

RLAIF-SPA: Structured AI Feedback for Semantic-Prosodic Alignment in Speech Synthesis

Qing Yang¹, Zhenghao Liu^{1†}, Yangfan Du¹, Pengcheng Huang¹, Tong Xiao^{1†}

¹School of Computer Science and Engineering, Northeastern University, China

Abstract—Recent advances in Text-To-Speech (TTS) synthesis have achieved near-human speech quality in neutral speaking styles. However, most existing approaches either depend on costly emotion annotations or optimize surrogate objectives that fail to adequately capture perceptual emotional quality. As a result, the generated speech, while semantically accurate, often lacks expressive and emotionally rich characteristics. To address these limitations, we propose RLAIF-SPA, a novel framework that integrates Reinforcement Learning from AI Feedback (RLAIF) to directly optimize both emotional expressiveness and intelligibility without human supervision. Specifically, RLAIF-SPA incorporates Automatic Speech Recognition (ASR) to provide semantic accuracy feedback, while leveraging structured reward modeling to evaluate prosodic-emotional consistency. RLAIF-SPA enables more precise and nuanced control over expressive speech generation along four structured evaluation dimensions: Structure, Emotion, Speed, and Tone. Extensive experiments on LibriSpeech, MELD, and Mandarin ESD datasets demonstrate consistent gains across clean read speech, conversational dialogue, and emotional speech. On LibriSpeech, RLAIF-SPA consistently outperforms Chat-TTS, achieving a 26.1% reduction in word error rate, a 9.1% improvement in SIM-O, and over 10% gains in human subjective evaluations.

Index Terms—Emotional Speech Synthesis, Reinforcement Learning, AI Feedback

I. INTRODUCTION

Recent advances in Large Language Models (LLMs) have enabled Text-To-Speech (TTS) systems to achieve near-human naturalness and intelligibility [1]–[3]. Nevertheless, human interaction is not solely about informational exchange but also involves the subtle expression of emotion. These emotional signals play a crucial role in shaping listener engagement and comprehension, particularly in applications such as conversational agents, audiobooks, and virtual assistants. In these settings, insufficient emotional expressiveness often leads to speech that sounds monotonous or flat, thereby reducing its effectiveness in accurately conveying human emotions [4]. Thus, developing TTS models with controllable emotional capabilities is essential for bridging this gap and has become a major focus in the field.

Despite this progress, emotional TTS still faces a core bottleneck: scalable supervision for controllable expressiveness while preserving content fidelity. Label-conditioned methods offer an intuitive control interface, but coarse emotion categories fail to capture fine-grained prosody and scaling typically relies on costly human annotations. Preference-based reinforcement learning provides another direction by optimizing

perceptual quality, yet a single holistic score offers weak credit assignment and does not reveal which prosodic attributes are misaligned, making targeted control difficult. Meanwhile, improving expressiveness often degrades intelligibility when content fidelity is not explicitly constrained. These challenges call for a framework that provides structured, attribute-level feedback at scale and jointly optimizes expressiveness and content fidelity.

We view emotional TTS post-training as a multi-attribute alignment problem where holistic preference scores provide weak credit assignment, motivating diagnostic AI feedback at the attribute level. To address these challenges, we propose **RLAIF-SPA**, a framework that incorporates Reinforcement Learning from AI Feedback (RLAIF) [5] for Semantic-Prosodic Alignment in Text-To-Speech synthesis. Unlike prior RLHF-style post-training that optimizes a holistic preference, RLAIF-SPA introduces a structured AI-feedback interface that decomposes alignment into semantic fidelity and attribute-wise prosodic alignment, enabling joint optimization of expressiveness and accuracy with automatically computed rewards. RLAIF-SPA leverages two core feedback components. For expressiveness, *Prosodic Label Alignment* judges the model output against automatically generated labels along four fine-grained dimensions: *Structure*, *Emotion*, *Speed*, and *Tone* [6]. For clarity, *Semantic Accuracy Feedback* assesses consistency between the transcribed output and the original input text. By combining these two signals, the AI Feedback mechanism provides a stable and scalable optimization target.

Experiments on LibriSpeech, MELD, and ESD show that RLAIF-SPA achieves a better trade-off between accuracy and expressiveness than strong baselines, reducing word error rate while improving speaker similarity and emotion alignment. Beyond improved metrics, our key contribution is a structured AI-feedback interface that decomposes emotional TTS post-training into attribute-wise prosodic alignment and semantic accuracy supervision, enabling diagnostic credit assignment and preventing expressiveness gains from degrading content fidelity. Our code is released in an repository: <https://github.com/Zoe-Mango/RLAIF-SPA>.

