.28
Slow LoRA (ours)	**2.30**	122.1	194.3	3.18	**0.650**	**3.05**	**4.56**	**4.24**
Pitch-axis control								
DSP pitch-up	3.65	150.9	239.9	2.59	0.173	1.57	1.40	2.28
High-pitch LoRA (ours)	3.61	**156.1**	**241.0**	3.01	**0.609**	**3.37**	**4.12**	**4.40**
DSP pitch-down	3.65	98.0	155.4	3.28	0.158	1.49	1.40	2.00
Low-pitch LoRA (ours)	3.68	**108.9**	**164.6**	3.11	**0.632**	**3.16**	**4.60**	**4.84**

We convert these rewards into group-relative advantages by normalizing them within the sampled group:

$$A_i = \frac{r_i - \mathrm{mean}_j(r_j)}{\mathrm{std}_j(r_j) + \epsilon}. \tag{2}$$

We then update only the LoRA parameters with a token-level PPO-clipped GRPO objective and a reference-policy penalty that keeps the adapted policy close to the frozen TTS backbone. The full token-level objective is given in Appendix B.

Because rewards are computed after waveform generation, the same loop can optimize non-differentiable signals such as ASR-based WER, generated token length, and estimated F_0.

3.3 Reward Design

Each reward combines a WER intelligibility anchor with a style term:

$$R_k(\mathbf{y}, \mathbf{x}) = \eta R_{\mathrm{WER}}(\mathbf{y}, \mathbf{x}) + (1-\eta) R_k^{\mathrm{style}}(\mathbf{y}), \tag{3}$$

where $R_{\mathrm{WER}} = 1 - \tanh(\gamma\, \mathrm{WER}(\mathbf{x}, \mathbf{y}))$ follows prior WER-based TTS rewards (Shin et al., 2026) and uses Whisper transcripts (Radford et al., 2023). For style, we min–max normalize the target statistic z_i within each GRPO group:

$$m(z_i) = \begin{cases} \dfrac{z_i - z_{\min}}{z_{\max} - z_{\min}}, & z_{\max} > z_{\min}, \\ 0.5, & \text{otherwise.} \end{cases} \tag{4}$$

Here $z_{\min}$ and $z_{\max}$ are group extrema, and 0.5 is used when all completions tie. Speed rewards use generated speech-token length, with fast and slow taking $1 - m(z_i)$ and $m(z_i)$, respectively; pitch rewards use utterance-level mean $F_{0,i}$, with high and low taking $m(z_i)$ and $1 - m(z_i)$, respectively.

3.4 LoRA Arithmetic

A LoRA adapter represents a low-rank update ΔW_k to frozen weights. Each adapter learns one signed acoustic direction while the speaker prompt and backbone remain fixed. Since all adapters share target modules and rank, their updates can be combined directly at inference time:

$$\Delta W(\mathbf{w}) = \sum_k w_k \Delta W_k. \tag{5}$$

The weights w_k are user-tunable controls, enabling same-axis interpolation such as $\Delta W(\alpha) = \alpha \Delta W_{\text{fast}} + (1-\alpha)\Delta W_{\text{slow}}$, and multi-axis composition such as $\Delta W = w_1 \Delta W_{\text{fast}} + w_2 \Delta W_{\text{high}}$ without retraining. We use w=0.5 as a stable composition point to avoid full-strength overshoot.

4 Experiments

4.1 Setup

We train one LoRA adapter per style direction on LibriTTS-R (Koizumi et al., 2023) using multi-speaker prompts. We evaluate zero-shot transfer on Seed-TTS-eval `test_en` (N=1088), reporting style-control metrics (syllables per second (SPS) and F_0) and quality metrics (WER, speaker similarity (SpkSim), and UTMOSv2), S-MOS, and N-MOS). This setup tests whether reward-trained LoRA adapters behave as speaker-agnostic style directions that transfer to unseen prompts, rather than memorizing training speakers. Baselines are the unmodified CosyVoice2-0.5B (Du et al., 2024) model and digital signal processing (DSP)-based transformations of its outputs: time-stretching for speed and pitch-shifting for pitch.

