Emo-LiPO: Listwise Preference Optimization for Fine-Grained Emotion Intensity Control in LLM-based Text-to-Speech

Yihang Lin[1], **Li Zhou**[1*], **Congwei Cao**[1,5], **Dongchu Xie**[1], **Xiaoxue Gao**[2], **Chen Zhang**[3] and **Haizhou Li**[1,3,4,5*]
[1]The Chinese University of Hong Kong, Shenzhen
[2]Agency for Science, Technology and Research
[3]National University of Singapore
[4]Shenzhen Research Institute of Big Data
[5]Shenzhen Loop Area Institute
yihanglin@link.cuhk.edu.cn, {lizhou21,haizhouli}@cuhk.edu.cn

Abstract

Large language model (LLM)-based text-to-speech (TTS) systems enable prompt-conditioned emotional control but struggle with fine-grained emotion intensity due to the semantic–acoustic gap between text and speech. To address this challenge, we formulate emotion intensity control in LLM-based TTS as a learning-to-rank problem and propose **Emo-LiPO**, a listwise preference optimization framework that aligns prompt-conditioned speech generation with relative emotion intensity expressed in text. Emo-LiPO explicitly models global intensity ordering within each emotion under fixed transcripts, enabling more faithful and continuous emotional expression. We further construct **ESD-plus**, a multi-speaker dataset with explicit emotion intensity variations, to support fine-grained emotion modeling and evaluation. Experiments on ESD-plus demonstrate that Emo-LiPO significantly improves emotion accuracy and intensity controllability over both supervised- and DPO-based LLM TTS baselines, with particularly pronounced gains at high intensity levels.

1 Introduction

Large language model (LLM)-based text-to-speech (TTS) systems have recently enabled prompt-conditioned control over speaking style and emotion [Leng *et al.*, 2023; Du *et al.*, 2024a; Du *et al.*, 2024b; Zhang *et al.*, 2025b; Yang *et al.*, 2025a; Zhou *et al.*, 2026], offering substantially greater flexibility than traditional signal-driven approaches [Zhou *et al.*, 2022b; Wang *et al.*, 2023; Schnell and Garner, 2021; Im *et al.*, 2022; Lei *et al.*, 2022; Cho *et al.*, 2024; Cho *et al.*, 2025]. Despite this progress, fine-grained control over emotion intensity remains a fundamental challenge for prompt-conditioned LLM-based TTS systems [Guo *et al.*, 2023b; Leng *et al.*, 2023; Du *et al.*, 2024a; Yang *et al.*, 2025a].

*Corresponding author.

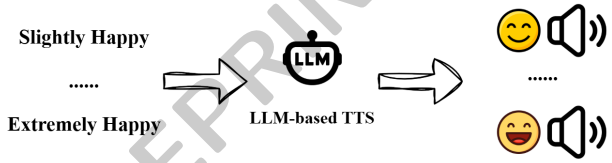

Figure 1: Illustration of the task of prompt-conditioned emotion intensity control in LLM-based TTS.

A key difficulty lies in the semantic–acoustic gap between textual emotion descriptions and their realization in speech. While natural language prompts can effectively convey emotion categories (e.g., *happy* or *sad*), they often fail to reliably induce perceptually distinguishable variations in emotion strength [Guo *et al.*, 2023b; Leng *et al.*, 2023; Zhang *et al.*, 2025b], as illustrated in Figure 1. As a result, existing models tend to realize emotion intensity implicitly at the acoustic level, leading to unstable or weakly differentiated intensity expressions, especially under high-intensity emotional prompts [Du *et al.*, 2024a; Du *et al.*, 2024b; Yang *et al.*, 2025a]. This limitation significantly constrains the controllability and expressiveness of emotional TTS systems in real-world applications.

Meanwhile, recent advances in preference optimization have demonstrated strong potential for aligning generative models with subjective human judgments. Pair-wise and list-wise preference learning frameworks provide principled mechanisms for modeling relative comparisons, making them particularly suitable for perceptual generation tasks [Rafailov *et al.*, 2023; Zhao *et al.*, 2023; Liu *et al.*, 2025a]. In the context of TTS, emerging preference-aligned methods have shown improved emotional expressiveness and perceptual quality by encouraging preferred outputs over less preferred ones [Zhang *et al.*, 2024b; Gao *et al.*, 2025; Liu *et al.*, 2025b; Yang *et al.*, 2025b; Li *et al.*, 2025]. However, existing preference-based TTS approaches primarily focus on controlling emotion at the categorical level, without explicitly modeling the ordinal structure of emotion intensity. Consequently, relative intensity information specified in text

prompts is not directly aligned with preference supervision during training.

In this work, we address the challenge of fine-grained emotion intensity control in LLM-based TTS by framing it as a learning-to-rank problem. We adopt a listwise preference optimization approach that explicitly aligns prompt-conditioned speech generation with relative emotion intensity conveyed by natural language. To facilitate systematic modeling and evaluation of emotion intensity, we additionally construct a multi-speaker dataset with explicit intensity variations and conduct extensive experiments to validate the effectiveness of our approach. Our main contributions can be summarized as follows:

- We propose Emo-LiPO[1], a listwise preference optimization framework that formulates emotion intensity control in LLM-based TTS as a learning-to-rank problem, enabling fine-grained and text-grounded intensity modeling.
- We construct ESD-plus[2], a multi-speaker emotional speech dataset with explicit emotion intensity variations to support fine-grained emotion intensity modeling and evaluation.
- Extensive experiments on ESD-plus demonstrate that EMO-LiPO outperforms supervised- and DPO-based baselines in emotion accuracy and intensity controllability, with especially pronounced gains at high intensity.