II. RELATED WORK

Early emotional TTS systems typically condition synthesis on explicit, coarse-grained emotion labels [7]–[9]. By associating discrete categories with acoustic and prosodic patterns, these methods modulate speech attributes to elicit the desired affect [10]. Despite their simplicity and controllability,

[†]Corresponding author.

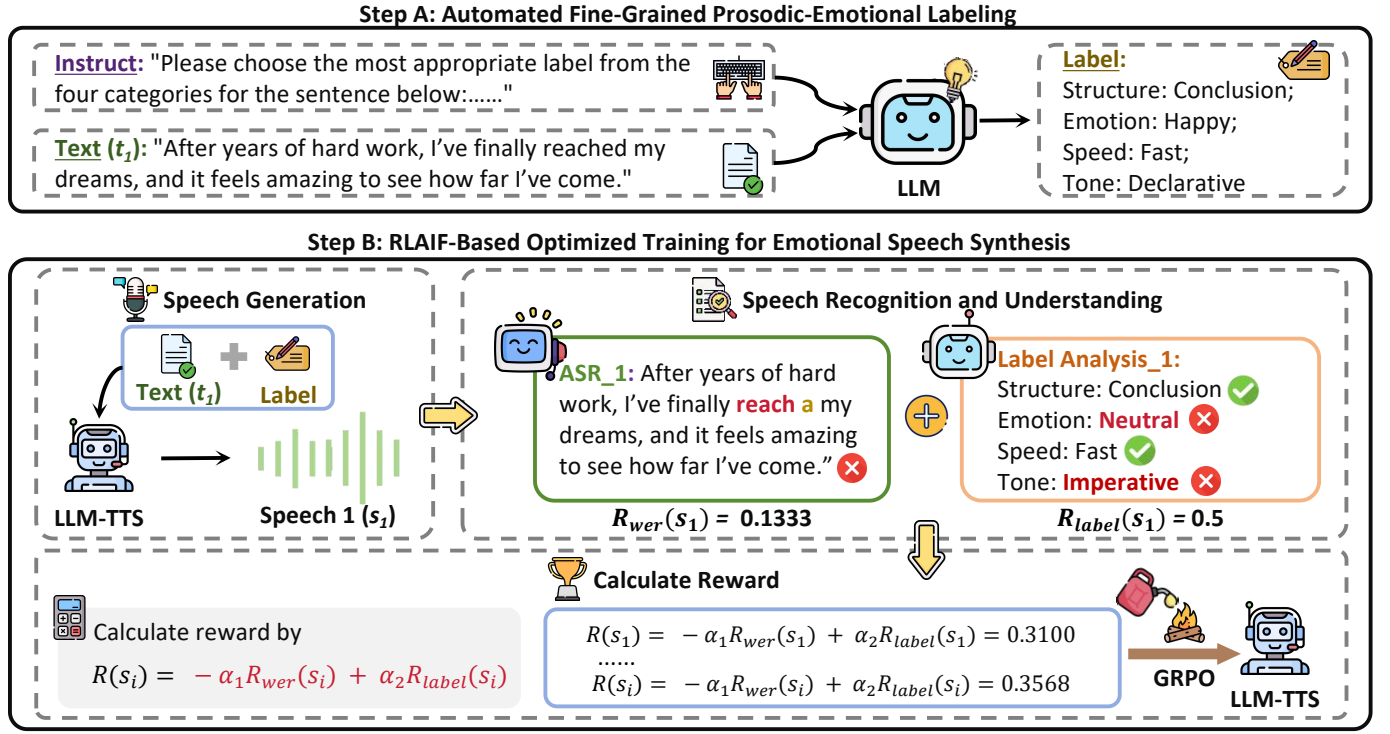


Fig. 1: Illustration of the proposed RLAIF-SPA framework. RLAIF-SPA optimizes both emotional expressiveness and accuracy through the AI Feedback mechanism, which is generated by AI models themselves.

this paradigm faces two inherent limitations. First, discrete labels oversimplify the continuous and compositional nature of human affect, making it difficult to represent subtle intensity changes or mixed emotions. Second, learning high-quality emotional control at scale typically relies on extensive manual annotations, rendering large-scale corpus construction labor-intensive and costly and thus limiting scalability [11]–[13].

To reduce dependence on categorical emotion labels and manual annotation, recent studies have explored Reinforcement Learning (RL) paradigms that optimize perceptual objectives from human or model-based feedback [14]–[16]. Leveraging pairwise preference annotations, these approaches can model human judgments of emotional expressiveness and produce more nuanced and diverse speech [17]–[19]. However, existing RLHF-based TTS methods predominantly rely on holistic preference signals. In practice, optimizing a single scalar score entangles multiple factors and provides limited diagnostic guidance on which prosodic dimensions should be adjusted [20]. As a result, these models often exhibit limited controllability and struggle to support precise, targeted optimization for emotional speech synthesis, especially when expressiveness improvements risk degrading semantic fidelity.

