DR.EMO: DE-BIASED REWARD OPTIMIZATION FOR EMOTIONALLY EXPRESSIVE TEXT-TO-SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Achieving natural and controllable emotional expression remains a key challenge in text-to-speech (TTS) synthesis. Reinforcement learning offers a promising approach through emotion reward optimization. However, such optimization relies on reward models that can reliably recognize emotions in synthesized speech. Since human-verified emotional labels for synthesized speech are scarce, existing methods typically employ speech emotion recognition (SER) models trained on real human speech as reward models. Yet, we identify a substantial misalignment between such models and human judgments on synthesized speech. To address this issue, we curate an open-source dataset of manually filtered synthesized speech with clear emotional expression and train a de-biased emotion reward model that better aligns with human judgments. Based on this model, we introduce DR.EMO, a multi-objective group relative policy optimization framework that enhances emotional expressiveness while preserving general TTS capabilities. Experiments demonstrate that DR.EMO achieves state-of-the-art performance in conveying target emotions, with over 90% of synthesized utterances exhibiting clear human-evaluated emotional expression. Our findings highlight the importance of reliable, de-biased rewards for effective emotional TTS training.[1]

1 INTRODUCTION

Driven by the rapid evolution of generative neural networks, text-to-speech (TTS) synthesis has achieved remarkable milestones (Ren et al., 2019; Li et al., 2023; Chen et al., 2025c; Wang et al., 2025c; Lee et al., 2025). In particular, neural codec token language modeling has emerged as a leading approach for speech generation with the emergence of Large Language Models (Du et al., 2024; Chen et al., 2024b; Anastassiou et al., 2024; Chen et al., 2025b; Nishimura et al., 2025). Despite the rich semantic and paralinguistic features inherent in codec tokens, it remains challenging to accurately predict tokens that harmonize correct content with nuanced emotional expression. This limits the ability of current TTS models to synthesize expressive speech that mirrors human interaction.

Recent studies have explored post-training methods to improve emotional expressiveness in TTS models (Yang et al., 2025; Zhou et al., 2026; Gao et al., 2025c; Chen et al., 2026). Off-policy controllable TTS approaches, represented by supervised fine-tuning (SFT), are known for training stability and data efficiency (Yang et al., 2025; Zhou et al., 2026; Gao et al., 2025c). However, off-policy post-training often suffers from catastrophic forgetting and limited generalization (Agarwal et al., 2024; Lu & Lab, 2025). In contrast, on-policy methods based on reinforcement learning (RL) sample from the model's own predictive distribution, enabling exploration of regions in the distribution tail that are typically neglected by SFT (Gao et al., 2025a; Chen et al., 2026).

However, a critical limitation of existing RL methods is that they overlook the gap between human and synthesized speech (Ge et al., 2026; Wang et al., 2025a). RL relies on reward signals to determine the optimization direction of the policy model. While current emotion controllable RL methods typically use the predictions of Speech Emotion Recognition (SER) models as reward signals, Ge et al. (2026) show that SER models trained on real human speech distributions fail to generalize effectively to synthesized speech distributions. Consequently, using SER models as reward signals or evaluation metrics yields biased estimates of human emotional preferences (An et al., 2024; Wang et al., 2025b; Ji et al., 2025; Chen et al., 2025a; Zhang et al., 2025b; Zhao et al., 2026).

[1] https://anonymous.4open.science/r/DrEMO-0926/README.md

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

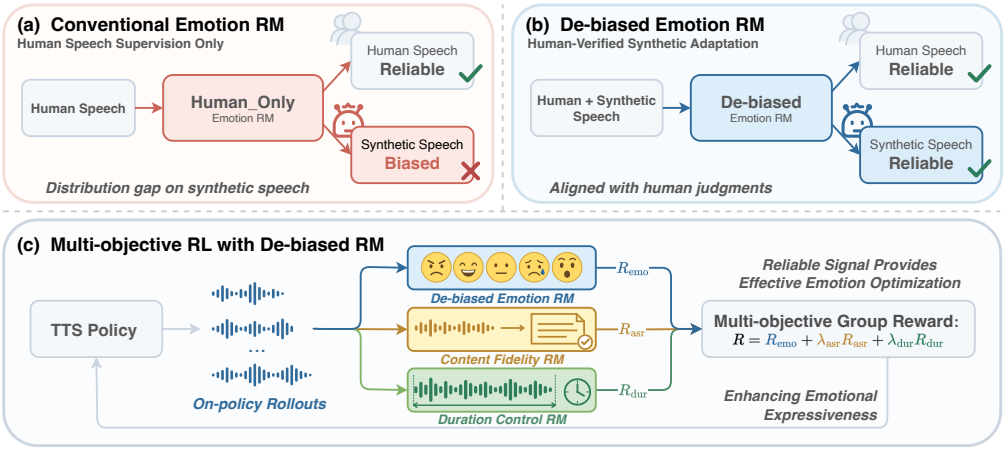


Figure 1: Overview of DR.EMO: reward model debiasing and multi-objective GRPO for emotional TTS. (a) Emotion reward models trained on human speech produce biased predictions on synthesized speech because of the distribution gap. (b) Training on mixed human and synthesized speech yields a de-biased reward model. (c) Multi-objective GRPO combines emotion, ASR, and duration rewards to improve emotional expressiveness while preserving general TTS capabilities.

After identifying the bias in SER reward models, we propose DR.EMO, a reward de-biasing and multi-objective GRPO framework that provides more reliable reward guidance for emotional TTS, as illustrated in Figure 1. Specifically, we first curate an open-source synthesized dataset with explicit emotional expression to develop Qwen2.5-Omni-SER, a de-biased reward model whose agreement with human judgments exceeds that of off-the-shelf SER models by over 25 percentage points. We then integrate this de-biased model into a multi-objective GRPO framework for effective emotional expression optimization. DR.EMO achieves stronger emotional expressiveness while preserving general TTS capabilities across emotion categories. Human evaluation further confirms this improvement, with 91.1% and 91.7% of utterances successfully expressing the target emotion on two backbones, respectively, achieving state-of-the-art performance and improving upon their original counterparts by at least 29 percentage points. In all, our contributions are:

- We identify the misalignment between SER predictions and human judgments of emotional expression in synthesized speech. The biased reward signals provide unreliable guidance for emotional TTS training, limiting emotional expressiveness.
- We curate an open-source dataset of 8K manually filtered synthesized utterances with clear emotional expression, supporting reliable reward model adaptation to synthesized speech.
- We develop Qwen2.5-Omni-SER, a de-biased emotion reward model trained on real and curated synthesized speech, which better aligns with human judgments, provides more reliable reward guidance, and generalizes to unseen TTS systems.
- We introduce DR.EMO, a multi-objective GRPO framework that jointly optimizes emotion, automatic speech recognition (ASR), and duration rewards. It achieves state-of-the-art human-evaluated emotional expressiveness across two TTS backbones while largely preserving general TTS capabilities.

