From Scores to Preferences: Redefining Evaluation Paradigm for Speech Quality Reward Modeling

Anonymous ACL submission

Abstract

Speech quality assessment (SQA) is typically formulated as a score regression task based on subjective ratings, such as the Mean Opinion Score (MOS), which inherently suffer from inconsistent standards and limit cross-dataset training and evaluation. To address these limitations, we reformulate SQA as a preference-based comparison paradigm and construct **MOS-Pref**, a large-scale MOS-derived preference dataset. Building on MOS-Pref, we systematically implement and evaluate three reward modeling paradigms: scalar, semi-scalar, and generative reward models, alongside existing SQA approaches. Our experiments reveal three key findings: (1) scalar models achieve the strongest overall performance, consistently exceeding 74% accuracy; (2) score regression-based approaches generally underperform preference-based methods in both overall performance and generalization; and (3) all reward models struggle on pairs with very small MOS gap. Motivated by these observations, we propose a MOS-aware GRM design that incorporates MOS gap into the reward function during reinforcement learning. Experimental results show that the MOS-aware GRM significantly improves fine-grained speech quality discrimination. We hope this work fosters more rigorous and scalable research in SQA.

1 Introduction

Assessing the perceptual quality of speech is crucial for guiding the development and refinement of speech generation models (Valentini-Botinhao and Yamagishi, 2018). The rapid progress of text-to-speech (TTS) and generative audio models has significantly improved the naturalness and expressiveness of synthetic speech (Wang et al., 2025b; Xu et al., 2025b), while also introducing new challenges for evaluating speech quality at scale (Lo et al., 2019; Huang et al., 2022).

Speech quality assessment (SQA) has traditionally relied on human subjective ratings such as the Mean Opinion Score (MOS) (Sector, 1996), and models are typically trained to regress these scores (Kondo et al., 2025). Under this formulation, inherent biases in human annotations lead to inconsistent subjective standards across datasets, hindering effective cross-dataset training and evaluation (Pieper and Voran, 2024). Moreover, the restricted joint utilization of multiple datasets leads to a loss of data diversity, as each dataset may cover different speech domains, thereby further constraining the generalization ability of models trained under the conventional MOS regression paradigm.

To address these limitations, we shift the paradigm of SQA from absolute score regression to preference-based comparisons within individual datasets. Instead of predicting MOS values, we reformulate speech quality modeling as a comparative task, where models learn to determine relative perceptual preferences between speech samples. Following this principle, we construct **MOS-Pref**, a large-scale MOS-derived preference dataset that covers diverse speech scenarios and languages, and is annotated with natural-language critiques describing perceptible comparisons in speech quality. This paradigm shift mitigates the inconsistency in subjective standards induced by absolute scores, while providing a unified supervision signal for cross-dataset training and evaluation of SQA models under a consistent comparison setting.

Building on MOS-Pref, we systematically construct and evaluate three paradigms for reward modeling: scalar reward models, semi-scalar reward models, and generative reward models (GRMs). Our analysis shows that scalar models achieve the strongest overall performance, consistently exceeding 74% accuracy and reaching around 80% on average across both in-domain and out-of-domain (OOD) test sets. We further observe that score regression-based approaches generally underperform preference-based methods in terms of both overall accuracy and cross-dataset generalization,