2 Related Work

2.1 Emotion Intensity Control in TTS

Early studies on emotion intensity control mainly operate in the acoustic domain, where emotion strength is explicitly modeled as a continuous variable derived from speech signals using relative ranking or learning-to-rank objectives [Zhou *et al.*, 2022b; Wang *et al.*, 2023]. Other approaches extract intensity representations at different temporal levels and inject them into TTS systems for fine-grained control [Schnell and Garner, 2021; Im *et al.*, 2022; Lei *et al.*, 2022]. More recent methods further disentangle emotion intensity from style using geometric representations or incorporate soft intensity guidance through diffusion-based models [Cho *et al.*, 2024; Cho *et al.*, 2025; Liang *et al.*, 2025; Guo *et al.*, 2023a; Wu *et al.*, 2024]. While effective in controlled settings, these approaches typically rely on explicit acoustic features or specialized architectures and lack flexible, text-driven intensity control. In contrast, recent prompt-conditioned and LLM-based TTS systems leverage natural language descriptions for expressive speech generation, but primarily control discrete emotion categories or coarse prosodic attributes, leaving emotion intensity implicitly realized [Guo *et al.*, 2023b; Leng *et al.*, 2023; Zhang *et al.*, 2025b; Du *et al.*, 2024a; Du *et al.*, 2024b; Yang *et al.*, 2025a]. Our approach directly grounds emotion intensity in natural language and aligns it with preference supervision, enabling explicit and fine-grained intensity control in LLM-based TTS.

[1]Code is available at hlt-cuhksz/Emo-LiPO.

[2]Dataset is available at hlt-cuhksz/ESD-plus.

2.2 Preference Optimization for Generation Tasks

Preference optimization aims to align pre-trained large language models (LLMs) with human preferences through gradient-based optimization during training. Existing approaches can be broadly categorized into point-wise, pair-wise, and list-wise methods. Point-wise methods optimize models using feedback defined on individual samples, typically in the form of scalar rewards or binary desirability signals, as exemplified by Proximal Policy Optimization (PPO) [Schulman *et al.*, 2017], Reward Maximization (ReMax) [Li *et al.*, 2024], and Kahneman–Tversky Optimization (KTO) [Ethayarajh *et al.*, 2024]. Pair-wise methods, such as Direct Preference Optimization (DPO) [Rafailov *et al.*, 2023; Zhang *et al.*, 2025a] and SLiC-HF [Zhao *et al.*, 2023], learn from relative preferences between two candidate outputs by contrasting a preferred response against a less preferred one, shaping the model's generation distribution accordingly [Zhang *et al.*, 2024a]. List-wise methods further generalize preference alignment as a ranking problem, directly optimizing list-structured preference signals within a learning-to-rank framework [Song *et al.*, 2024; Liu *et al.*, 2025a]. These methods motivate us by providing a principled approach to learning from subjective human judgments in natural language processing, making them especially well suited for perceptual generation tasks such as emotional speech synthesis.

2.3 Preference Alignment for TTS

Recent studies have begun to incorporate preference alignment into TTS systems to better match human perceptual judgments of speech quality, expressiveness, and emotional appropriateness. SpeechAlign [Zhang *et al.*, 2024b] is an early attempt in this direction, aligning codec-based speech language models by learning preferences between golden and synthetic codec tokens to mitigate distribution mismatches between training and inference. Building upon this paradigm, several works apply preference optimization to emotional speech synthesis. Emo-DPO [Gao *et al.*, 2025] and ARDM-DPO [Liu *et al.*, 2025b] adopt pair-wise preference objectives to encourage emotionally preferred outputs over less preferred ones, while RLAIF-SPA [Yang *et al.*, 2025b] and EMORL-TTS [Li *et al.*, 2025] employ reinforcement learning with task-specific rewards to improve emotional expressiveness and controllability. These approaches demonstrate the effectiveness of preference-based optimization for enhancing perceptual speech quality. However, existing preference-aligned TTS methods mainly focus on selecting better samples rather than explicitly modeling the ordinal structure of emotion intensity. As a result, emotion strength is typically optimized implicitly at the acoustic level, and relative intensity specified in text prompts is not directly aligned with preference signals.

3 Method

3.1 Problem Formulation

Task Definition. We investigate fine-grained control of emotion intensity in LLM-based TTS systems. Given a text

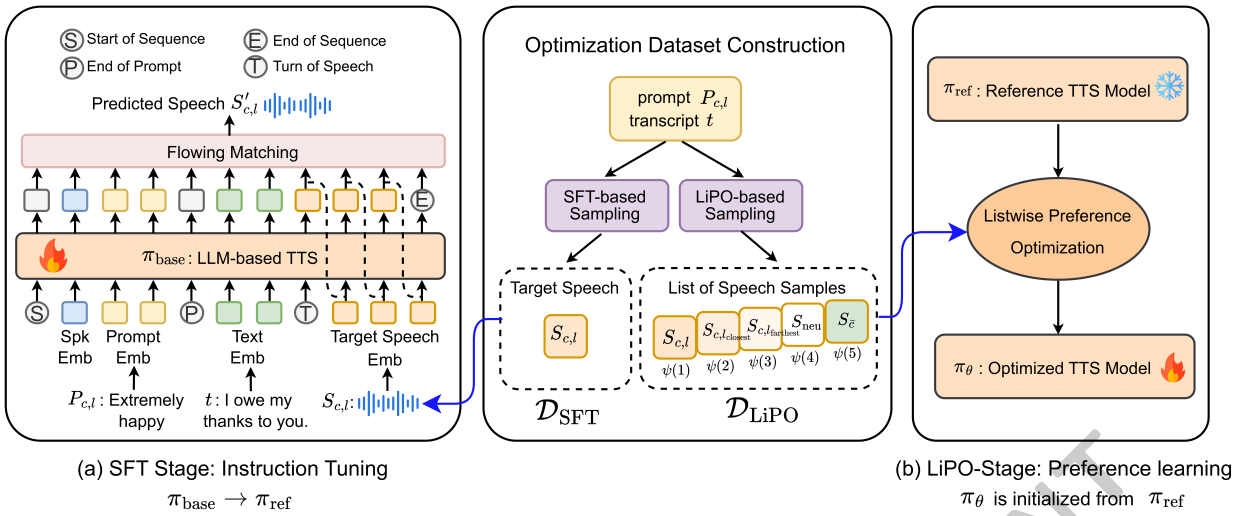

Figure 2: Emo-LiPO framework. The LLM-based model is trained with SFT followed by LiPO for fine-grained emotion intensity control.