2 Related Work

2.1 Emotion Controllable Speech Synthesis

The goal of text-to-speech is to synthesize intelligible and natural speech, with emotion playing a crucial role (Tan et al., 2021; Xie et al., 2025). Emotion controllable TTS models can broadly be categorized into: (i) *Zero-shot Synthesis* facilitates controllable synthesis by leveraging the paralin-

guistic attributes of a reference audio, such as emotion and timbre, akin to the mechanisms of voice conversion (Mohammadi & Kain, 2017; Sisman et al., 2020). (ii) *Description-based Synthesis* utilizes additional category labels or natural language for emotion control. Natural language allows users to exercise more nuanced and fine-grained emotion control over the synthesized speech. The advanced TTS system combines two emotion control methods (Zhu et al., 2026), yet it still faces the problem of insufficiently expressive emotional delivery. Therefore, recent studies also explore post-training methods to improve emotional expressiveness in TTS models.

2.2 RL IN TEXT-TO-SPEECH SYNTHESIS

Reinforcement learning is a promising preference alignment method in the field of LLMs (Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024) and has garnered significant attention in text-to-speech tasks in recent years. Current works explore leveraging Direct Preference Optimization (DPO) (Rafailov et al., 2023) to improve overall speech quality in TTS systems, including intelligibility, naturalness, and speaker consistency (Anastassiou et al., 2024; Hussain et al., 2025; Tian et al., 2025; Zhang et al., 2025a). Emo-DPO further extends DPO to enhance fine-grained emotional controllability in speech synthesis (Gao et al., 2025c). UNO (Chen et al., 2024a) and RIO (Hu et al., 2024) improve the DPO algorithm by incorporating subjective uncertainty in reward signals and introducing self-supervised preference signals that eliminate the need for manual annotations, respectively. In addition, CosyVoice3 (Du et al., 2025) and DiffRO (Gao et al., 2025a) directly compute differentiable rewards on speech tokens instead of relying on synthesized audio, improving efficiency and stability in large-scale text-to-speech model post-training.

Compared to DPO, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) enables on-policy, group-relative optimization without requiring a separate learned value model, while also improving sample efficiency during policy training. Building on this, Gao et al. (2025b) explores differentiable reward-based GRPO to further enhance TTS performance, while FlexiVoice (Chen et al., 2026) extends GRPO to a multi-objective framework for more advanced instruction-following capabilities across diverse and competing generation objectives.

However, recent works using SER models trained on human speech as the reward model often overlook the gap in emotion understanding between human and synthesized speech (Ge et al., 2026), resulting in biased estimates of true human emotional preferences.

3 EMOTION REWARD MODELS ADAPTED TO SYNTHESIZED SPEECH

3.1 MISALIGNMENT BETWEEN SER PREDICTIONS AND HUMAN JUDGMENTS

Existing RL-based TTS methods often use speech emotion recognition (SER) models trained on real human speech as reward models, implicitly assuming that they generalize to synthesized speech. However, the distribution gap between real and synthesized speech may undermine their reliability (Ge et al., 2026). We therefore examine whether commonly used SER models agree with human judgments of emotional expression on synthesized speech.

We select 200 distinct texts and synthesize each with CosyVoice2 (Du et al., 2024) under five target emotions: *angry*, *happy*, *neutral*, *sad*, and *surprised*, yielding a held-out set of 1K utterances. Human annotators then judge whether each utterance successfully expresses its target emotion. We evaluate Emotion2vec (Ma et al., 2024), a widely used SER model for TTS evaluation and RL alignment, and Qwen2.5-Omni-3B (Xu et al., 2025), an open-source multimodal model with strong speech understanding. For each emotion, samples are divided into *positive* and *negative* subsets according to human judgments, denoting whether the target emotion is successfully expressed.

As demonstrated in Table 1 (I), neither off-the-shelf model aligns reliably with human judgments. Emotion2vec achieves only 58.7% average agreement, despite its widespread use in emotional TTS evaluation and RL training, while Qwen2.5-Omni reaches just 36.4%. These results suggest that SER models trained primarily on human speech do not reliably generalize to synthesized speech, and that general speech understanding does not necessarily translate into robust emotion recognition. Accordingly, using such models directly as rewards may provide unreliable supervision for RL-based emotional TTS training, highlighting the need for reward models that better align with human perception of emotional expression in synthesized speech.

3.2 Reward De-biasing for Synthesized Speech

We therefore adapt both SER backbones to synthesized emotional speech for more reliable reward modeling. In this work, we use *bias* to refer to the systematic misalignment between SER predictions and human judgments on synthesized speech, and define *de-biasing* as adapting the reward model to better reflect human-perceived emotional expression in generated TTS outputs.

To obtain reliable and human-grounded supervision, we construct a human-verified training set containing 8K synthesized utterances with clear emotional expression and 4K real human speech utterances. Since existing SER models are themselves unreliable on synthesized speech, directly using their predictions as labels could propagate the same misalignment into subsequent reward model training. We therefore manually screen the synthesized data to ensure that each retained utterance clearly expresses its target emotion. Both subsets are balanced across the five emotion categories and do not overlap with the 1K held-out evaluation set described in Section 3.1.

We then update the speech encoders together with their SER prediction heads during adaptation. Specifically, we unfreeze the embedding parameters of Emotion2vec (Ma et al., 2024) and the Whisper encoder of Qwen2.5-Omni (Xu et al., 2025), along with their corresponding prediction heads. Jointly optimizing the encoders and prediction heads helps the models capture emotion-relevant prosodic patterns in synthesized speech and better align their predictions with human judgments.

We evaluate both adapted models on the same 1K held-out set constructed in Section 3.1 for a direct comparison. As shown in Table 1 (II), Emotion2vec-SFT achieves 63.9% overall agreement, improving by 5.2 percentage points over its off-the-shelf counterpart. Qwen2.5-Omni-SER achieves 84.1% agreement, corresponding to improvements of 47.7 percentage points over the original Qwen2.5-Omni and 25.4 percentage points over off-the-shelf Emotion2vec. We therefore adopt Qwen2.5-Omni-SER as the final emotion reward model in DR.EMO.

Table 1: Emotion consistency (%) between SER predictions and human judgments on synthesized speech. We compare Emotion2vec (Ma et al., 2024) and Qwen2.5-Omni (Xu et al., 2025) before (I) and after (II) reward de-biasing. For each target emotion, Positive denotes utterances judged to express the target emotion, where consistency requires predicting that emotion; Negative denotes unsuccessful utterances, where consistency requires not predicting the target emotion.