Expressive speech is inherently multi-dimensional and is therefore better modeled through complementary prosodic attributes [21]. Speaking rate and pitch variation correlate strongly with arousal and subtle affective cues [22], [23], while higher-level structures such as discourse rhythm contribute to contextual coherence [24], [25]. These findings motivate decomposing emotional alignment into attribute-wise super-

vision rather than relying on a single overall preference. In parallel, group-relative formulations such as Group Relative Policy Optimization (GRPO) compare multiple candidates for the same input [26], [27], yielding a more stable learning signal than noisy absolute scoring. RLAIF-SPA builds on these insights by introducing structured, attribute-level AI feedback and jointly constraining semantic fidelity, enabling scalable and controllable emotional TTS without requiring human preference labels.

III. METHODOLOGY

This section presents the proposed RLAIF-SPA framework. We first describe the overall training strategy and optimization objectives, and then introduce the two core components of the AI feedback mechanism, Prosodic Label Alignment and Semantic Accuracy Feedback, as illustrated in Fig. 1.

A. Training Strategy and Optimization Framework

The training of RLAIF-SPA is formulated as a multi-objective optimization problem guided by feedback provided by an automated evaluator. Specifically, the feedback signals correspond to two key objectives: emotional expressiveness and semantic accuracy. The optimization aims to improve prosodic-emotional label alignment while simultaneously reducing the Word Error Rate (WER), thereby encouraging the model to generate speech that is both emotionally expressive and accurate.

Given an input text t_i , we first construct a target prosodic label vector y_i with four dimensions, automatically generated by a LLM. The policy model π_θ then produces a speech

sample $s_i \sim \pi_\theta(\cdot | t_i)$, which is scored by a composite reward function $R(s_i)$ capturing two complementary objectives: (i) prosodic-emotional alignment to y_i , evaluated by an external speech understanding model, and (ii) semantic accuracy with respect to t_i , quantified using a transcription produced by an Automatic Speech Recognition (ASR) system. These signals are combined through a weighted linear form:

$$R(s_i) = -\alpha_1 R_{\text{wer}}(s_i) + \alpha_2 R_{\text{label}}(s_i), \quad (1)$$

where $R_{\text{label}}(s_i)$ denotes the reward for prosodic-emotional alignment (Section III-B), and $R_{\text{wer}}(s_i)$ is a penalty term based on the WER of the generated speech (Section III-C). The non-negative hyperparameters α_1 and α_2 control the trade-off between accuracy and emotional expressiveness.

To optimize the policy model, we adopt Group Relative Policy Optimization (GRPO) [26], which evaluates the relative quality of multiple candidates within a group of generated outputs rather than scoring each output independently. Formally, within the GRPO framework, for a given input text t_i , the policy π_θ generates a group of G candidate speech outputs $\{s_{i,g}\}_{g=1}^G$. The model's parameters θ are then updated to maximize the following objective function:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{t_i \sim D, \{s_{i,g}\}_{g=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | t_i)} \left[\frac{1}{G} \sum_{g=1}^G L_{i,g}(\theta) \right] - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\theta_{\text{old}}}), \quad (2)$$

where $L_{i,g}(\theta) = A_{i,g}^G \log \pi_\theta(s_{i,g} | t_i)$ is an advantage-weighted log-likelihood surrogate objective for the g -th candidate output of the i -th input. Here $A_{i,g}^G$ measures the performance of $s_{i,g}$ relative to the average quality of the group, so candidates with higher-than-average reward receive positive advantages and are assigned higher probability under π_θ . The term $D_{\text{KL}}(\pi_\theta \| \pi_{\theta_{\text{old}}})$ is a KL divergence penalty that regularizes the policy update, preventing large deviations from the previous policy $\pi_{\theta_{\text{old}}}$ and ensuring training stability. The hyperparameter β controls the strength of this regularization.

B. Prosodic Label Alignment

The first component of the AI Feedback mechanism is prosodic-emotional label matching. To guide the model toward generating emotionally expressive speech, we adopt a fine-grained labeling strategy that annotates speech along four distinct prosodic-emotional dimensions: *Structure*, *Emotion*, *Speed*, and *Tone*. Here, *Structure* captures the utterance-level discourse function, *Emotion* is a five-way category, *Speed* denotes speaking rate, and *Tone* describes pragmatic speaking style. These four labels represent key and complementary aspects of emotional expression, thereby forming a comprehensive yet manageable framework [28].