transcript and an emotion specification, the TTS model generates speech conditioned on both linguistic content and emotional attributes. Natural language prompts specify an emotion category and, for non-neutral emotions, a relative intensity level. Our goal is to train an LLM-based TTS system to generate speech that accurately follows these specifications.

Formally, let $\mathcal{C} = \mathcal{C}_{\text{emo}} \cup \{\textit{neutral}\}$ denote the set of emotion categories, where $\mathcal{C}_{\text{emo}} = \{\textit{happy}, \textit{sad}, \textit{angry}, \ldots\}$ and *neutral* corresponds to emotionally flat speech. For each $c \in \mathcal{C}_{\text{emo}}$, we define an ordered set of intensity levels $\mathcal{L} = \{l_1, l_2, \ldots, l_K\}$, where $l_i < l_j$ for $i < j$ indicates strictly increasing emotional intensity (e.g., slightly, moderately, extremely). Crucially, the weakest level l_1 must remain perceptually distinct from neutral speech, preserving a clear boundary between emotional and non-emotional outputs.

We define the model input as $x = (t, P_{c,l})$, consisting of a text transcript t and an emotion prompt $P_{c,l}$ specifying an emotion category c and intensity level l. The TTS model generates speech $S = \pi_\theta(x)$, which should satisfy: (i) **Content fidelity**, meaning the speech faithfully conveys the linguistic content of the transcript t; (ii) **Category correctness**, meaning the generated speech follows emotion c exactly; and (iii) **Intensity ordering** for $c \in \mathcal{C}_{\text{emo}}$, such that speech generated with higher intensity levels is perceived as strictly stronger than that generated with lower levels.

Learning-to-Rank Formulation via LiPO (Emo-LiPO). To meet these requirements, we adopt the Listwise Preference Optimization (LiPO) framework [Liu *et al.*, 2025a] and formulate fine-grained emotion intensity control as a Learning-to-Rank (LTR) problem. In this formulation, all candidate speech samples compared under the same prompt are generated from the same text transcript. LiPO learns from listwise preference data defined over multiple candidates generated for the same prompt, making it well suited for modeling both emotion category correctness and global intensity ordering.

In contrast, Supervised Fine-Tuning (SFT) relies on independent prompt–target pairs, $\mathcal{D}_{\text{SFT}} = \{(x = (t, P_{c,l}),\ y = S_{c,l})\}$, which provide no supervision on relative emotional intensity across samples. Direct Preference Optimization (DPO) extends SFT by introducing pairwise preferences, $\mathcal{D}_{\text{DPO}} = \{(x = (t, P_{c,l}),\ y = S_{c,l}, \tilde{S}_{c,l}))\}$, where $\tilde{S}_{c,l}$ denotes a negative sample. While DPO captures local preference information, it remains limited to binary comparisons. In emotional TTS, existing DPO-based methods such as Emo-DPO [Gao *et al.*, 2025] similarly rely on pairwise emotional preferences[3]. As a result, these approaches are inherently constrained to local comparisons and cannot enforce a global ordering of emotional intensity.

Notably, Emo-LiPO adopts a listwise formulation, $\mathcal{D}_{\text{LiPO}} = \{(x = (t, P_{c,l}),\ y = (\mathcal{T}_{c,l}, \psi_{c,l}))\}$, where $\mathcal{T}_{c,l}$ is a list of speech samples generated by the TTS model, and $\psi_{c,l} \in [0, 1]^{|\mathcal{T}_{c,l}|}$ is a vector of real-valued preference scores associated with the corresponding samples, with larger values indicating stronger preference. This listwise supervision enables Emo-LiPO to explicitly model fine-grained emotional intensity rankings across multiple candidates.

3.2 Emo-LiPO Framework

We present the Emo-LiPO framework by specifying the construction of listwise preference data and the corresponding optimization objective under the LiPO formulation.

Rule-Based Preference Construction. For a prompt $P_{c,l}$, we construct a listwise preference set $\mathcal{T}_{c,l}$ of $K + 2$ speech candidates with the same text transcript t using a **rule-based ranking strategy**. The list consists of: (i) a *target* sample $S_{c,l}$ with the exact emotion category c and intensity level l; (ii) $K - 1$ *same-emotion* samples with the same category c but different intensity levels $l' \in \mathcal{L} \setminus \{l\}$; (iii) one *neutral* speech sample; and (iv) one *negative* sample with a randomly selected non-target emotion category. Under this rule-based

[3] In Emo-DPO, each training instance is defined as $x = (t, P_c)$, $y = (S_c, S_{\tilde{c}})$.

strategy, the $K-1$ *same-emotion* samples are ordered according to their absolute intensity distance $|l' - l|$, with samples closer to the target intensity being preferred; ties are resolved by random ordering. The resulting preference list follows the order

$$\mathcal{T}_{c,l} = [S_{c,l} \succ S_{c,l_{\text{closest}}} \succ \cdots \succ S_{c,l_{\text{farthest}}} \succ S_{\text{neu}} \succ S_{\bar{c}}], \quad (1)$$

where l_{closest} and l_{farthest} denote intensity levels of the same emotion category ordered by increasing distance from l.