	Category	Angry	Happy	Neutral	Sad	Surprised
I. Vanilla Speech Emotion Recognition Models						
Emotion2vec	Positive	72.4	57.0	74.2	29.2	18.3
	Negative	100.0	68.1	0.0	90.6	95.9
Qwen2.5-Omni	Positive	9.7	11.7	80.3	1.2	1.6
	Negative	100.0	80.6	0.0	96.9	100.0
II. Fine-tuning SER Models with Synthesized Speech						
Emotion2vec-SFT	Positive	82.7	34.4	84.3	31.0	48.4
	Negative	50.0	83.3	0.0	90.6	83.8
Qwen2.5-Omni-SER	Positive	100.0	78.1	96.0	82.7	92.1
	Negative	25.0	50.0	50.0	53.1	60.8

4 Learning Emotional Expression Through De-biased Rewards

4.1 Multi-objective Reward Optimization

GRPO. Building on the de-biased emotion reward model introduced in Section 3.2, DR.EMO optimizes emotional TTS with multi-objective Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Given an input text D_T and emotion control signal C, the TTS policy S_θ samples a group of K candidates, $D_S^{(k)} \sim S_\theta(\cdot \mid D_T, C)$, where $k = 1, \ldots, K$. Each candidate is assigned a

multi-objective reward $r^{(k)}$, from which GRPO computes the group-relative advantage:

$$A^{(k)} = \frac{r^{(k)} - \mathrm{mean}(\{r^{(j)}\}_{j=1}^{K})}{\mathrm{std}(\{r^{(j)}\}_{j=1}^{K})}. \tag{1}$$

The policy is then optimized using the standard GRPO objective with clipping and KL regularization against the reference policy. As we adopt GRPO without modifying its optimization procedure, we focus below on the reward design for emotional TTS.

Rewards. In the GRPO stage, we instantiate an emotion reward for emotional expressiveness and an ASR reward for content fidelity, which are described as follows.

(1) De-biased Emotion Reward. The emotion reward encourages the generated speech to express the target emotion C. We use Qwen2.5-Omni-SER from Section 3.2 as the reward model R_ϕ, which predicts $\hat{C} = \arg\max_c R_\phi(c \mid D_S)$. We then assign a discrete reward based on whether the target emotion is correctly recognized:

$$R_{\mathrm{emo}}(D_S, C) = \begin{cases} +3, & \hat{C} = C, \\ -1, & \text{otherwise.} \end{cases} \tag{2}$$

This discrete formulation reduces sensitivity to confidence calibration, while the asymmetric scale emphasizes successful target-emotion realization.

(2) ASR Reward. Optimizing emotional expression alone may encourage exaggerated acoustic patterns that improve emotion recognition while degrading linguistic fidelity. We therefore introduce an ASR-based reward to ensure that emotional optimization preserves the input content. Specifically, we use SenseVoiceSmall [2] as the ASR model A to transcribe the generated speech as $\hat{D}_T = A(D_S)$, and define

$$R_{\mathrm{asr}}(D_T, \hat{D}_T) = \max\left(0, 1 - \mathrm{WER}(D_T, \hat{D}_T)\right). \tag{3}$$

This reward provides graded supervision for content fidelity, decreasing as transcription errors increase. In addition, when $R_{\mathrm{asr}} < 0.5$, indicating severe content corruption, we override the emotion reward and replace R_{emo} with -2 to prevent emotionally expressive but linguistically incorrect generations from receiving a positive optimization signal.

4.2 Duration Reward in Emotional TTS Optimization

Although the ASR reward encourages content fidelity, it also introduces an undesirable optimization shortcut. The policy tends to reduce the speaking rate, making the resulting synthesized speech easier for the ASR model to transcribe. However, this produces overlong utterances and trailing speech (Gao et al., 2025b). We therefore introduce a group-relative duration reward that penalizes unusually long or short candidates within each sampled group.

For the K speech candidates sampled for the same input (D_T, C), let $l_S^{(k)}$ denote the duration of the k-th utterance and let

$$\widetilde{l}_S = \mathrm{median}\left(\left\{l_S^{(j)}\right\}_{j=1}^{K}\right) \tag{4}$$

denote the median duration of the sampled group. Following Gao et al. (2025b), we define the duration reward as

$$R_{\mathrm{dur}}\left(D_S^{(k)}\right) = -\left|\frac{l_S^{(k)} - \widetilde{l}_S}{\widetilde{l}_S}\right|. \tag{5}$$

This relative formulation is scale-normalized by the group-median duration and penalizes candidates whose durations deviate substantially from the typical duration within the sampled group. In particular, it suppresses abnormally long generations associated with slow speech and trailing artifacts without imposing a fixed duration across utterances of different lengths. Finally, we combine the three reward signals as $R = R_{\mathrm{emo}} + \lambda_{\mathrm{asr}} R_{\mathrm{asr}} + \lambda_{\mathrm{dur}} R_{\mathrm{dur}}$. We set the coefficient of R_{emo} to 1 and tune the remaining coefficients relative to its reward scale. We set $\lambda_{\mathrm{asr}} = 1$ and determine λ_{dur} through the ablation study in Appendix A. This configuration is used in all subsequent experiments.

[2] https://huggingface.co/FunAudioLLM/SenseVoiceSmall

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and Baselines. We apply DR.EMO to the official CosyVoice2-0.5B and CosyVoice3-0.5B backbones (Du et al., 2024; 2025), using their original checkpoints as direct baselines. To isolate the effect of the emotion reward model, we compare DR.EMO with GRPO variants that replace Qwen2.5-Omni-SER with Emotion2vec (Ma et al., 2024) or Qwen2.5-Omni (Xu et al., 2025). We further evaluate IndexTTS2, GPT-4o-TTS, and Gemini-2.5-Pro as external baselines. The prompts for emotion reward models and TTS systems are provided in Appendix B.

Datasets. For reward model SFT, we combine 8K synthesized utterances with 4K real-speech utterances. The synthesized subset is generated using the textual content from LibriTTS-test[3] and SeedTTS-test (Anastassiou et al., 2024). Specifically, we use CosyVoice2 (Du et al., 2024) to synthesize emotional speech conditioned on prompt audio sampled from RAVDESS (Livingstone & Russo, 2018), then manually filter and retain 8K utterances with explicit emotional characteristics. For the real-speech subset, we randomly sample 4K utterances from TESS[4] and CREMA-D[5]. Each subset is balanced across five emotions: *angry*, *happy*, *neutral*, *sad*, and *surprised*.

For GRPO training, we randomly sample 15K training and 3K validation texts from LibriTTS-train[6]. During rollout generation, we pair the training texts with emotional audio prompts sampled from RAVDESS. We follow the same prompt construction procedure used for reward model training. The training set contains 3K instances per emotion. For held-out evaluation, we reuse the 200 held-out texts introduced in Section 3.1 and synthesize each text under all five target emotions, yielding 1K utterances. These texts are sampled from LibriTTS-train without overlap with the training set.