To implement this strategy at scale, we first generate target prosodic-emotional labels for the entire training dataset using an LLM [29], and then optimize the policy with these labels as supervision. We define the match indicator as:

$$m_k(s_i) = \mathbb{I}[\hat{y}_{i,k}(s_i) = y_{i,k}], \quad k \in \mathcal{K}, \quad (3)$$

where $\mathcal{K} = \{\text{Structure}, \text{Emotion}, \text{Speed}, \text{Tone}\}$ denotes the evaluation fields, each field $k \in \mathcal{K}$ has a discrete label set $\mathcal{Y}_k$, $y_{i,k} \in \mathcal{Y}_k$ is the target label for sample i , and $\hat{y}_{i,k}(s_i) \in \mathcal{Y}_k$ is the predicted label given speech s_i .

The reward function for this prosodic-emotional label alignment is formally defined as:

$$R_{\text{label}}(s_i) = \sum_{k \in \mathcal{K}} w_k \cdot m_k(s_i), \quad (4)$$

where $m_k(s_i)$ is a binary match indicator for each field, and w_k is the corresponding weight for that field.

C. Semantic Accuracy Feedback

The second key component of the AI Feedback mechanism is designed to ensure speech accuracy. This is achieved by quantifying the semantic accuracy of the synthesized speech. Let t_i be the original input text and s_i be a generated speech sample. We employ an ASR model to obtain the transcription $\text{ASR}(s_i)$.

The accuracy-driven reward component, R_{wer} , is then formulated as the WER computed between the original text t_i and the transcribed text $\text{ASR}(s_i)$:

$$R_{\text{wer}}(s_i) = \text{WER}(t_i, \text{ASR}(s_i)), \quad (5)$$

where $\text{WER}(\cdot)$ calculates the standard word error rate between the two input texts. This value serves as a direct cost within our composite reward function (Eq. 1), effectively penalizing any deviation from the source text. By integrating this semantic accuracy penalty, the framework ensures that any improvements in emotional expressiveness do not come at the expense of clarity or content accuracy, leading to speech that is both articulate and emotionally resonant.

IV. EXPERIMENTAL METHODOLOGY

In this section, we describe the datasets, baselines, evaluation metrics, and implementation details of our experiments.

Datasets. During training, we use a subset of 1,000 utterances from the LibriSpeech train-clean set [30]. We annotate each transcript with the same four structured attributes using GPT-4o following the labeling strategy in Sec. III-B to ensure consistency between training supervision and evaluation. Since LibriSpeech has no emotion labels, all attribute labels are inferred from text by GPT-4o. Instead of uniform sampling, we select a more expressive subset by ranking utterances with an LLM-derived expressiveness score from the same labeler and keeping the top 1,000.

To ensure a comprehensive evaluation, we employ three datasets. For LibriSpeech test-clean, we annotate its transcripts with the same four attributes using GPT-4o to ensure consistency with the training supervision. MELD [31], derived from the Friends TV series, provides a challenging conversational English setting characterized by complex and dynamic emotional interactions. ESD [32] is used for Mandarin evaluation and contains parallel emotional speech from native speakers across five emotion categories.

TABLE I: Objective and Subjective Evaluation Results Comparison of RLAIF-SPA with Baselines on WER, SIM-O, CMOS, Emotion MOS, and Speech Emotion Recognition across three datasets. **Bold** indicates the best result in each column, while underlining indicates the second best.