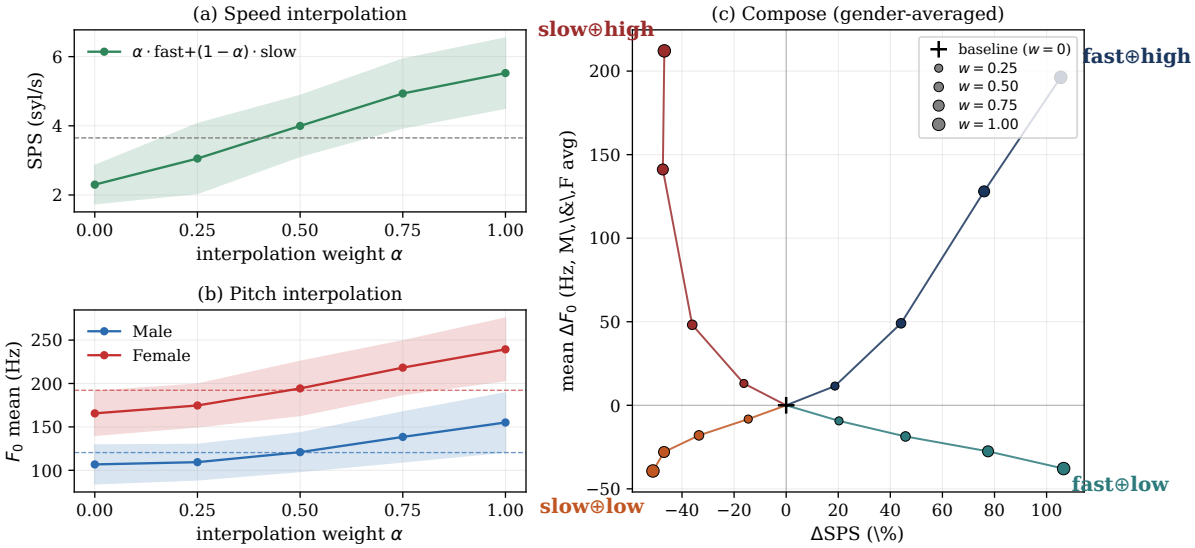

Figure 2: Continuous interpolation and gender-averaged multi-axis composition via linear LoRA arithmetic. (a,b) Opposite-adapter interpolation smoothly changes SPS and F_0. (c) Speed–pitch composition moves outputs toward intended quadrants; w=0.5 is stable before larger pitch overshoot.

In our implementation, LoRA is attached only to the autoregressive token generation module, yielding 1.08M trainable parameters, or 0.22% of the 495M-parameter AR backbone. Figure 1 summarizes the training and inference workflow. Additional details on implementation, metrics, and data subsets are provided in Appendix A.

4.2 Individual style control

Table 1 shows that LoRA adapters reach DSP-comparable target shifts (SPS, F_0) while keeping off-axis statistics near the baseline, indicating largely separable speed and pitch directions. In particular, DSP baselines collapse UTMOS to 1.45–1.57 and, for pitch shifting, drop SpkSim below 0.18, whereas LoRA stays within 0.23 UTMOS and 0.05 SpkSim of the baseline. S-MOS/N-MOS also favor LoRA over DSP (4.12–4.72/4.24–4.84 vs. 1.40–3.08/2.00–2.76). WER stays within 0.7 pp of the baseline across all adapters.

4.3 Continuous interpolation

We linearly blend opposite adapters with $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$ on a balanced subset. Figure 2(a,b) shows smooth control: SPS shifts from 2.30 to 5.52, while male/female F_0 moves from 107/166 Hz to 155/239 Hz. This demonstrates that interpolation between opposite LoRA adapters provides continuous control over the target acoustic style; Appendix C reports utterance-level monotonicity and quality across the sweep.

4.4 Multi-axis composition

Finally, we compose one speed adapter and one pitch adapter across four combinations: fast⊕high, fast⊕low, slow⊕high, and slow⊕low. At w_A=w_B=0.5, all four combinations move toward their intended speed–pitch quadrant (Figure 2(c)), with each axis retaining 80–121% of its single-axis effect (Appendix D, Table 3). Appendix D also provides gender-specific views and shows why full-strength composition overshoots. The result demonstrates that independently reward-trained adapters can be reused as composable acoustic style directions.