Metrics. We evaluate WER computed by Whisper-Large-V3 (Radford et al., 2022), speaker similarity using the cosine similarity of WavLM-large (Chen et al., 2022) speaker embeddings, and speech quality measured by DNSMOS (Reddy et al., 2021) and UTMOSv2 (Baba et al., 2024). For automatic emotion evaluation, we report E-SIM, the cosine similarity between Emotion2vec (Ma et al., 2024) embeddings of generated speech and prompt audio (Zhang et al., 2025b), and E2V-Acc, the percentage of TTS utterances whose Emotion2vec predictions match the target emotion (Lin et al., 2026). Given the limitations of automatic SER identified in Section 3.1, we use the human-evaluated emotion success rate (ESR) as the primary measure of emotional expressiveness. To improve ESR reliability, each utterance is evaluated by two blinded annotators in randomized order, with independent adjudication of disagreements as detailed in Appendix C.

Hyper-parameters. In the SFT stage, we fine-tune vanilla Emotion2vec and Qwen2.5-Omni for 20 epochs using a batch size of 4 and a learning rate of 1×10^{-5} on 4 NVIDIA A800 GPUs. In the GRPO stage, we optimize the backbone models with a global batch size of 32 and a learning rate of 1×10^{-6} on 2 NVIDIA A800 GPUs. These settings are shared across models in each training stage.

5.2 EXPERIMENTAL RESULTS

5.2.1 MAIN RESULTS

Table 2 compares DR.EMO with the original backbones, alternative emotion reward models, and representative open-source and commercial TTS systems. DR.EMO achieves the highest human-evaluated ESR on both backbones, reaching 91.1% on CosyVoice2 and 91.7% on CosyVoice3. These gains are accompanied by lower WER and higher speaker similarity, but a moderate decrease in UTMOSv2 and a slight decrease in DNSMOS, a trade-off also reported in prior expressive TTS studies (Feng et al., 2026; Lin et al., 2026). Given the sensitivity of automatic MOS prediction to unseen systems and acoustic distributions (Huang et al., 2024), these score reductions may not translate directly into an equivalent degradation in human-perceived naturalness. Overall, DR.EMO

[3] https://openslr.trmal.net/resources/60/test-clean.tar.gz
[4] https://huggingface.co/datasets/myleslinder/tess
[5] https://huggingface.co/datasets/AbstractTTS/CREMA-D
[6] https://openslr.trmal.net/resources/60/train-clean-100.tar.gz

Table 2: Comparison of the backbone models, their GRPO variants trained with different emotion reward models, and representative open-source and commercial TTS systems in terms of speech quality and emotional expressiveness. We report word error rate (WER), speaker similarity (SS), DNSMOS (MOS_D), and UTMOSv2 (MOS_U) for speech naturalness. E-SIM (%) measures the cosine similarity between Emotion2vec embeddings of each synthesized utterance and its corresponding prompt audio. E2V-Acc (%) denotes the percentage of utterances whose Emotion2vec prediction matches the target emotion, while ESR (%) is determined by human evaluation.

Model	Speech Quality				Emotion Expressiveness		
	WER↓	SS↑	MOS_D↑	MOS_U↑	E-SIM↑	E2V-Acc↑	ESR↑
CosyVoice2-0.5B	3.24	45.4	**3.90**	**3.48**	95.8	46.3	61.5
+ Emotion2vec	3.22	**50.9**	3.87	3.19	**96.0**	**85.8**	79.3
+ Qwen2.5-Omni	3.18	46.8	3.87	3.29	95.8	56.5	32.2
+ Qwen2.5-Omni-SER (Ours)	**3.15**	48.9	3.85	3.21	95.9	80.8	**91.1**
CosyVoice3-0.5B	4.01	48.8	**3.91**	**3.47**	96.0	35.8	56.7
+ Emotion2vec	2.98	50.2	3.84	3.16	96.0	**96.1**	83.7
+ Qwen2.5-Omni	3.00	50.2	3.62	3.00	96.0	75.1	74.9
+ Qwen2.5-Omni-SER (Ours)	**2.79**	**50.9**	3.89	3.25	**96.1**	82.0	**91.7**
IndexTTS2	3.48	60.9	3.75	3.08	96.1	50.6	60.5
GPT-4o-TTS	2.99	–	3.82	3.56	–	45.5	41.2
Gemini-2.5-Pro	2.83	–	4.06	3.63	–	50.4	77.7

improves emotional expressiveness while preserving content fidelity and speaker identity, with some reduction in predicted naturalness.

Despite these improvements, SER-based evaluation metrics do not consistently reflect human judgments. For CosyVoice2, the Emotion2vec-reward variant achieves a higher E2V-Acc than DR.EMO (85.8% vs. 80.8%), yet its human-evaluated ESR is considerably lower (79.3% vs. 91.1%). A similar ranking reversal appears on CosyVoice3, where Emotion2vec reaches 96.1% E2V-Acc compared with 82.0% for DR.EMO, despite lower human-evaluated ESR. This shows that stronger agreement with Emotion2vec does not necessarily indicate stronger human-perceived emotional expression. E-SIM shows a different limitation, remaining tightly clustered between 95.8 and 96.1 across systems even when ESR varies substantially. For example, among the CosyVoice2 variants, ESR ranges from 32.2% to 91.1% while E-SIM changes by only 0.2 points. Together, these results reveal distinct limitations in both classification-based and embedding-based automatic emotion metrics, reinforcing the importance of human evaluation for emotional expressiveness.

Beyond comparisons within each backbone, DR.EMO achieves higher human-evaluated ESR than the other evaluated TTS systems. The DR.EMO-trained CosyVoice2 and CosyVoice3 models exceed all external baselines in ESR, and the CosyVoice3 variant also attains the lowest WER in Table 2. These results demonstrate that DR.EMO enhances emotional expressiveness while maintaining competitive content fidelity and speech naturalness.

5.2.2 EMOTION-WISE EXPRESSIVENESS ACROSS TTS SYSTEMS

Figure 2 compares emotion-wise ESR across the official backbones, the backbones post-trained with DR.EMO, and representative open-source and commercial TTS systems. DR.EMO improves expressiveness across all non-neutral emotions while maintaining the already high ESR on neutral speech. Despite differences in their initial performance, the two backbones exhibit similar emotion-wise ESR profiles after training, suggesting that the benefits of reward debiasing extend across backbone models. Consistent with these quantitative improvements, the DR.EMO-trained models produce substantially more explicit and recognizable emotional speech than the official backbones.

Compared with IndexTTS2, GPT-4o-TTS, and Gemini-2.5-Pro, the DR.EMO-trained models achieve stronger overall emotional expressiveness and more consistent performance across emotion categories. These results complement the aggregate ESR in Table 2, showing that the overall improvements extend across multiple emotions rather than being concentrated in a single category.