Model	Objective		Subjective		Speech Emotion Recognition					
	WER↓	SIM-O↑	CMOS↑	Emotion MOS↑	Neutral↑	Happy↑	Sad↑	Angry↑	Surprise↑	Avg↑
<i>LibriSpeech test-clean (En)</i>										
Chat-TTS	7.85	0.66	5.99 ± 0.56	5.75 ± 0.41	76.43	9.84	37.31	1.47	0.00	25.01
F5-TTS	4.87	0.70	6.97 ± 0.50	5.93 ± 0.36	80.85	15.03	5.47	7.35	2.94	22.33
MegaTTS3	6.90	<u>0.71</u>	<u>7.10 ± 0.39</u>	<u>6.12 ± 0.44</u>	82.00	8.29	9.45	5.88	0.00	21.12
Spark-TTS	9.25	0.68	6.18 ± 0.58	5.73 ± 0.54	36.33	49.22	3.48	<u>7.35</u>	<u>2.94</u>	19.86
RLAIF-SPA	<u>5.80</u>	0.72	7.23 ± 0.40	6.53 ± 0.52	<u>80.85</u>	<u>16.06</u>	38.80	8.82	5.88	30.08
<i>MELD (En)</i>										
Chat-TTS	19.64	<u>0.54</u>	6.17 ± 0.46	5.85 ± 0.35	61.17	29.37	21.60	3.44	20.56	27.23
F5-TTS	<u>9.07</u>	0.44	<u>6.63 ± 0.43</u>	6.10 ± 0.32	71.57	24.60	14.40	36.26	15.56	<u>32.48</u>
MegaTTS3	9.57	0.53	6.58 ± 0.39	<u>6.25 ± 0.33</u>	67.13	22.22	29.60	3.82	28.89	30.33
Spark-TTS	15.54	0.53	6.18 ± 0.49	5.97 ± 0.41	27.79	50.79	9.60	2.29	38.33	25.76
RLAIF-SPA	8.92	0.55	6.73 ± 0.45	6.30 ± 0.37	<u>68.65</u>	<u>29.76</u>	41.60	<u>5.73</u>	<u>29.44</u>	35.04
<i>ESD (Zh)</i>										
Chat-TTS	6.70	0.67	6.06 ± 0.44	5.92 ± 0.34	63.14	6.35	2.57	2.57	2.86	15.50
F5-TTS	4.01	0.69	<u>6.45 ± 0.44</u>	6.06 ± 0.35	<u>87.71</u>	0.00	0.00	10.29	0.57	19.71
MegaTTS3	3.86	0.72	6.40 ± 0.41	6.28 ± 0.40	54.00	51.14	0.00	0.86	0.00	21.20
Spark-TTS	2.40	0.70	6.20 ± 0.41	5.99 ± 0.35	82.29	5.56	0.00	4.00	0.00	18.37
RLAIF-SPA	<u>3.68</u>	0.74	6.50 ± 0.48	6.29 ± 0.40	89.71	<u>7.14</u>	4.57	<u>4.86</u>	3.14	21.88

Baselines. We compare RLAIF-SPA against four representative TTS baselines: Chat-TTS, which is also used as the underlying speech synthesis engine in our system; F5-TTS, a non-autoregressive flow-matching model based on Diffusion Transformers; MegaTTS3, a latent diffusion transformer with sparse speech-text alignment; and Spark-TTS, an LLM-based autoregressive TTS system built upon the Qwen2.5 backbone. Our overall framework is built upon MiniCPM-o 2.6, with Chat-TTS instantiated as its speech synthesis component.

Evaluation Metrics. We assess model performance using both objective and subjective metrics to evaluate accuracy, speaker consistency, and emotional expressiveness.

For objective evaluation, we consider three aspects. Intelligibility is measured using the Word Error Rate (WER), obtained by transcribing synthesized speech with Whisper-Large-v3 and comparing it against the reference text. Speaker similarity is quantified using SIM-O, computed via WavLM-Large [33], where scores range from $[-1, 1]$ and higher values indicate closer similarity between the generated speech and the prompt. Emotional correctness is evaluated using a Speech Emotion Recognizer, emotion2vec-large [34], which predicts an emotion label for each synthesized utterance. We report SER accuracy for each emotion category and the unweighted macro-average across the five categories.

For subjective evaluation, we conduct human listening tests to assess perceptual quality and emotional fidelity. Speech naturalness is evaluated using the Mean Opinion Score (MOS), including CMOS for overall quality (covering clarity, naturalness, and high-frequency details) and Emotion MOS for perceived emotional similarity between synthesized and ground-truth speech. We report 95% confidence intervals for both metrics using a two-way cluster bootstrap that jointly resamples raters and utterances, presenting results as mean $\pm$ CI half-width. In total, 50 randomly selected samples are evaluated, each by at least 20 listeners.

Implementation Details. RLAIF-SPA is built upon the

MiniCPM-O 2.6 model. During GRPO training, the reward signal is computed automatically: Whisper-Large-v3 is used to compute WER for accuracy assessment, while Qwen2-Audio evaluates alignment with the four prosodic-emotional labels. The corresponding weights are set to $\alpha_1 = 0.3$ for the WER penalty and $\alpha_2 = 0.7$ for the label-based reward, with uniform weights (w_k) applied across all label dimensions. All components are initialized from pre-trained MiniCPM-O 2.6 checkpoints. Training is conducted for 7 epochs using a learning rate of 5×10^{-6} .

V. EXPERIMENTAL RESULTS

A. Overall Performance

We evaluate the effectiveness of RLAIF-SPA by benchmarking it against four strong baselines: Chat-TTS, F5-TTS, MegaTTS3, and Spark-TTS. Table I summarizes the comparative performance across three diverse datasets, detailing the results from general objective (WER, SIM-O) and subjective (CMOS, Emotion MOS) metrics, as well as Speech Emotion Recognition (SER) accuracy.