5 Conclusion

We presented GLASS, a reward-guided framework for modular acoustic style control in AR-TTS. By freezing the TTS backbone and training one lightweight LoRA adapter per style direction with GRPO, GLASS learns speaker-agnostic speed and pitch controls from post-generation rewards rather than style-labeled speech. Experiments show that the resulting adapters modify speaking rate and pitch while preserving intelligibility, speaker similarity, and predicted naturalness better than DSP baselines. Since styles are encoded as LoRA weight updates, the same adapters support swapping, smooth interpolation, and multi-axis composition, suggesting a practical path toward extensible and composable style control for zero-shot TTS.

Limitations

This work focuses on two measurable acoustic style axes, speaking rate and pitch. These axes are interpretable and useful, but they do not cover richer prosodic or paralinguistic attributes such as emotion, speaking effort, accent, or conversational style. Extending GLASS to such attributes would require reliable automatic rewards or human preference feedback.

Our rewards and evaluations rely on automatic proxy measurements: speech-token length for speaking rate, voiced-frame mean F_0 for pitch, Whisper-based WER for intelligibility, and automatic speaker-similarity and naturalness metrics. These proxies make the method reproducible and scalable, but they may not capture all perceptual aspects of style transfer.

References

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

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 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.

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

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. 2024. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer.

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.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, and 1 others. 2023. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36:15890–15902.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Promptttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2023. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *International Conference on Learning Representations*.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802*.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

Yinghao Aaron Li, Xilin Jiang, Fei Tao, Cheng Niu, Kaifeng Xu, Juntong Song, and Nima Mesgarani. 2026. Dmospeech 2: Reinforcement learning for duration prediction in metric-optimized speech synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):31814–31822.

Chang Liu, Ya-Jun Hu, Ying-Ying Gao, Shi-Lei Zhang, and Zhen-Hua Ling. 2026. Group relative policy optimization for text-to-speech with large language models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16132–16136. IEEE.

Haowei Lou, Helen Paik, Wen Hu, and Lina Yao. 2024. Stylespeech: Parameter-efficient fine tuning for pre-trained controllable text-to-speech. *arXiv preprint arXiv:2408.14713*.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.

Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. 2024. Orthogonal adaptation for modular customization of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7964–7973. IEEE Computer Society.

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.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

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*.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. 2023. Ziplora: Any subject in any style by effectively merging loras. *arXiv preprint arXiv:2311.13600*.

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

Seungyoun Shin, Dongha Ahn, Jiwoo Kim, and Sungwook Jeon. 2026. No verifiable reward for prosody: Toward preference-guided prosody learning in tts. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2256–2260. IEEE.

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*.

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*.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, pages 5180–5189. PMLR.

Dongchao Yang and 1 others. 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *arXiv preprint arXiv:2301.13662*.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE.

Yicheng Zhong, Peiji Yang, and Zhisheng Wang. 2025. Multi-reward grpo for stable and prosodic single-codebook tts llms at scale. *arXiv preprint arXiv:2511.21270*.

A Implementation and Evaluation Details

Backbone and LoRA. The backbone is CosyVoice2-0.5B. LoRA is attached only to the Qwen2 autoregressive token model, while speech embeddings, token decoder, flow-matching acoustic model, and vocoder remain frozen. We target $\{q_{\text{proj}}, v_{\text{proj}}\}$ attention projections with rank r=16, scaling α=32, and dropout 0.05.

Training data and optimization. For each style adapter, we sample prompts from 50 randomly selected LibriTTS-R (Koizumi et al., 2023) `train-clean-100` speakers (25 male, 25 female; seed 42), each with at least 20 utterances. Target texts are drawn from a 3,000-sentence pool from the same corpus. For every batch item, we sample a fresh speaker uniformly and keep that speaker fixed across the G=8 generations in the GRPO group, making within-group reward normalization comparable under the same text and speaker condition. This multi-speaker prompt training encourages each LoRA to capture a transferable style direction rather than a speaker-specific voice transformation.

Each adapter is trained for 500–750 update steps with AdamW, batch size 4, group size G=8, two PPO epochs, ε=0.2, β=0.01, η=0.5, and γ=1. Whisper-large-v3 (Radford et al., 2023) provides transcripts for R_{WER}; token length is computed from generated speech tokens, and pitch rewards use voiced-frame F_0 estimated with `pyworld`.