Based on the rule-based ranking described above, we assign a real-valued preference label vector $\psi_{c,l} \in [0,1]^{|\mathcal{T}_{c,l}|}$ to the speech samples in $\mathcal{T}_{c,l}$. Specifically, we adopt an **index-based** preference scoring scheme, where the preference label of each sample is determined by its position in the ordered list $\mathcal{T}_{c,l}$. Let i denote the index of a sample in the list, with smaller indices indicating stronger preference. The preference label of the i-th sample is defined as

$$\psi_{c,l}(i) = 1 - \frac{i-1}{K+2}, \quad i = 1, \ldots, K+2. \quad (2)$$

Multi-Stage Optimization. As illustrated in Figure 2, following Emo-DPO [Gao *et al.*, 2025], we optimize the TTS model in a multi-stage manner, starting with SFT to learn basic instruction following, and subsequently applying LiPO for fine-grained emotion intensity control. The LLM-based TTS model formulates text-to-speech as a conditional autoregressive generation task [Du *et al.*, 2024a], in which speech tokens are generated sequentially conditioned on the joint input $x = (t, P_{c,l})$, together with a speaker embedding.

Specifically, in the SFT stage, we perform supervised fine-tuning on a paired prompt–speech dataset $\mathcal{D}_{\text{SFT}}$ to initialize the backbone TTS model π_{base}:

$$\mathcal{L}_{\text{SFT}}(\pi_{\text{base}}) = \mathbb{E}_{(x,S)\sim\mathcal{D}_{\text{SFT}}} [-\log \pi_{\text{base}}(S \mid x)], \quad (3)$$

where $\log \pi_{\text{base}}(S \mid x)$ is computed using teacher forcing with token-level cross-entropy over the autoregressively generated speech tokens. The resulting model is denoted as the reference TTS policy π_{ref} in the subsequent stage.

In the LiPO stage, initialized from π_{ref}, we optimize the model using the listwise preference dataset $\mathcal{D}_{\text{LiPO}}$ as defined in §3.1. Given a preference list $\mathcal{T}_{c,l}$ and its associated preference label vector $\psi_{c,l}$, we apply Listwise Preference Optimization to encourage the model to generate speech samples that better align with the desired emotion intensity ordering. The corresponding optimization objective is defined as:

$$\mathcal{L}_{\text{LiPO}}(\pi_\theta; \pi_{\text{ref}}, \beta) = \mathbb{E}_{(x,\mathcal{T}_{c,l},\psi_{c,l})\sim\mathcal{D}_{\text{LiPO}}} [r(\psi_{c,l}, \mathrm{s})], \quad (4)$$

where $r(\cdot)$ denotes a listwise learning-to-rank loss under the LiPO framework, and s represents the ranking scores induced by the current policy π_θ relative to the reference policy π_{ref}, The score of each candidate $S_i \in \mathcal{T}_{c,l}$ is computed as:

$$s_i = \beta \log \frac{\pi_\theta(S_i \mid x)}{\pi_{\text{ref}}(S_i \mid x)}. \quad (5)$$

Following the original LiPO formulation [Liu *et al.*, 2025a], $r(\cdot)$ is defined as:

$$r(\boldsymbol{\psi}_{c,l}, \mathbf{s}) = -\sum_{(i,j)\in\psi_{c,l}} \lambda_{i,j} (s_i - s_j), \quad (6)$$

where $\psi_{c,l} = \{(i,j) \mid \psi_{c,l}(i) > \psi_{c,l}(j)\}$, the relative orders of candidate pairs (S_i, S_j). The weighting term $\lambda_{i,j}$ serves as an intensity-aware coefficient that scales each pairwise loss term according to the implied emotion-intensity discrepancy between (S_i, S_j), injecting fine-grained intensity information beyond the binary preference order and promoting ordinal consistency for more reliable intensity control. We then define the gain and discount functions, $G(\cdot)$ and $D(\cdot)$, based on the position index i:

$$G(i) = 2^{\psi_{c,l}(i)} - 1, \qquad D(i) = \frac{1}{\log(1+i)}. \quad (7)$$

and $\lambda_{i,j}$ is derived as:

$$\lambda_{i,j} = |G(i) - G(j)| \cdot \left| \frac{1}{D(i)} - \frac{1}{D(j)} \right|. \quad (8)$$

A larger $\lambda_{i,j}$ implies a larger implied intensity gap between S_i and S_j under the list ranking, so the model is penalized more strongly for violating their order.

4 ESD-plus Dataset

4.1 Dataset Overview

To evaluate the effectiveness of the proposed Emo-LiPO method, we construct **ESD-plus**, a speech dataset with explicit emotion intensity variations designed to support research on fine-grained emotion control in TTS systems. ESD-plus is built upon the English portion of the ESD corpus [Zhou *et al.*, 2022a], from which we select 350 parallel sentences as synthesis scripts. We retain the original ESD emotion categories, including four non-neutral emotions (happy, surprise, sad, and angry) and one neutral emotion. For each non-neutral emotion, we further introduce three distinct intensity levels, resulting in a total of 13 fine-grained emotion labels.

In total, ESD-plus contains 45,500 fine-grained emotional speech samples, with an average duration of 2.92 seconds per sample, amounting to approximately 36.89 hours of audio for training and evaluation. Following the official data partitioning strategy of ESD [Zhou *et al.*, 2022a], we split the dataset into training, development, and test sets with 300, 20, and 30 utterances, respectively, while ensuring that all speaker–emotion variants of each utterance are assigned to the same split.