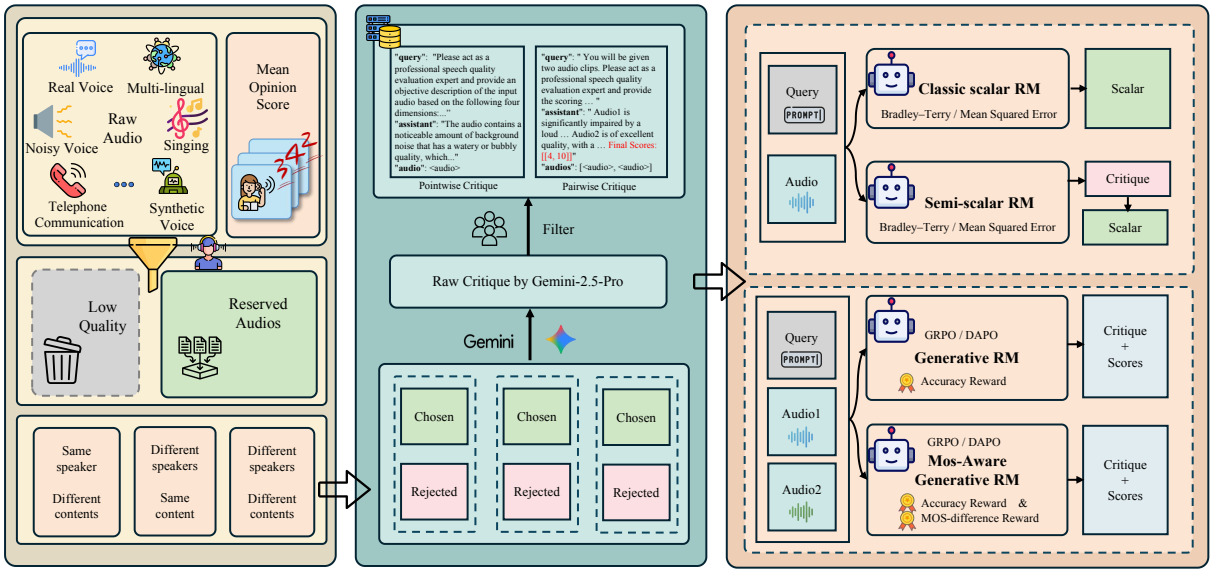


Figure 1: Overview of MOS-Pref and reward modeling paradigms. Data from multiple MOS datasets are filtered and grouped, then converted into pairwise comparisons with natural-language critiques generated by Gemini-2.5-Pro. The resulting dataset supports training and evaluation of reward models in a consistent and reproducible setting.

highlighting the advantage of preference comparison for modeling speech quality judgments. Finally, the performance of reward models degrades notably when distinguishing speech pairs with very small MOS differences, indicating that fine-grained quality discrimination remains a key challenge.

To further improve performance on challenging cases, we propose a MOS-aware GRM design that incorporates MOS gap information into the reward function during reinforcement learning, enabling the model to adaptively scale rewards according to the difficulty of each sample pair. Experimental results show that the MOS-aware GRM consistently improves performance across all evaluation datasets, with accuracy gains of up to 3% on samples with highly similar speech quality.

We hope this work will help establish a unified evaluation framework and provide methodological insights to foster more rigorous and scalable research in automatic speech quality assessment. Overall, our contributions are threefold:

1. We construct MOS-Pref, a large-scale preference dataset with natural-language critiques, covering diverse speech scenarios and supporting consistent training and evaluation.
2. We implement a range of reward modeling paradigms to develop stronger SQA models, and conduct systematic evaluation in comparison with existing methods, providing insights into their relative strengths and limitations.
3. We propose a MOS-aware GRM design that incorporates MOS gap information into the reward function, achieving measurable improvements in fine-grained speech quality discrimination across all evaluation datasets.

2 Related Work

2.1 Automatic Speech Quality Assessment

Research on automatic speech quality assessment has evolved from early CNN- and RNN-based predictors such as MOSNet and Quality-Net (Lo et al., 2019; Fu et al., 2018) to self-supervised learning approaches based on wav2vec 2.0 and HuBERT (Baevski et al., 2020; Hsu et al., 2021), and more recently to Audio Language Models (ALMs) (Wang et al., 2025a; Deshmukh et al., 2024; Chen et al., 2025; Zezario et al., 2025). Recent state-of-the-art MOS prediction systems, including UTMOS (Saeki et al., 2022) and LE-SSL-MOS (Qi et al., 2023), have shown strong results on in-domain and OOD tracks of the VoiceMOS challenges. To address the inconsistencies in subjective ratings across datasets, recent work has explored bias-aware training losses (Mittag et al., 2021c) and dataset score alignment frameworks such as AlignNet (Pieper and Voran, 2024). While these methods improve cross-dataset performance, progress remains limited by heterogeneous MOS annotation protocols, motivating the need for a unified setting for consistent evaluation.