Evaluation. Seed-TTS-eval `test_en` contains 1,088 prompt–text pairs from Common Voice and is out-of-domain with respect to the LibriTTS-R training data. This setting tests whether LoRA style adapters trained with LibriTTS-R speakers transfer to unseen speaker prompts at inference. For interpolation sweeps, we use a gender-balanced 200-utterance subset (100 male, 100 female) to keep inference cost tractable.

We report SPS, mean voiced-frame F_0, WER, speaker similarity, UTMOSv2 (Saeki et al., 2022), S-MOS, and N-MOS. SpkSim is the cosine similarity between prompt and generated speech embeddings extracted by a WavLM-large speaker-verification model (Chen et al., 2022), following the Seed-TTS evaluation protocol. For MOS, 15 human raters scored 25 utterances per transformed system on 1–5 S-MOS/N-MOS scales for prompt speaker similarity and naturalness; each utterance received five ratings, and Table 1 reports means. Following Spark-TTS (Wang et al., 2025), SPS denotes syllables per second, computed as the target-text syllable count divided by the generated waveform duration in seconds; higher SPS indicates faster speaking rate. For DSP baselines, we apply `librosa` time-stretching with rates 1.5 and 0.6 for speed and pitch shifts of ± 4 semitones for pitch.

B Token-Level GRPO Objective

For each prompt, GLASS samples G completions and computes scalar rewards r_i after waveform generation. The group-relative advantage used for completion i is

$$A_i = \frac{r_i - \mu_r}{\sigma_r + \epsilon_{\text{adv}}}, \qquad \mu_r = \frac{1}{G}\sum_{j=1}^{G} r_j, \tag{6}$$

where σ_r is the within-group reward standard deviation. This normalization makes the update depend on relative quality among samples from the same text and speaker prompt, rather than on the absolute reward scale.

Let $\ell_{\theta,i,t} = \log \pi_{\theta_k}(y_{i,t} \mid \mathbf{x}, \mathbf{y}_{i,<t})$ be the current token log-probability for generated token t in completion i. We define $\ell_{\text{old},i,t}$ analogously for the sampling policy and $\ell_{0,i,t}$ for the frozen reference policy. In implementation, π_0 is obtained by disabling LoRA layers in the same model, avoiding a separate reference copy. The token-level likelihood ratio and its clipped version are

$$\begin{aligned} \rho_{i,t} &= \exp(\ell_{\theta,i,t} - \ell_{\text{old},i,t}), \\ \bar{\rho}_{i,t} &= \text{clip}(\rho_{i,t}, 1-\varepsilon, 1+\varepsilon). \end{aligned} \tag{7}$$

The sequence-level advantage A_i is broadcast to all generated tokens, and losses are averaged within each generated sequence before averaging across completions. With $\Delta_{i,t} = \ell_{0,i,t} - \ell_{\theta,i,t}$ and generated length T_i, the objective minimized for each LoRA adapter is

$$\begin{aligned} \mathcal{L}_{\text{GRPO}} = &\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i} \\ &\Big[-\min(\rho_{i,t}A_i, \bar{\rho}_{i,t}A_i) \\ &\quad + \beta\left(e^{\Delta_{i,t}} - \Delta_{i,t} - 1\right)\Big]. \end{aligned} \tag{8}$$

The first term is the PPO-clipped policy-gradient loss, while the second term penalizes drift from the frozen backbone at the token level. Only LoRA parameters receive gradients; all backbone, acoustic decoder, and vocoder parameters remain frozen.

Table 2: Continuous interpolation between opposite LoRA adapters on the 200-utterance subset. The target style metric (α vs. SPS for speed; vs. F_0 for pitch) varies smoothly while off-axis metrics remain near baseline. WER reaches its minimum at the central blend α=0.5 on both axes.

	Speed axis (α·fast+(1−α)·slow)					*Pitch axis* (α·high+(1−α)·low)				
α	SPS	WER% ↓	SpkSim ↑	UTMOS ↑	F_0 (M/F)	SPS	WER% ↓	SpkSim ↑	UTMOS ↑	F_0 (M/F)
0.00	2.30	4.50	0.647	3.00	123/196	3.72	3.36	0.629	3.16	107/166
0.25	3.05	4.25	0.649	3.18	121/193	3.71	2.60	0.644	3.19	109/175
0.50	4.00	**2.16**	0.645	3.29	121/194	3.64	**2.14**	0.644	3.28	121/194
0.75	4.93	2.91	0.624	3.32	121/193	3.64	3.02	0.634	3.34	138/218
1.00	5.52	3.51	0.614	3.31	119/192	3.62	2.96	0.614	3.35	155/239

Table 3: Per-axis retention at w_A=w_B=0.5. Both axes recover 80–121% of the single-axis effect, indicating near-additive composition.