4.2 Dataset Construction and Quality Control

We construct ESD-plus using a prompt-based emotional speech synthesis pipeline built upon `gpt-4o-mini-tts`[4]. For each emotion–intensity category, we design a structured prompt that combines a natural-language emotion description with a scalar intensity indicator, providing unified and reusable conditional control signals. To ensure perceptually distinguishable emotional variations across different intensity levels, the scalar indicators are deliberately spaced (e.g., 1/5, 3/5, and 5/5) rather than uniformly incremental.[5] Moreover,

[4] https://openai.fm/

[5] Details on dataset construction, quality control, and dataset statistics are provided in the supplementary material (Appendix A).

Model	Speech Quality				Emotion Relevance		
	WER ↓	NISQA ↑	DNSMOS ↑	UTMOS ↑	EmoSIM ↑	Recall ↑	Recall-ft ↑
CosyVoice	4.47	4.71	3.16	**4.30**	81.87	25.10	29.90
EmoVoice	5.40	4.79	**3.29**	4.27	89.84	20.56	28.51
Emo-DPO (R)	12.78	4.37	3.04	3.90	91.52	24.77	33.46
Emo-DPO (E)	4.78	4.66	3.23	4.06	91.73	24.08	34.92
Emo-DPO (I)	6.79	4.60	3.21	4.00	91.85	26.87	37.21
Emo-LiPO	4.26	**4.79**	3.26	4.18	91.93	**27.56**	**39.54**
- w/o λ	**4.15**	4.79	3.24	4.17	**92.30**	26.21	37.59

Table 1: Comparison of different prompt-conditioned TTS models in terms of speech quality and emotion relevance.

we generate speech using 10 official English voices, comprising 4 male and 6 female speakers, enabling multi-speaker emotional synthesis.

To ensure data reliability, we perform human verification on the development and test sets of ESD-plus, which are used exclusively for model evaluation. Annotators perform pairwise comparisons within each emotion category to verify that speech samples generated at different intensity levels exhibit consistent and monotonically increasing emotional intensity in accordance with their prompts. Samples that fail to demonstrate the expected intensity ordering are flagged and regenerated until the target ordering is satisfied, thereby ensuring the correctness of emotion–intensity labels in the development and test sets. Overall, 91.23% of the evaluated samples pass verification, demonstrating that ESD-plus provides high-quality and reliable data for model training and evaluation.[6]

5 Experiments

5.1 Experimental Setup

Baselines and Implementation. We adopt several LLM-based TTS models as baselines. Specifically, we include two supervised LLM-based TTS models: **CosyVoice** [Du *et al.*, 2024a], which is built upon supervised semantic tokens, and **EmoVoice** [Yang *et al.*, 2025a], which supports freestyle text prompting for emotional control. In addition, we consider preference-optimized LLM-based TTS methods following Emo-DPO [Gao *et al.*, 2025], which primarily target coarse-grained emotion category control. To evaluate their effectiveness in modeling fine-grained emotion intensity, we design three Emo-DPO variants: (i) **Emo-DPO (R)**, trained with randomly sampled pairwise preferences $(S_{c,l}, \bar{S}_{c,l})$; (ii) **Emo-DPO (E)**, trained using pairwise preferences across different emotion categories at the same intensity level $(S_{c,l}, S_{\bar{c},l})$; and (iii) **Emo-DPO (I)**, trained using pairwise preferences across different intensity levels within the same emotion category $(S_{c,l}, S_{c,\bar{l}})$. Our **Emo-LiPO** model is built upon **CosyVoice** as the backbone, specifically using the `CosyVoice-300M-Instruct` version.

Evaluation Metric. We evaluate the generated speech from two complementary perspectives: speech quality and emotion relevance. At the speech level, we assess intelligibility and perceptual quality. Speech intelligibility is measured using Word Error Rate (**WER**), computed with the Whisper-Large-v3 ASR model [Radford *et al.*, 2023]. Speech naturalness and overall perceptual quality are evaluated using three objective metrics: **NISQA** [Mittag *et al.*, 2021], **DNSMOS** [Reddy *et al.*, 2021], and **UTMOS** [Saeki *et al.*, 2022]. At the emotion level, we evaluate how well the generated speech conveys the intended emotional content. We employ the emotion2vec model [Ma *et al.*, 2024] for speech emotion recognition (SER) and adopt two metrics: **Emotion Similarity** (**EmoSIM**) and **Recall**. EmoSIM is computed as the cosine similarity between emotion embeddings extracted from the generated and ground-truth speech. Recall measures the emotion classification accuracy of emotion2vec on the generated speech. Due to its limited out-of-domain performance, we further fine-tune emotion2vec on ESD-plus and report the resulting score as **Recall-ft**.

[6]The training set is not modified during this process, as its overall label correctness is already high.

5.2 Emotion Category Control Performance

Effect of Intensity-Aware Preference Modeling. Table 1 summarizes the performance of different prompt-conditioned TTS models in terms of emotion category control, together with their speech quality metrics. Overall, Emo-LiPO demonstrates the strongest ability to generate speech that aligns with the intended emotion categories, consistently outperforming both supervised baselines and DPO-based methods, while maintaining comparable speech quality. Compared with supervised LLM-based TTS models, preference-optimized approaches show clear advantages in emotion relevance, indicating that incorporating preference signals is critical for learning effective emotional control without sacrificing speech naturalness. Among the DPO-based variants, models trained with structured emotion-related preferences achieve better emotion category recognition than those trained with randomly sampled pairs, highlighting the importance of informative preference construction. Notably, introducing intensity-aware preference modeling further improves emotion category control. Emo-DPO trained with intensity-level preferences performs better than variants relying solely on cross-category comparisons, suggesting that modeling relative intensity ordering within emotions provides transferable supervision for category-level emotion recognition. Finally, Emo-LiPO consistently achieves the best performance across emotion relevance metrics, demonstrating that listwise pref-