2.2 Advances in Reward Modeling

Recent advances in reward modeling (RM), originally introduced to align model outputs with human preferences (Ouyang et al., 2022), have inspired diverse paradigms in language and vision domains. In the language domain, research has progressed from scalar reward models to generative reward modeling. The Critique-out-Loud (CLoud) framework (Ankner et al., 2024) introduces natural language critiques prior to scalar scoring, bridging generative judgment and reward modeling. More recently, Liu et al. (2025) proposes Self-Principled Critique Tuning (SPCT), enabling GRMs to generate adaptive principles and critiques during inference, achieving state-of-the-art results across RM benchmarks. In the vision domain, RM has been extended to multimodal evaluation, with proprietary systems such as GPT-4V (Achiam et al., 2023) demonstrating strong agreement with human judgments and open-source efforts like LLaVA-Critic (Xiong et al., 2025) unifying pointwise and pairwise scoring. By contrast, research on speech RM remains sparse, with no established frameworks for scalable or fine-grained preference alignment.

3 Dataset

To support systematic training and evaluation of reward models for SQA, we construct MOS-Pref, a large-scale MOS-derived preference dataset. As shown in Figure 1, MOS-Pref integrates diverse MOS datasets into a unified format, addresses inconsistencies in scoring standards by using preference pairs rather than absolute MOS scores, and covers a wide range of speech scenarios. This chapter introduces the dataset in three aspects: data sources, the annotation strategy for preference data, and overall dataset statistics.

3.1 Dataset Source

MOS-Pref is constructed based on several widely used speech quality datasets. Specifically, we select BVCC (Cooper and Yamagishi, 2021), NISQA (Mittag et al., 2021a), SingMOS (Tang et al., 2024), SOMOS (Maniati et al., 2022), and TMHINT-QI (Chen and Tsao, 2021) as the primary sources for training and in-domain evaluation. To further assess the generalization of models, we incorporate the VMC'23 (Cooper et al., 2023) dataset as an OOD evaluation dataset. Detailed descriptions of each dataset are provided in Appendix A.

3.2 Preference Annotation

Given the variability in subjective scoring standards across different datasets, we adopt a preference-based annotation strategy that converts absolute scores into pairwise comparisons.

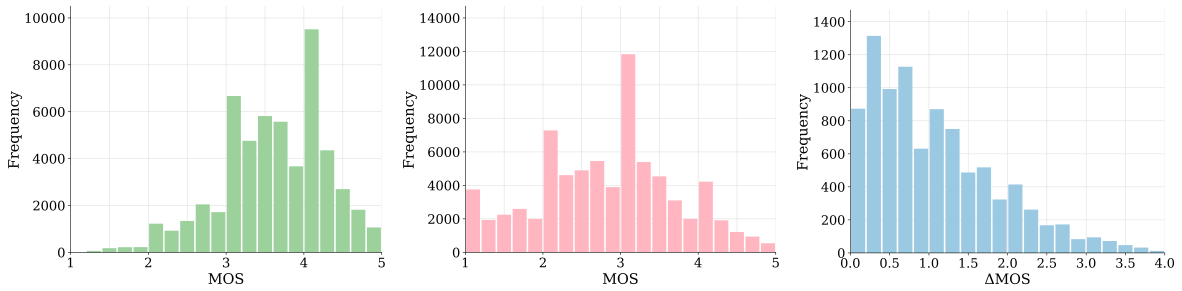
We first convert all samples into a unified 16 kHz WAV format and filter out samples with unreliable annotations, defined as those containing incomplete metadata such as speaker, content, or system identifiers, which are essential for preference construction. The remaining data from each dataset are then partitioned into three categories: (i) samples that share the same speech content but differ in speaker or system; (ii) samples that share the same speaker or are generated by the same system but differ in speech content; and (iii) samples that differ in both content and speaker or system.