Combination	ret_{SPS}	$\text{ret}_{F_0,\text{M}}$	$\text{ret}_{F_0,\text{F}}$
fast⊕high	83%	121%	114%
fast⊕low	86%	90%	91%
slow⊕high	98%	109%	117%
slow⊕low	91%	80%	82%

Table 4: Composition at w_A=w_B=1.0. Pitch overshoots to non-physiological registers when paired with `high` (male F_0 up to 5.6× the single-axis shift); voicing ratio drops sharply when paired with `low`, indicating creak/aperiodic phonation.

Combination	SPS	F_0(M)	F_0(F)	voicing
fast⊕high	7.50	318.9	385.5	0.64
fast⊕low	7.54	106.2	147.0	0.37
slow⊕high	1.94	326.3	392.0	0.58
slow⊕low	1.78	106.8	136.1	0.32
Baseline	3.65	120.4	192.2	0.65

C Continuous Interpolation Details

We linearly blend opposite adapters as $\Delta W(\alpha) = \alpha\,\Delta W_+ + (1-\alpha)\,\Delta W_-$ with $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$, where $+$ denotes the fast or high-pitch adapter and $-$ denotes its opposite. For tractability, we evaluate on a gender-balanced 200-utterance subset of Seed-TTS-eval `test_en` (100 M, 100 F; seed 42). The same prompts and texts are used across all α, so trends are within-utterance.

Monotonicity. We compute per-utterance Spearman ρ between α and the target acoustic statistic (SPS for speed; mean voiced-frame F_0 for pitch, restricted to utterances with voicing ratio > 0.3), then average across utterances. The averaged correlations are strongly positive: 0.954 ± 0.088 for speed, 0.874 ± 0.154 for male pitch, and 0.923 ± 0.099 for female pitch. Thus, the target statistic usually increases with α at the utterance level; the curves in Figure 2(a,b) are not artifacts of averaging.

Quality across α. Table 2 reports the full sweep. WER follows a U-shape with a minimum at α=0.5 on both axes (2.16% speed, 2.14% pitch), below both endpoints. UTMOSv2 remains within 0.35 of baseline, indicating that interpolation does not degrade predicted naturalness. Speaker similarity is essentially flat for $\alpha \leq 0.5$ and drops modestly at the extremes (0.61–0.62 at α=1.0). Overall, intermediate weights are not a degenerate average of two adapters but a stable operating regime.

D Composition Retention and Overshoot

We compose one speed adapter and one pitch adapter via linear LoRA arithmetic $\Delta W = w_A\,\Delta W_A + w_B\,\Delta W_B$ with w_A=w_B=w for $w \in \{0.25, 0.5, 0.75, 1.0\}$, evaluated on the full Seed-TTS-eval `test_en` (N=1088).

Retention at w=0.5. To quantify whether the composed adapter preserves each axis's effect, we define

$$\text{ret}_X = \frac{\Delta X_{\text{compose}(w=0.5)}}{\Delta X_{\text{single-axis LoRA}}} \times 100\%,$$

where ΔX is the signed change in target statistic X (SPS % or F_0 in Hz) relative to the baseline. Table 3 shows that at w=0.5, both axes retain 80–121% of their respective single-axis effect. Retention is consistently close to and occasionally above 100%, suggesting that the two adapters act on largely separable directions in token-policy space rather than canceling each other. Mild overshoot (>100%) appears on the pitch axis when paired with `high`, which may reflect longer high-pitched intervals contributing more voiced frames to the F_0 estimate.