As shown in the results, RLAIF-SPA maintains high accuracy across testing scenarios, balancing semantic accuracy with expressive generation. While non-autoregressive baselines exhibit transcription accuracy on clean data, RLAIF-SPA demonstrates superior robustness in complex, emotionally dynamic contexts, where it achieves lower WER compared to competitive models. This stability is a direct consequence of our methodology, which explicitly incorporates a WER-based penalty into the AI Feedback mechanism. By optimizing for semantic alignment alongside emotional expressiveness, our model is guided to produce speech that preserves articulation precision even when modulating intense prosody.

Beyond accuracy, RLAIF-SPA excels in emotional expressiveness and overall speech quality. This advancement is principally driven by our fine-grained, label-driven reward component, which enables the model to precisely modulate

TABLE II: Ablation Study on Model Performance. Base denotes the original model without post-training, and w/o Label Reward removes the prosodic label reward during post-training while keeping GRPO.

Model	Objective		Subjective		Speech Emotion Recognition					
	WER↓	SIM-O↑	CMOS↑	Emotion MOS↑	Neutral↑	Happy↑	Sad↑	Angry↑	Surprise↑	Avg↑
<i>LibriSpeech test-clean (En)</i>										
RLAIF-SPA	5.80	0.72	7.23 ± 0.40	6.53 ± 0.52	80.85	16.06	38.80	8.82	5.88	30.08
w/o Label Reward	8.08	0.65	5.78 ± 0.38	5.31 ± 0.48	78.23	23.32	18.91	1.47	0.00	24.39
Base (No Post-training)	8.89	0.63	5.59 ± 0.28	5.33 ± 0.48	76.10	15.87	21.89	2.94	0.00	23.36
<i>MELD (En)</i>										
RLAIF-SPA	8.92	0.55	6.73 ± 0.45	6.30 ± 0.37	68.65	29.76	41.60	5.73	29.44	35.04
w/o Label Reward	18.90	0.54	6.17 ± 0.41	5.52 ± 0.29	70.56	25.40	41.60	3.44	17.22	31.64
Base (No Post-training)	17.32	0.53	5.46 ± 0.26	5.90 ± 0.39	67.64	25.79	36.8	3.44	18.33	30.4
<i>ESD (Zh)</i>										
RLAIF-SPA	3.68	0.74	6.50 ± 0.48	6.29 ± 0.40	89.71	7.14	4.57	4.86	3.14	21.88
w/o Label Reward	3.83	0.73	5.96 ± 0.29	5.66 ± 0.24	88.57	3.14	4.29	1.43	1.14	19.71
Base (No Post-training)	3.75	0.73	5.89 ± 0.35	6.05 ± 0.38	91.14	1.14	5.14	0.86	2.29	20.11

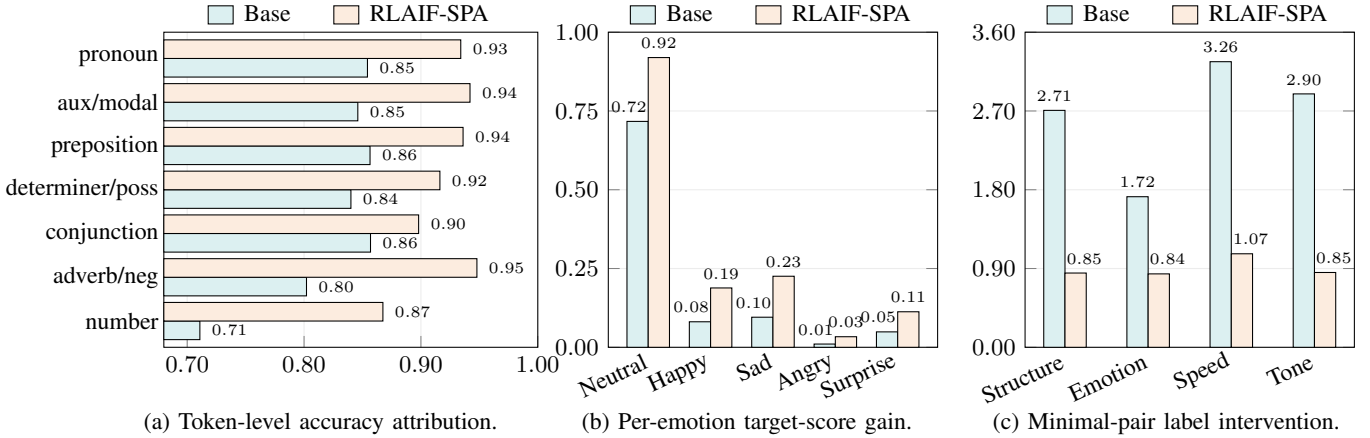


Fig. 2: Fine-grained analyses of the ablation study. (a) Token-level accuracy change based on ASR alignment. (b) Mean increase in target-emotion recognizer score, grouped by emotion. (c) Minimal-pair label interventions showing normalized changes on target vs. non-target dimensions.