To ensure meaningful and consistent comparisons, we construct preference pairs within each dataset based on the intrinsic relationships between samples. Specifically, for the first two categories, we group samples by shared content or by shared speaker/system, respectively, then construct preference pairs within each group and ensure that the samples in each pair have different MOS scores. The sample with the higher MOS is labeled as "chosen" and the other as "rejected". For the third category, since explicit grouping is not feasible, we construct preference pairs directly within the dataset and apply the same "chosen-rejected" labeling rule. Finally, we ensure that the number of preference pairs is balanced across datasets to avoid over-representation of any particular source.

This annotation strategy yields two primary benefits. First, it minimizes cross-dataset interference while preserving structured and interpretable preference comparisons to the greatest extent possible, and simultaneously incorporates more unconstrained and realistic comparison scenarios. Such a design reflects practical speech quality assessment settings, where the assessment is not restricted to identical content or the same speaker/system. Second, by converting absolute MOS values into relative preferences, the strategy mitigates dataset-specific biases and enables a unified supervision signal for both training and evaluation of speech quality assessment models across diverse scenarios. To further ensure reliability, we additionally perform human verification of the constructed preference pairs, as detailed in Appendix C.1.

Dataset	Train	Dev	Test	Scenario	Language
BVCC	9,948	2,132	1,000	natural speech, TTS, VC	English
NISQA	11,571	2,796	2,240	natural & distorted speech	English, German
SingMOS	10,000	2,720	1,000	natural speech, SVS, SVC	Chinese, Japanese
SOMOS	13,814	2,257	1,000	natural speech, TTS	English
TMHINT-QI	10,000	–	1,000	natural & noisy speech, SE	Chinese
VMC'23	–	–	3,000	natural & noisy speech, TTS, SVS, SVC, SE	French, English, Chinese
Overall	55,333	9,905	9,240	–	–

Table 1: Statistics of MOS-Pref across datasets.



(a) Distribution of chosen samples (b) Distribution of rejected samples (c) Distribution of ΔMOS

Figure 2: Distribution of MOS scores in MOS-Pref. Figures (a) and (b) show the MOS distributions of chosen and rejected samples. Figure (c) presents the distribution of ΔMOS in the test set.

3.3 Critiques Annotation

To support generative SQA modeling, we augment MOS-Pref with natural-language critiques, annotating both individual samples and preference pairs along four dimensions: noise, distortion, continuity, and naturalness. We further validate the reliability of the annotations, as detailed in Appendix C.1.

Single-sample critiques. For each audio sample, we use Gemini-2.5-Pro (Comanici et al., 2025) to generate an overall quality description covering the four dimensions above.

Pairwise comparative critiques. For each sample pair, we use Gemini-2.5-Pro to produce comparative critiques along the same dimensions and assign an overall quality score to each sample. The annotations are validated to ensure consistency with preference labels, such that the chosen sample receives a higher score than the rejected one.

3.4 Dataset Statistics

Overview. MOS-Pref is a large-scale preference dataset comprising 55,333 training samples, 9,905 development samples, and 6,240 in-domain test samples from five source datasets, along with 3,000 OOD test samples from the VMC'23 challenge, as summarized in Table 1. The dataset covers a broad spectrum of speech scenarios, including natural and synthetic speech, singing voice, VC, SVC, SE, and speech with real or simulated noise and distortions, spanning five languages: English, Chinese (Mainland and Taiwanese Mandarin), Japanese, French, and German. This diversity in acoustic conditions and linguistic coverage establishes a solid foundation for SQA models, enabling them to achieve more robust perceptual discrimination and to generalize effectively across domains and languages.