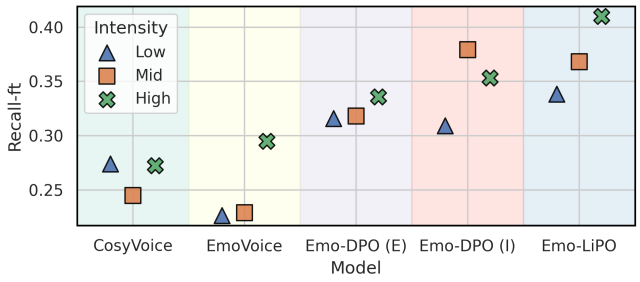

Figure 3: Emotion recognition accuracy (Recall-ft) across different emotion intensity levels.

erence optimization enables more accurate and robust emotion category control under fine-grained emotional prompting, without degrading speech quality.

The impact of weighting term λ. To analyze the role of the distance-based weighting term λ (Eq. 8) in learning stable intensity ordering, we conduct an ablation study in which λ is fixed to 1 for all preference pairs, resulting in a uniformly weighted LiPO loss, while keeping all other training settings unchanged. As shown in Table 1, adopting uniform weighting leads to a slight improvement in emotion similarity but consistently degrades emotion control metrics compared with the full model. These findings indicate that λ does not primarily improve category-level alignment; rather, it provides more informative supervision for intensity-aware ranking by emphasizing ordinal consistency across candidates, thereby enhancing controllability and generalization during evaluation.

5.3 Emotion Intensity Control Performance

Automatic Evaluation Comparison. Intuitively, higher emotional intensity leads to more perceptually salient speech and should therefore result in higher emotion recognition accuracy. Accordingly, we evaluate emotion recognition accuracy (Recall-ft) across different emotion intensity levels, as shown in Figure 3. Emo-LiPO achieves the strongest overall performance and is the only model that exhibits a clear and stable monotonic increase in Recall-ft from low to high intensity levels. In contrast, supervised baselines and DPO-based variants display weaker or non-monotonic trends, indicating inconsistent alignment between textual intensity descriptions and the generated speech. These results demonstrate that Emo-LiPO more faithfully translates relative emotion intensity specifications into perceptually distinguishable speech, validating its effectiveness in fine-grained emotion intensity control.

Human Evaluation Comparison. To further validate the effectiveness of Emo-LiPO in emotion intensity control, we conduct a human Arena-style evaluation. We perform pairwise comparisons between Emo-LiPO and three representative baselines (CosyVoice, Emo-DPO (E), and Emo-DPO (I)) along three perceptual dimensions: *Speech Quality*, *Emotion Expression*, and *Intensity Control*. For each comparison, annotators are presented with the same text prompt and two candidate models, each providing a pair of speech samples generated under the same script and emotion category at adjacent

	CosyVoice	Emo-DPO (E)	Emo-DPO (I)
Speech Quality	94.29	79.37	89.28
Emotion Expression	90.34	80.65	78.24
Intensity Control	86.08	66.44	58.33

Table 2: Arena win rate (%) of Emo-LiPO compared against baseline models across three evaluation dimensions.

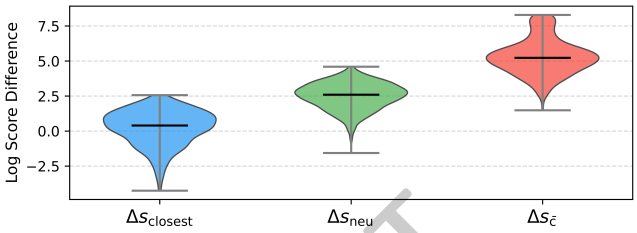

Figure 4: Score Margin Distributions Between Target and Different Candidates.

intensity levels. The evaluated intensity transitions include $(S_{\text{neu}}, S_{c,l_{\text{low}}})$, $(S_{c,l_{\text{low}}}, S_{c,l_{\text{mid}}})$, and $(S_{c,l_{\text{mid}}}, S_{c,l_{\text{high}}})$. Annotators select the model that performs better with respect to a single evaluation dimension, with a tie option allowed when performance is comparable. The order of models is randomized, and five annotators participate in the evaluation. [7]

Table 2 reports the Arena win rates of Emo-LiPO compared with three baseline models across three perceptual dimensions. Overall, Emo-LiPO consistently outperforms all baselines in all evaluation dimensions, demonstrating strong perceptual advantages under human judgment. When compared with Emo-DPO (I), the advantage of Emo-LiPO in intensity control is less pronounced, suggesting that incorporating intensity-aware pairwise preferences already captures part of the fine-grained intensity information. However, Emo-LiPO consistently outperforms Emo-DPO (I) in *Speech Quality*, indicating that listwise preference optimization better preserves acoustic naturalness while maintaining competitive intensity controllability. This comparison highlights a favorable trade-off achieved by Emo-LiPO, where global intensity ordering is modeled without sacrificing speech quality.

Score Distance Visualization. To diagnose whether the Emo-LiPO policy π_θ induces a separable preference structure consistent with the rule-based list supervision in Eq. 1, we visualize the distribution of score margins between the *target* sample $S_{c,l}$ and the candidates in $\mathcal{T}_{c,l}$. Based on Eq. 5, we compute three absolute margins aligned with the list components: (i) the margin to closest same emotion sample, $\Delta s_{\text{closest}} = |s(S_{c,l}) - s(S_{c,l_{\text{closest}}})|$; (ii) the margin to neutral sample, $\Delta s_{\text{neu}} = |s(S_{c,l}) - s(S_{\text{neu}})|$; and (iii) the margin to other emotion sample, $\Delta s_{\bar{c}} = |s(S_{c,l}) - s(S_{\bar{c}})|$. For comparability across scales, we plot the logarithm of these margins. As shown in Fig. 4, the three margins exhibit a stable hierarchy in both location and central tendency: $\Delta s_{\bar{c}}$ is the largest

[7] Detailed human evaluation settings, including the evaluation scale, platform, and annotation guidelines, are provided in the supplementary material (Appendix B).