prosodic nuances across dimensions such as Structure, Emotion, and Tone. The model’s superiority is validated through a suite of evaluations. Objectively, it consistently achieves the highest speaker similarity across all datasets. Regarding Speech Emotion Recognition, while certain baselines may show competitive performance in specific emotion categories, RLAIF-SPA achieves the highest average accuracy across all testing sets. This indicates that our model possesses a superior generalized ability to synthesize a diverse range of emotions distinctly, rather than over-optimizing for a single emotional style. Subjectively, human listeners award RLAIF-SPA higher ratings for both overall naturalness and emotional fidelity.

B. Ablation Study

We conduct an ablation to isolate the effect of structured feedback by comparing Base, w/o Label Reward, and RLAIF-SPA. The w/o Label Reward variant uses only the ASR-based fidelity reward during GRPO post-training, while RLAIF-SPA additionally includes attribute-wise label rewards. Table II shows that optimizing with the ASR-based fidelity reward can improve transcription accuracy on clean read speech, but its effect is less consistent under domain shift. Adding attribute-wise label rewards yields the most reliable overall trade-off, improving expressiveness while keeping WER competitive.

To complement the aggregate metrics, we design three diagnostic analyses in Fig. 2 that directly test the key claims of our method: the fidelity feedback improves accuracy, the attribute-wise reward strengthens the intended emotion, and the structured decomposition reduces cross-attribute entanglement. First, Fig. 2a breaks down transcription accuracy by token categories, showing broadly distributed gains rather than improvements concentrated in a single token type, which supports that the fidelity signal improves overall content preservation. Second, Fig. 2b reports increased recognizer confidence for the ground-truth emotion across classes, confirming that attribute-wise supervision effectively amplifies the intended affect. Finally, Fig. 2c evaluates minimal-pair label flips and observes lower non-target interference, demonstrating that decomposed feedback yields better disentanglement with reduced cross-attribute leakage.

VI. CONCLUSION AND FURTHER WORK

This paper presents RLAIF-SPA, a novel framework that autonomously optimizes for both emotional expressiveness and accuracy in speech synthesis. By employing an AI Feedback mechanism with GRPO to enforce fine-grained prosodic consistency and semantic accuracy, RLAIF-SPA significantly outperforms strong baseline models on the LibriSpeech, MELD

and ESD datasets. Crucially, our work demonstrates the feasibility of generating emotionally rich and highly intelligible speech without reliance on costly manual annotations, paving the way for more scalable and data-efficient emotional TTS systems. Future work will focus on refining the reward mechanism and assessing the framework’s scalability across a broader range of acoustic environments and languages. Another promising direction involves modeling how a speaker’s transient emotional state dynamically shapes prosody.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation of China (Nos. 62276056 and U24A20334), the Yunnan Fundamental Research Projects (No.202401BC070021), the Yunnan Science and Technology Major Project (No. 202502AD080014), the Fundamental Research Funds for the Central Universities (Nos. N25BSS054 and N25BSS094), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No.B16009).