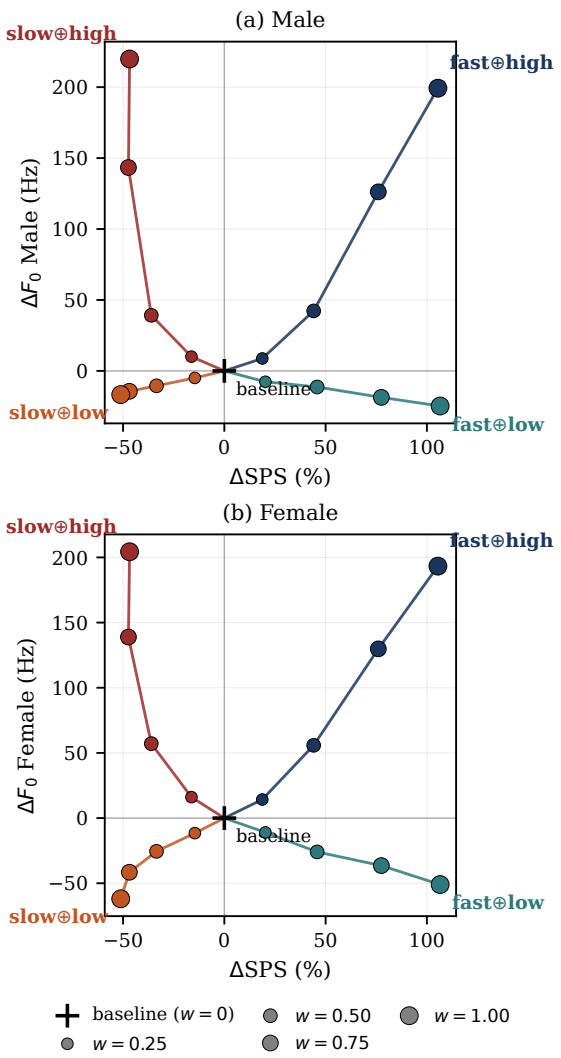

Figure 3: Gender-specific multi-axis speed–pitch composition results corresponding to the averaged view in Figure 2(c). Both male and female subsets move toward the intended quadrants as w increases.

Why w=0.5, not w=1.0. While naive intuition suggests using each adapter at its trained strength (w=1.0), summing two adapters at full strength produces pronounced overshoot, especially on the pitch axis. Table 4 reports the w=1.0+1.0 regime: fast⊕high drives male mean F_0 from 120 Hz at baseline to 319 Hz—roughly 5.6× the shift produced by the single-axis high adapter (35.7 Hz), entering an unrealistically high register. Female mean F_0 similarly reaches 386 Hz. The ∗⊕low combinations show a different failure mode: mean voicing ratio collapses to 0.32–0.37 (baseline ≈ 0.65), indicating creaky or aperiodic phonation. We therefore adopt w=0.5 as the default composition operating point in the main paper, where retention is near-additive without the severe overshoot observed at full strength.

E Additional Related Work

LoRA provides a parameter-efficient way to adapt large pretrained models by learning low-rank weight updates while keeping the base model frozen (Hu et al., 2022). Beyond its use as an efficient fine-tuning method, recent work has shown that weight updates can be treated as modular control directions. Task arithmetic represents the difference between a pretrained model and a fine-tuned model as a task vector and composes such vectors in weight space (Ilharco et al., 2023). LoRAHub similarly combines multiple LoRA modules at inference time to adapt to unseen tasks without updating the base model (Huang et al., 2023).

This adapter-arithmetic view has been explored extensively in text-to-image diffusion models, where independently trained LoRAs are merged or composed for subject, style, and multi-concept control. ZipLoRA merges subject and style LoRAs to balance subject fidelity and style fidelity (Shah et al., 2023); Mix-of-Show fuses multiple concept LoRAs while reducing concept conflicts (Gu et al., 2023); Orthogonal Adaptation encourages independently trained customization updates to be orthogonal (Po et al., 2024); and Concept Sliders learn interpretable low-rank directions that can be continuously modulated and composed (Gandikota et al., 2024). Together, these studies show that low-rank weight updates can act as reusable control handles for generative models.

GLASS brings this adapter-arithmetic perspective to acoustic style control in autoregressive zero-shot TTS. Instead of training adapters from subject or style examples, GLASS learns speaker-agnostic LoRA directions from post-generation acoustic rewards such as speaking rate and mean F_0. The resulting adapter library supports discrete swapping, continuous interpolation between opposite directions, and multi-axis composition while preserving the speaker prompt.