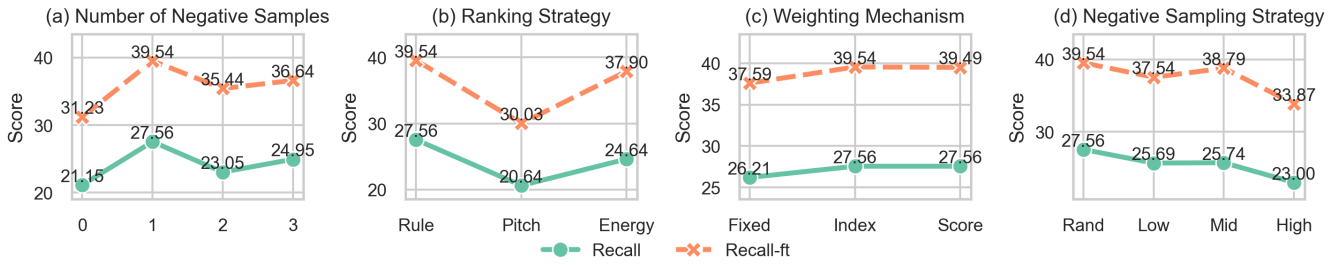

Figure 5: Ablation study of Emo-LiPO on different design choices.

and most concentrated, followed by Δs_{neu}, while $\Delta s_{\text{closest}}$ is the smallest and closer to zero with a longer tail, reflecting the finer and more challenging discrimination between nearest intensities within the same emotion. Overall, this distributional hierarchy is consistent with the rule-based order in Eq. 1, indicating that Emo-LiPO shapes a scoring landscape that prioritizes cross-emotion separation while preserving fine-grained intensity hierarchy modeling within the same emotion.

5.4 Ablation Studies and Analysis

Effect of the Number of Negative Samples $S_{\bar{c}}$. We investigate how the number of negative samples affects intensity learning in listwise optimization by varying the number of negatives while keeping the target intensity chain fixed. Each negative sample is drawn from a non-target emotion category under the same text condition. As shown in Figure 5 (a), introducing a single negative sample yields the best overall performance and produces a clearer monotonic trend across intensity levels. In contrast, increasing the number of negative samples to two or three does not lead to further gains and instead degrades performance. These results suggest that a limited amount of cross-emotion contrast is sufficient for effective intensity learning, whereas excessive negative samples weaken the supervision signal for modeling intensity ordering.

Effect of Ranking Strategy for Preference List Construction. To study how the ranking strategy for constructing preference lists $\mathcal{T}_{c,l}$ affects intensity learning, we compare three ordering schemes on the same candidate set. Specifically, we evaluate the proposed rule-based strategy against two acoustic-driven alternatives, where candidates are ranked by their distance to the target speech in terms of average pitch or energy. As shown in Figure 5 (b), the rule-based ranking achieves the best overall performance among the three strategies. In comparison, pitch-based ranking performs the worst, while energy-based ranking yields intermediate results. These findings suggest that rankings derived from a single acoustic attribute provide weaker supervision signals, whereas the rule-based strategy offers a more effective and stable ordering for listwise preference optimization.

Effect of Weighting Mechanism in Listwise Optimization. We study the effect of the weighting mechanism λ in listwise optimization under identical training settings. By default, Emo-LiPO adopts an index-based preference labeling scheme for $\psi_{c,l}$, from which the distance-aware weighting term $\lambda_{i,j}$ is derived. We compare this default setting with two variants: (i) **fixed-λ**, which disables weighting by setting $\lambda_{i,j} = 1$ for all pairs; and (ii) **score-λ**, where $\lambda_{i,j}$ is computed from pre-assigned preference score gaps. As shown in Figure 5 (c), introducing distance-aware weighting consistently improves performance over the fixed-λ variant. Index-based and score-based weighting achieve similar results, indicating that the performance gains mainly stem from the weighting mechanism itself rather than the specific definition of $\psi_{c,l}$.

Impact of Negative Sampling Strategies. We study how the sampling strategy for the single negative sample $S_{\bar{c}}$ affects intensity learning in listwise optimization. By default, our method samples the negative uniformly at random across all intensity levels (*Rand*). Each preference list includes one negative sample drawn from a non-target emotion category. We further compare this default strategy with three alternatives that restrict the negative sample to a specific intensity level (*Low*, *Mid*, or *High*). As shown in Figure 5 (d), the default random strategy achieves the best overall performance, while constraining negatives to a single intensity level consistently degrades performance. These results indicate that sampling negatives across diverse intensity levels provides more balanced supervision, whereas single-level sampling introduces bias and weakens stable intensity ordering.

6 Conclusion

This work studies fine-grained emotion intensity control in prompt-conditioned LLM-based text-to-speech systems and identifies the limitations of existing supervised and preference-aligned methods in modeling perceptually meaningful intensity variations. To address this challenge, we propose **Emo-LiPO**, a listwise preference optimization framework that formulates emotion intensity control as a learning-to-rank problem and explicitly models global intensity ordering under fixed transcripts. We further introduce **ESD-plus**, a multi-speaker emotional speech dataset with explicit intensity variations to support systematic training and evaluation. Extensive automatic and human evaluations demonstrate that Emo-LiPO consistently outperforms strong supervised and DPO-based LLM TTS baselines in emotion accuracy and intensity controllability, while maintaining high speech quality. Overall, this work highlights the importance of listwise preference modeling for fine-grained emotional control and provides a principled direction for expressive and controllable speech generation.