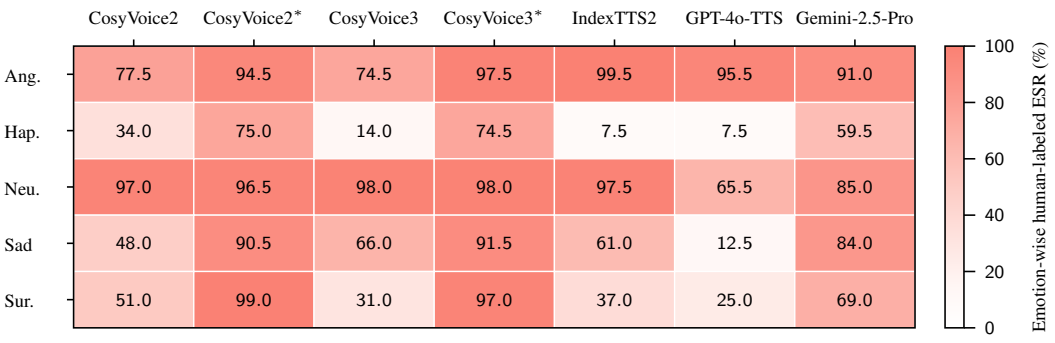


Figure 2: Emotion-wise human-labeled ESR (%) for the official CosyVoice2 and CosyVoice3 backbones, representative open-source and commercial TTS models, and our Dr.EMO-trained variants. CosyVoice2* and CosyVoice3* denote models post-trained with DR.EMO, which uses Qwen2.5-Omni-SER as the emotion reward model. Darker colors indicate better emotional expressiveness.

To further examine the acoustic changes associated with these improvements, Appendix D provides matched mel-spectrogram analyses of prosodic characteristics beyond overall energy differences.

6 GENERALIZATION ANALYSIS

6.1 GENERALIZATION OF DE-BIASED REWARD MODEL

A reliable emotion reward model should transfer to synthesized speech from unseen TTS systems. Previous studies have reported limited out-of-domain generalization of Emotion2vec (Lin et al., 2026; Ge et al., 2026). We therefore examine whether our de-biased reward model can accurately recognize emotions in speech generated by unseen TTS systems.

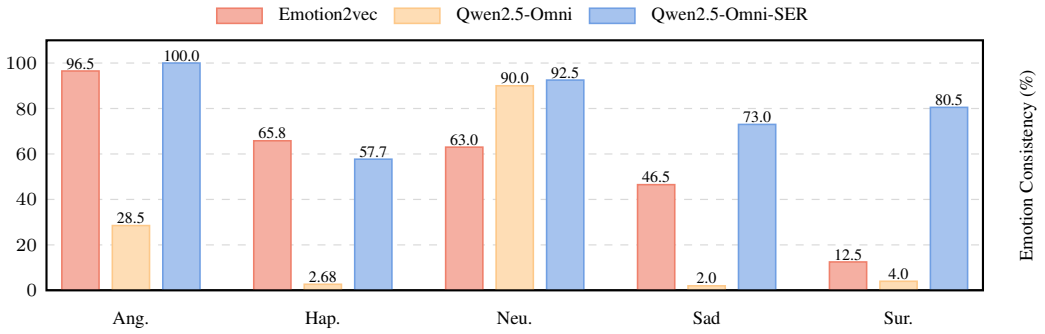


Figure 3: Emotion-wise generalization of emotion reward models on the human-labeled *Positive* subset of speech synthesized by IndexTTS2, GPT-4o-TTS, and Gemini-2.5-Pro. This subset retains only utterances judged by annotators to successfully express the target emotion. Emotion consistency is the percentage of predictions matching the target emotion. Higher is better.

We construct an out-of-domain test set using speech generated by IndexTTS2, GPT-4o-TTS, and Gemini-2.5-Pro in Table 2. None of these systems contributes speech to reward model training. Following the definition of *Positive* in Section 3.1, we retain up to 200 utterances per emotion from the human-labeled *Positive* subset. Since only 149 *happy* utterances satisfy this criterion, the resulting test set contains 949 utterances. This positive-only evaluation isolates conditional emotion-recognition accuracy from generation failures. We evaluate Emotion2vec, vanilla Qwen2.5-Omni, and Qwen2.5-Omni-SER without further adaptation, and measure emotion consistency as the percentage of predictions matching the human-labeled target emotion.

Figure 3 shows that Qwen2.5-Omni-SER achieves 81.98% sample-weighted overall consistency, compared with 56.38% for Emotion2vec. It outperforms vanilla Qwen2.5-Omni in all five emotion categories and achieves the highest consistency in four. Its gains over Emotion2vec are largest for *sad* and *surprised* speech, reaching 73.0% and 80.5%, respectively. Qwen2.5-Omni-SER underperforms Emotion2vec on *happy* speech (57.7% vs. 65.8%), indicating a remaining category-specific limitation. Nevertheless, its overall and per-category results show that reward model adaptation substantially improves emotion recognition on speech generated by unseen TTS systems.

6.2 General TTS Capabilities Retention

Beyond reward model generalization, we evaluate whether Dr.EMO preserves the general TTS capabilities of the policy on out-of-domain benchmarks. We evaluate the official CosyVoice2, CosyVoice3 and their Dr.EMO-trained checkpoints on the English subsets of Seed-TTS-Eval[7] and CV3-Eval (Du et al., 2025) without benchmark-specific adaptation. Following their protocols, WER is computed using Whisper-Large-V3 for both benchmarks. Seed-tts-eval measures speaker similarity using WavLM-large, whereas CV3-Eval uses ERes2Net (Chen et al., 2024c) embeddings and additionally evaluates audio quality with DNSMOS (Reddy et al., 2021). We omit the official CV3-Eval emotional voice cloning metric because it relies on Emotion2vec (Ma et al., 2024), which is not reliably aligned with human judgments of synthesized speech as demonstrated in Section 3.1.

As illustrated in Table 3, Dr.EMO broadly preserves general synthesis quality across both backbones. WER consistently decreases in all comparisons, covering zero-shot TTS, voice conversion, and cross-lingual synthesis. On Seed-TTS-Eval, both models achieve lower WER with comparable SS. On CV3-Eval, CosyVoice2 lowers WER by 0.75 and 1.50 percentage points on TTS-EN and cross-lingual synthesis, respectively, while slightly improving speaker similarity. CosyVoice3 also reduces WER in both settings, although its SS slightly decreases within 0.5 points. These results indicate that Dr.EMO improves emotional expressiveness while largely preserving the policy's general TTS capabilities on held-out benchmarks. Appendix E further evaluates the hard subsets of CV3-Eval, with results that generally support the same conclusion.

Table 3: Out-of-domain evaluation on the English subsets of Seed-TTS-Eval and CV3-Eval. CosyVoice2* and CosyVoice3* denote models post-trained with Dr.EMO, which uses Qwen2.5-Omni-SER as the emotion reward model. Seed-TTS-Eval includes zero-shot TTS (TTS-EN) and zero-shot voice conversion (VC-EN), while CV3-Eval includes zero-shot synthesis (TTS-EN) and cross-lingual synthesis to English (X-ling. $\rightarrow$ EN). MOS denotes the DNSMOS score.