REFERENCES

- [1] Huda Barakat, Oytun Turk, and Cenk Demiroglu, “Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, pp. 11, 2024.
- [2] Xiaoqian Liu, Xiyan Gui, Zhengkun Ge, Yuan Ge, Chang Zou, Jiacheng Liu, et al., “Waveex: Accelerating flow matching-based speech generation via wavelet-guided extrapolation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, vol. 40, pp. 32177–32185.
- [3] Li-Wei Chen, Shinji Watanabe, and Alexander Rudnicky, “A vector quantized approach for text to speech synthesis on real-world spontaneous speech,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 12644–12652.
- [4] Se Jin Park, Yeonju Kim, Hyeonseop Rha, Bella Godiva, and Yong Man Ro, “Av-emodialog: Chat with audio-visual users leveraging emotional cues,” *arXiv preprint arXiv:2412.17292*, 2024.
- [5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, et al., “Constitutional ai: Harmlessness from ai feedback,” *arXiv preprint arXiv:2212.08073*, 2022.
- [6] Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, et al., “Controlspeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control,” in *ACL*, 2025.
- [7] Daria Diatlova and Vitalii Shutov, “Emospeech: guiding fastspeech2 towards emotional text to speech,” in *12th ISCA Speech Synthesis Workshop, SSW 2023*.
- [8] Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu, “Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance,” in *ICASSP*, 2023.
- [9] Minki Kang, Wooseok Han, Sung Ju Hwang, and Eunho Yang, “Zet-speech: Zero-shot adaptive emotion-controllable text-to-speech synthesis with diffusion and style-based models,” in *Interspeech*, 2023.
- [10] Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao, “Emomix: Emotion mixing via diffusion models for emotional speech synthesis,” in *Interspeech*, 2023.
- [11] Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, Chunghyun Ahn, and Hong-Goo Kang, “Emotional speech synthesis with rich and granularized control,” in *ICASSP*, 2020.
- [12] Wenhao Guan, Yishuang Li, Tao Li, Hukai Huang, Feng Wang, Jiayan Lin, et al., “Mm-tts: Multi-modal prompt based style transfer for expressive text-to-speech synthesis,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 18117–18125.
- [13] Haibin Wu, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Daniel Tompkins, Chung-Hsien Tsai, et al., “Laugh now cry later: Controlling time-varying emotional states of flow-matching-based zero-shot text-to-speech,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 690–697.
- [14] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn, “Direct preference optimization: Your language model is secretly a reward model,” in *NeurIPS*, 2023.
- [15] Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi Lee, and Ivan Buluko, “Align-slm: Textless spoken language models with reinforcement learning from ai feedback,” in *ACL*, 2025.
- [16] Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Qinmei Xu, et al., “BATON: aligning text-to-audio model using human preference feedback,” in *IJCAI*, 2024, pp. 4542–4550, ijcai.org.
- [17] 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.
- [18] Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu, “Speechalign: Aligning speech generation to human preferences,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 50343–50360, 2024.
- [19] Geoffrey Cideron, Sertan Girgin, Mauro Verzett, Damien Vincent, Matej Kastelic, Zalán Borsos, et al., “Musicrl: Aligning music generation to human preferences,” in *ICML*, 2024.
- [20] Yuan Ge, Haishu Zhao, Aokai Hao, Junxiang Zhang, Bei Li, Xiaoqian Liu, et al., “On the emotion understanding of synthesized speech,” *arXiv preprint arXiv:2603.16483*, 2026.
- [21] Pauline Larrouy-Maestri, David Poeppel, and Marc D Pell, “The sound of emotional prosody: Nearly 3 decades of research and future directions,” *Perspectives on Psychological Science*, vol. 20, no. 4, pp. 623–638, 2025.
- [22] Kari Kallinen and Niklas Ravaja, “Emotion-related effects of speech rate and rising vs. falling background music melody during audio news: the moderating influence of personality,” *Personality and Individual Differences*, vol. 37, no. 2, pp. 275–288, 2004.
- [23] Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee, “Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech,” in *Interspeech*, 2024.
- [24] Julio Galdino, Ariadne Matos, Flaviane Svartman, and Sandra Aluisio, “The evaluation of prosody in speech synthesis: a systematic review,” *Journal of the Brazilian Computer Society*, vol. 31, pp. 466–487, 07 2025.
- [25] Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, et al., “Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark,” in *Interspeech*, 2024.
- [26] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, et al., “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” *arXiv preprint arXiv:2402.03300*, 2024.
- [27] Chang Liu, Ya-Jun Hu, Ying-Ying Gao, Shi-Lei Zhang, and Zhen-Hua Ling, “Group relative policy optimization for text-to-speech with large language models,” *arXiv preprint arXiv:2509.18798*, 2025.
- [28] Haishu Zhao, Aokai Hao, Yuan Ge, Zhenqiang Hong, Tong Xiao, and Jingbo Zhu, “Stylebench: Evaluating speech language models on conversational speaking style control,” *arXiv preprint arXiv:2603.07599*, 2026.
- [29] Yuan Ge, Junxiang Zhang, Xiaoqian Liu, Bei Li, Xiangnan Ma, Chenglong Wang, et al., “Sagelm: A multi-aspect and explainable large language model for speech judgement,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, vol. 40, pp. 30807–30815.
- [30] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *ICASSP*, 2015.
- [31] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, “MELD: A multimodal multi-party dataset for emotion recognition in conversations,” in *ACL*, July 2019.
- [32] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li, “Emotional voice conversion: Theory, databases and esd,” *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [33] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.
- [34] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen, “emotion2vec: Self-supervised pre-training for speech emotion representation,” in *Findings of ACL*, 2024.