Ethical Statement

There are no ethical issues.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No. 62271432), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams(Grant No. 2023ZT10X044), Shenzhen Science and Technology Program(Shenzhen Key Laboratory, Grant No. ZDSYS20230626091302006), and the Shenzhen Stability Science Program 2023. This research was also supported by the Internal Project of the Shenzhen Research Institute of Big Data (SRIBD), the internal project of the Guangdong Provincial Key Laboratory of Big Data Computing under Grant B10120210117-KP02 (The Chinese University of Hong Kong, Shenzhen), and Shenzhen Key Lab of Multi-Modal Cognitive Computing.

Contribution Statement

Yihang Lin and Li Zhou are co-first authors and contributed equally.

References

[Cho *et al.*, 2024] 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. *arXiv preprint arXiv:2406.07803*, 2024.

[Cho *et al.*, 2025] Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. Emosphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Transactions on Affective Computing*, 2025.

[Du *et al.*, 2024a] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

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

[Ethayarajh *et al.*, 2024] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[Gao *et al.*, 2025] Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[Guo *et al.*, 2023a] Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Guo *et al.*, 2023b] Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: 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, 2023.

[Im *et al.*, 2022] Chae-Bin Im, Sang-Hoon Lee, Seung-Bin Kim, and Seong-Whan Lee. Emoq-tts: Emotion intensity quantization for fine-grained controllable emotional text-to-speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6317–6321. IEEE, 2022.

[Lei *et al.*, 2022] Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:853–864, 2022.

[Leng *et al.*, 2023] Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, et al. Prompttts 2: Describing and generating voices with text prompt. *arXiv preprint arXiv:2309.02285*, 2023.

[Li *et al.*, 2024] Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *International Conference on Machine Learning*, pages 29128–29163. PMLR, 2024.

[Li *et al.*, 2025] Haoxun Li, Yu Liu, Yuqing Sun, Hanlei Shi, Leyuan Qu, and Taihao Li. Emorl-tts: Reinforcement learning for fine-grained emotion control in llm-based tts. *arXiv preprint arXiv:2510.05758*, 2025.

[Liang *et al.*, 2025] Shixiong Liang, Ruohua Zhou, and Qingsheng Yuan. Ece-tts: A zero-shot emotion text-to-speech model with simplified and precise control. *Applied Sciences*, 15(9):5108, 2025.

[Liu *et al.*, 2025a] Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. Lipo: Listwise preference optimization through learning-to-rank. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2404–2420, 2025.

[Liu *et al.*, 2025b] Zhijun Liu, Dongya Jia, Xiaoqiang Wang, Chenpeng Du, Shuai Wang, Zhuo Chen, and Haizhou Li. Direct preference optimization for speech autoregressive diffusion models. *arXiv preprint arXiv:2509.18928*, 2025.

[Ma *et al.*, 2024] 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 the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, 2024.

[Mittag *et al.*, 2021] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*, 2021.

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

[Rafailov *et al.*, 2023] 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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[Reddy *et al.*, 2021] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE, 2021.

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

[Schnell and Garner, 2021] Bastian Schnell and Philip N Garner. Improving emotional tts with an emotion intensity input from unsupervised extraction. In *Proc. 11th ISCA Speech Synth. Workshop*, pages 60–65, 2021.

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

[Song *et al.*, 2024] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998, 2024.

[Wang *et al.*, 2023] Shijun Wang, Jón Gunason, and Damian Borth. Fine-grained emotional control of text-to-speech: Learning to rank inter-and intra-class emotion intensities. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Wu *et al.*, 2024] Haibin Wu, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Daniel Tompkins, Chung-Hsien Tsai, Canrun Li, Zhen Xiao, Sheng Zhao, Jinyu Li, 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)*, pages 690–697. IEEE, 2024.

[Yang *et al.*, 2025a] Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, et al. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10748–10757, 2025.

[Yang *et al.*, 2025b] Qing Yang, Zhenghao Liu, Junxin Wang, Yangfan Du, Pengcheng Huang, and Tong Xiao. Rlaif-spa: Optimizing llm-based emotional speech synthesis via rlaif. *arXiv preprint arXiv:2510.14628*, 2025.

[Zhang *et al.*, 2024a] Chen Zhang, Chengguang Tang, Dading Chong, Ke Shi, Guohua Tang, Feng Jiang, and Haizhou Li. TS-align: A teacher-student collaborative framework for scalable iterative finetuning of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8926–8946, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Zhang *et al.*, 2024b] 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*, 37:50343–50360, 2024.

[Zhang *et al.*, 2025a] Chen Zhang, Dading Chong, Feng Jiang, Chengguang Tang, Anningzhe Gao, Guohua Tang, and Haizhou Li. Aligning language models using follow-up likelihood as reward signal. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25832–25841, Apr. 2025.

[Zhang *et al.*, 2025b] Shaozuo Zhang, Ambuj Mehrish, Yingting Li, and Soujanya Poria. Proemo: Prompt-driven text-to-speech synthesis based on emotion and intensity control. *arXiv preprint arXiv:2501.06276*, 2025.

[Zhao *et al.*, 2023] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

[Zhou *et al.*, 2022a] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18, 2022.

[Zhou *et al.*, 2022b] Kun Zhou, Berrak Sisman, Rajib Rana, Björn W Schuller, and Haizhou Li. Emotion intensity and its control for emotional voice conversion. *IEEE Transactions on Affective Computing*, 14(1):31–48, 2022.

[Zhou *et al.*, 2026] Li Zhou, Hao Jiang, Junjie Li, Tianrui Wang, and Haizhou Li. Emoshift: Lightweight activation steering for enhanced emotion-aware speech synthesis. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17262–17266, 2026.