Model	Seed-TTS-Eval				CV3-Eval					
	TTS-EN		VC-EN		TTS-EN			X-ling. $\rightarrow$ EN		
	WER↓	SS↑	WER↓	SS↑	WER↓	SS↑	MOS_D↑	WER↓	SS↑	MOS_D↑
CosyVoice2	3.67	64.5	3.96	64.7	7.68	70.3	3.86	10.49	70.6	3.82
CosyVoice2*	3.64	64.5	3.88	64.6	6.93	70.5	3.86	8.99	70.8	3.83
CosyVoice3	2.36	68.3	2.29	68.8	5.02	74.3	3.83	4.22	64.5	3.81
CosyVoice3*	2.21	68.6	2.14	68.6	4.93	73.8	3.82	3.78	64.2	3.80

7 Conclusions

In this work, we present Dr.EMO to address the misalignment between SER-based emotion rewards and human judgments of synthesized speech. We curate an open-source, human-verified synthesized speech dataset with explicit emotion expression and develop a de-biased reward model tailored to synthesized speech. Dr.EMO integrates this reward model into multi-objective GRPO for emotional TTS training. Experiments demonstrate state-of-the-art emotional expressiveness while largely preserving general TTS capabilities. These findings highlight the importance of reliable, de-biased reward modeling for aligning emotional speech generation with human perception.

[7] https://github.com/BytedanceSpeech/seed-tts-eval

AI USE STATEMENT

We used ChatGPT to assist with writing and language polishing of the manuscript. We did not use generative AI for other research tasks requiring disclosure. The authors reviewed and edited all AI-assisted text to ensure its accuracy and consistency with the research findings, and take full responsibility for the final content of the paper, including its claims and conclusions.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*, 2024.

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

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The T05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, pp. 818–824. IEEE, 2024. doi: 10.1109/SLT61566.2024.10832315. URL https://doi.org/10.1109/SLT61566.2024.10832315.

Chen Chen, Yuchen Hu, Wen Wu, Helin Wang, Eng Siong Chng, and Chao Zhang. Enhancing zero-shot text-to-speech synthesis with human feedback. *arXiv preprint arXiv:2406.00654*, 2024a.

Dekun Chen, Xueyao Zhang, Yuancheng Wang, Kenan Dai, Li Ma, and Zhizheng Wu. Flexivoice: Enabling flexible style control in zero-shot tts with natural language instructions. *arXiv preprint arXiv:2601.04656*, 2026.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. Emova: Empowering language models to see, hear and speak with vivid emotions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5455–5466, 2025a.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://doi.org/10.1109/JSTSP.2022.3188113.

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

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025b.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, Shiliang Zhang, and Junjie Li. Eres2netv2: Boosting short-duration speaker verification performance with computational efficiency, 2024c. URL https://arxiv.org/abs/2406.02167.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Zhao Jian, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025c.

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.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Fangming Feng, Dongjie Fu, Zequn Xie, Yu Zhang, Yangyang Wu, Zhou Zhao, and Tao Jin. Rectifying the emotional flow: Aligning priors and dynamic guidance for high-arousal text-to-speech. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21874–21888, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.998. URL https://aclanthology.org/2026.acl-long.998/.

Changfeng Gao, Zhihao Du, and Shiliang Zhang. Differentiable reward optimization for llm based tts system. *arXiv preprint arXiv:2507.05911*, 2025a.

Changfeng Gao, Yabin Li, Keyu An, Zhifu Gao, Zhihao Du, Han Zhao, and Xiangang Li. Explore the reinforcement learning for the llm based asr and tts system. *arXiv preprint arXiv:2509.18569*, 2025b.

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)*, pp. 1–5. IEEE, 2025c.

Yuan Ge, Haishu Zhao, Aokai Hao, Junxiang Zhang, Bei Li, Xiaoqian Liu, Chenglong Wang, Jianjin Wang, Bingsen Zhou, Bingyu Liu, et al. On the emotion understanding of synthesized speech. *arXiv preprint arXiv:2603.16483*, 2026.

Yuchen Hu, Chen Chen, Siyin Wang, Eng Siong Chng, and Chao Zhang. Robust zero-shot text-to-speech synthesis with reverse inference optimization. *arXiv preprint arXiv:2407.02243*, 2024.

Wen-Chin Huang, Szu-Wei Fu, Erica Cooper, Ryandhimas E. Zezario, Tomoki Toda, Hsin-Min Wang, Junichi Yamagishi, and Yu Tsao. The voicemos challenge 2024: Beyond speech quality prediction, 2024. URL https://arxiv.org/abs/2409.07001.

Shehzeen Samarah Hussain, Paarth Neekhara, Xuesong Yang, Edresson Casanova, Subhankar Ghosh, Roy Fejgin, Mikyas T. Desta, Rafael Valle, and Jason Li. Koel-TTS: Enhancing LLM based speech generation with preference alignment and classifier free guidance. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21219–21234, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1076. URL https://aclanthology.org/2025.emnlp-main.1076/.

Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. ControlSpeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6966–6981, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.346. URL https://aclanthology.org/2025.acl-long.346/.

Keon Lee, Dong Won Kim, Jaehyeon Kim, Seungjun Chung, and Jaewoong Cho. Ditto-tts: Diffusion transformers for scalable text-to-speech without domain-specific factors. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in neural information processing systems*, 36:19594–19621, 2023.

Yihang Lin, Li Zhou, Congwei Cao, Dongchu Xie, Xiaoxue Gao, Chen Zhang, and Haizhou Li. Emo-lipo: Listwise preference optimization for fine-grained emotion intensity control in llm-based text-to-speech, 2026. URL https://arxiv.org/abs/2606.13006.

Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391, 2018. doi: 10.1371/journal.pone.0196391.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

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*, pp. 15747–15760, 2024.

Seyed Hamidreza Mohammadi and Alexander Kain. An overview of voice conversion systems. *Speech Communication*, 88:65–82, 2017.

Yuto Nishimura, Takumi Hirose, Masanari Ohi, Hideki Nakayama, and Nakamasa Inoue. Hall-e: Hierarchical neural codec language model for minute-long zero-shot text-to-speech synthesis. In *The Thirteenth International Conference on Learning Representations*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

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.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pp. 6493–6497. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9414878. URL https://doi.org/10.1109/ICASSP39728.2021.9414878.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

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

Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM transactions on audio, speech, and language processing*, 29:132–157, 2020.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.

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

Cong Wang, Changfeng Gao, Yang Xiang, Zhihao Du, Keyu An, Han Zhao, Qian Chen, Xiangang Li, Yingming Gao, and Ya Li. Rrpo: Robust reward policy optimization for llm-based emotional tts. *arXiv preprint arXiv:2512.04552*, 2025a.

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

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. In *The Thirteenth International Conference on Learning Representations*, 2025c.

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. Towards controllable speech synthesis in the era of large language models: A systematic survey. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 764–791, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.40. URL https://aclanthology.org/2025.emnlp-main.40/.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

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*, pp. 10748–10757, 2025.

Xueyao Zhang, Yuancheng Wang, Chaoren Wang, Ziniu Li, Zhuo Chen, and Zhizheng Wu. Advancing zero-shot text-to-speech intelligibility across diverse domains via preference alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12251–12270, 2025a.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, and Mingbo Ma. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *ICLR*. OpenReview.net, 2025b.

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.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35139–35148, 2026.

Han Zhu, Lingxuan Ye, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhifeng Han, Weiji Zhuang, Long Lin, and Daniel Povey. Omnivoice: Towards omnilingual zero-shot text-to-speech with diffusion language models. *arXiv preprint arXiv:2604.00688*, 2026.

A Ablation of multi-objective rewards

We conduct an ablation study to investigate the effects of the ASR reward and the duration-reward coefficient λ_{dur} during GRPO training. All ablation models are trained using the same 15K training data described in Section 5.1 and otherwise identical training settings. Except for the ablation without the ASR reward, all training configurations differ only in λ_{dur}.

For evaluation, we report WER computed by Whisper-Large-V3, Average Duration Ratio (ADR), Prolonged Speech Rate (PSR), and mean opinion score (MOS) predicted by UTMOSv2. To assess duration consistency, we define ADR as the average ratio of synthesized speech duration to its corresponding real reference:

$$\text{ADR} = \frac{1}{N} \sum_{i=1}^{N} \frac{d_i^{\text{gen}}}{d_i^{\text{ref}}}, \tag{6}$$

where d_i^{gen} and d_i^{ref} denote the durations of the synthesized and reference speech, respectively. An ADR closer to 1 indicates better overall duration consistency. To further capture severe speech prolongation, we define PSR as the percentage of synthesized utterances whose duration exceeds 1.5 times that of their corresponding reference speech.

Table 4: Ablation study of the ASR reward and duration-reward coefficient λ_{dur}. ADR denotes the average ratio between the duration of generated speech and that of the reference speech. PSR(%) denotes the percentage of generated utterances whose duration exceeds 1.5 times that of the reference speech. – indicates metrics not reported because the generated speech is largely unintelligible.

Setting	**WER↓**	**ADR→ 1**	**PSR↓**	**MOS↑**
R_{emo}	78.95	–	–	–
$R_{\text{emo}} + R_{\text{asr}}$	3.62	1.16	7.7	3.19
$+R_{\text{dur}}$ ($\lambda_{\text{dur}} = 0.5$)	3.50	1.13	3.5	3.17
$+R_{\text{dur}}$ ($\lambda_{\text{dur}} = 1.0$)	3.39	1.13	0.0	3.18

Table 4 demonstrates that the ASR reward is essential for preserving content fidelity. Optimizing only R_{emo} increases WER to 78.95%, making the generated speech largely unintelligible. We therefore include R_{asr} in all remaining configurations. Increasing λ_{dur} effectively mitigates abnormal speech prolongation without compromising content fidelity or perceptual quality. In particular, setting $\lambda_{\text{dur}} = 1.0$ reduces PSR from 7.7% to 0.0%, while maintaining a low WER and an ADR of 1.13. The predicted MOS also remains comparable to that without the duration reward (3.18 vs. 3.19). Based on this trade-off, we set $\lambda_{\text{dur}} = 1.0$ in the final training configuration.

B Prompts for SER and TTS Inference

B.1 Emotion Reward Model Prompts

We use the following prompt for emotion label prediction with both Qwen2.5-Omni and Qwen2.5-Omni-SER during emotional TTS training.

Input: `[AUDIO]`

Given the audio input, classify the speaker's emotion.

Emotion categories (choose exactly one):
– angry – happy – neutral – sad – surprised

Rule: Output only the emotion label in lowercase, without punctuation.

Output format: `<emotion_label>`

B.2 Commercial TTS Model Inference Prompts

GPT-4o-TTS. GPT-4o-TTS accepts the speaking instruction and synthesis text through separate `instructions` and `text` fields. We use the following emotion-specific instructions.

> **Angry:** Speak with a clearly angry, irritated tone, raised voice, and sharp, tense delivery.
>
> **Happy:** Speak in a cheerful, bright, and energetic tone, clearly sounding happy.
>
> **Neutral:** Speak in a neutral, calm, and balanced tone, with no strong emotion.
>
> **Sad:** Use a quiet, heavy voice, as if you are deeply disappointed and emotionally tired.
>
> **Surprised:** Speak with clear surprise and sudden excitement, with noticeable changes in pitch.

Gemini-2.5-Pro. For Gemini-2.5-Pro, we combine the emotion and text in a single instruction.

> Say the following sentence in a `[EMOTION]` tone: `[TEXT]`

C Human Evaluation Protocol and Annotation Guidelines

Annotation task. Human evaluation is used to determine whether each synthesized utterance successfully expresses its target emotion. Annotators are shown the target emotion and assign a binary judgment to each utterance: *Positive* if the target emotion is successfully expressed, and *Negative* otherwise. Judgments are based primarily on perceptible prosodic and paralinguistic characteristics, including pitch, energy, speaking rate, stress, and voice quality, rather than textual sentiment alone.

Emotion-wise guidelines. Annotators are provided with the following reference cues for the five target emotions. These cues are intended to support consistent judgments rather than serve as strict requirements, since emotional realization may vary across speakers and linguistic content.

> **Angry:**
> A forceful and intense delivery, typically characterized by increased vocal energy, a tense voice quality, prominent stress, or a shout-like manner.
>
> **Happy:**
> A positive and animated delivery, often conveyed through a bright voice quality, lively rhythm, rising pitch movements, or an audible smiling or laughter-like quality.
>
> **Neutral:**
> An even and controlled delivery with stable pitch, energy, and speaking rate, without salient emotional emphasis or coloration.
>
> **Sad:**
> A subdued delivery, commonly associated with reduced energy, a slower speaking rate, lower or falling pitch, and a soft, heavy, or sigh-like voice quality.
>
> **Surprised:**
> A sudden and marked prosodic change, such as an abrupt rise in pitch or energy, emphatic stress, or a brief exclamatory delivery that conveys an unexpected reaction.

Annotation and adjudication. For each model and evaluation condition, two annotators independently listen to each synthesized utterance once and assign a label. Annotators were blinded to system identity, and utterances were presented in randomized order. If their judgments agree, the shared annotation is used as the final label. If the two initial judgments disagree, the utterance is independently evaluated by an additional annotator, a fourth annotator is consulted when further adjudication is required.The Emotion Success Rate is then computed from the adjudicated labels as

$$\mathrm{ESR} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[y_i = \mathrm{Positive}] \times 100\%, \tag{7}$$

where N is the number of evaluated utterances and y_i is the final human judgment for the i-th utterance. This two-stage annotation procedure reduces individual subjectivity while ensuring that every disagreement receives an additional independent evaluation.

D Case Study on Emotional Speech

We conduct a matched case study using the same text synthesized under the *Happy* condition across eight model configurations. As shown in Figures 4 and 5, each panel shows an 80-bin mel-spectrogram normalized over $[-80, 0]$ dB, with the red curve denoting F_0. Relative to the corresponding Emotion2vec-reward variant, CosyVoice2 + Qwen2.5-Omni-SER increases median F_0 by 5.6%, F_0 range by 40.2%, and spectral flux by 3.8%. The mel-spectrogram shows greater variation in spectral energy, accompanied by a wider pitch range in the overlaid F_0 contour.

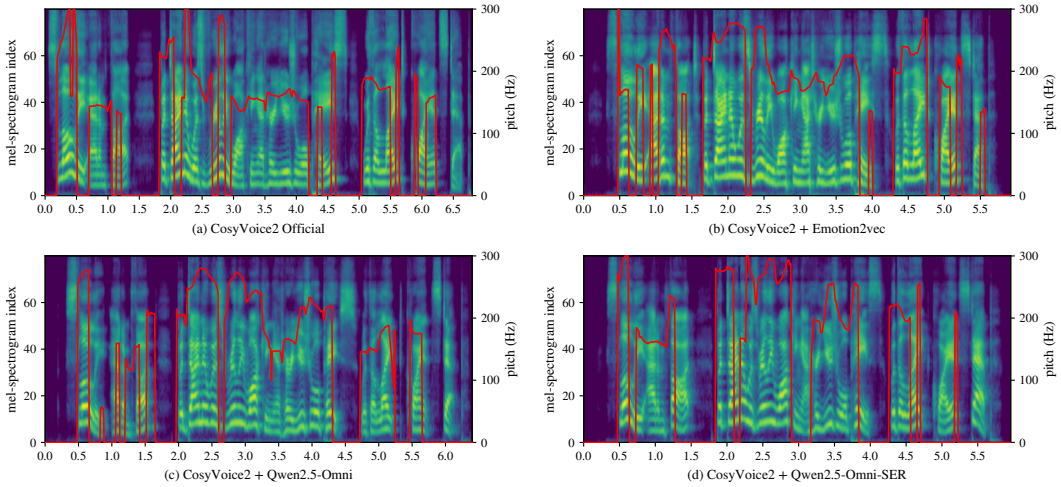


Figure 4: Mel-spectrograms and F_0 contours for the same text synthesized with *Happy* emotion by the CosyVoice2-based models: (a) Official, (b) Emotion2vec, (c) Qwen2.5-Omni, and (d) Qwen2.5-Omni-SER. The red curves represent F_0 according to the right vertical axis.

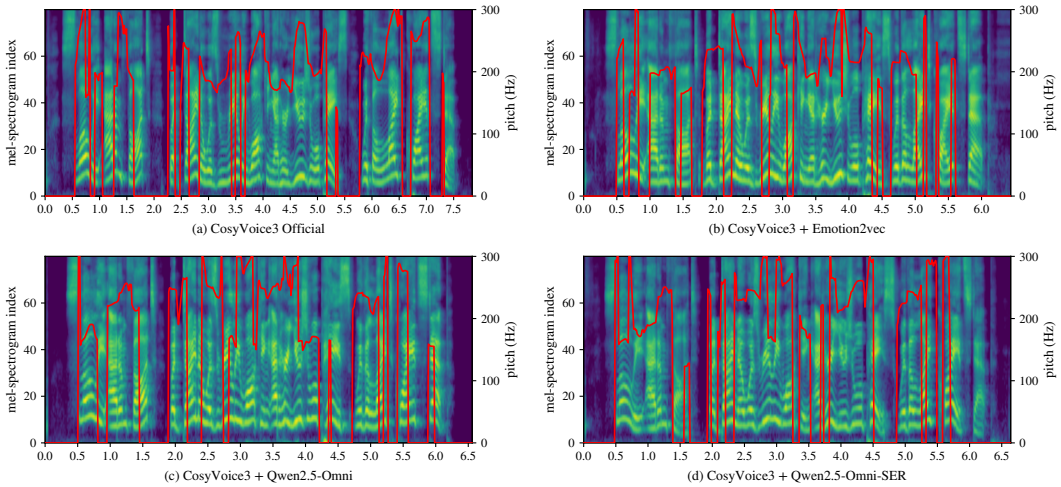


Figure 5: Mel-spectrograms and F_0 contours for the same text synthesized with *Happy* emotion by the CosyVoice3-based models: (a) Official, (b) Emotion2vec, (c) Qwen2.5-Omni, and (d) Qwen2.5-Omni-SER. The red curves represent F_0 according to the right vertical axis.

CosyVoice3 + Qwen2.5-Omni-SER also exhibits greater pitch variation, accompanied by richer high-frequency detail in the mel-spectrogram. Relative to its baseline, it increases median F_0 by

10.0%, F_0 range by 18.8%, and spectral flux by 3.4%. Whereas CosyVoice2 + Qwen2.5-Omni-SER shows pronounced variation in energy and pitch, CosyVoice3 + Qwen2.5-Omni-SER combines pitch variation with finer spectral detail. Together, these observations suggest that the gains in emotional expressiveness are accompanied by complementary acoustic changes across the two backbones.

E RESULTS ON THE CV3-EVAL HARD SUBSETS

Table 5 generally supports Section 6.2, despite minor regressions in specific settings. On hard TTS-EN, WER decreases from 14.3 to 13.7 for CosyVoice2 and from 10.19 to 7.88 for CosyVoice3. For cross-lingual synthesis, CosyVoice2 reduces WER from 11.5 to 10.2, whereas CosyVoice3 shows a slight increase from 5.86 to 6.05. Speaker similarity decreases slightly on TTS-EN but improves in cross-lingual synthesis for both backbones. MOS changes are at most 0.01. The observed regressions remain limited to individual metrics and settings. Together with the emotional expressiveness gains in Section 5.2, these results suggest that general TTS capabilities are largely preserved.

Table 5: Evaluation results on the hard subsets of CV3-Eval. CosyVoice2* and CosyVoice3* denote models post-trained with DR.EMO, which uses Qwen2.5-Omni-SER as the emotion reward model.

Model	**TTS-EN (Hard)**			**X-ling. → EN (Hard)**		
	WER↓	**SS↑**	**MOS↑**	**WER↓**	**SS↑**	**MOS↑**
CosyVoice2	14.3	70.5	3.96	11.5	70.3	3.93
CosyVoice2*	13.7	69.5	3.95	10.2	70.4	3.94
CosyVoice3	10.19	74.8	3.96	5.86	64.5	3.95
CosyVoice3*	7.88	74.4	3.95	6.05	64.7	3.96