MOS Distribution. Figure 2 illustrates the distribution of MOS scores within MOS-Pref. Overall, the MOS ratings of the chosen samples are predominantly concentrated in the range of 3 and above, whereas the rejected samples are largely distributed below 3.5. This clear separation between the two categories reflects the internal consistency and reliability of the preference annotations. For the test set, the MOS difference (ΔMOS) between paired samples is mostly within 1.5 points, indicating that a large portion of the pairs exhibit very similar perceptual quality, which poses a substantial challenge for fine-grained speech quality discrimination.

4 Experimental Setup

To assess model performance in SQA, we adopt a unified preference-based setting built upon MOS-Pref. Within this setting, we train and evaluate

multiple reward modeling paradigms and compare them with existing SQA methods on the MOS-Pref test set. This section describes the evaluation task and the models under evaluation.

4.1 Evaluation Task

Task Description. The evaluation is formulated as a binary preference comparison task: given a pair of speech samples, the model must either assign quality scores to both samples or determine which one exhibits higher perceptual quality. The model's performance reflects its ability to assess speech quality. To reduce potential position bias, the presentation order of the two audio clips in each pair is randomly swapped during evaluation. A decision is counted as correct only if the model assigns a strictly higher score to the sample annotated as chosen than to the one annotated as rejected.

Evaluation Metrics. Model performance is reported in terms of accuracy (Acc). For each dataset, we calculate the proportion of evaluation pairs in which the model's scores correctly reflect the annotated preference (i.e., the chosen sample receives the higher score). The overall accuracy is computed as the proportion of correctly judged pairs across all samples from all datasets, which provides a comprehensive measure of model performance across the diverse speech scenarios covered by MOS-Pref.

4.2 Evaluated Models

We implement and evaluate a range of reward modeling paradigms based on Qwen2-Audio-7B (Chu et al., 2024b), covering scalar, semi-scalar, and GRMs. In addition, we include representative MOS prediction models such as UTMOS (Saeki et al., 2022), UTMOSv2 (Baba et al., 2024), and NISQA v2.0 (Mittag et al., 2021b), as well as the LLM-as-a-judge (Zheng et al., 2023) paradigm. The following describes the models under each paradigm, with detailed configurations provided in Appendix B.

Classic scalar reward models. These models output a single quality score for each audio sample. We adopt two training objectives: the Bradley-Terry(BT) (Bradley and Terry, 1952) loss, which maximizes the likelihood that the sample annotated as chosen receives a higher predicted score, and the mean squared error (MSE) loss, which minimizes the squared difference between the predicted score and the corresponding MOS value.

Semi-scalar reward models. These models extend the scalar paradigm by incorporating textual descriptions of audio quality. The model is first trained on the single-sample critiques of MOS-Pref to generate these descriptions, after which its outputs are passed through a reward head to produce scalar scores. Similar to classic scalar models, both BT and MSE objectives are used for training.

Generative reward models. GRMs take two audio samples as input simultaneously. As in the semi-scalar setting, the model first learns to produce descriptive quality assessments. It is refined through supervised fine-tuning (SFT) on the pairwise comparative critiques of MOS-Pref, learning to generate textual judgments that express quality preferences between paired samples. The model is further optimized using reinforcement learning methods, including Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO) (Yu et al., 2025). In both cases, training employs a rule-based reward function, denoted as the R_{ACC}, which is determined solely by whether the model correctly judges the quality of a speech sample pair. Using S_c and S_r to denote the scores assigned to the chosen and rejected samples, respectively, the R_{ACC} is defined as follows:

$$R_{ACC} = \begin{cases} 1, & if\ S_c > S_r, \\ -1, & otherwise \end{cases} \quad (1)$$

LLM-as-a-judge. These models directly compare two audio samples and output a preference judgment without producing explicit scalar scores. We evaluate this paradigm using Gemini-2.5-Pro, Qwen2-Audio-7B-Instruct (Chu et al., 2024a), Qwen2.5-Omni-7B (Xu et al., 2025a), and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025c).

5 Experimental Results

This section presents an empirical evaluation of different paradigms based on MOS-Pref. We first report the overall evaluation results, then conduct an error analysis across samples with varying MOS gaps, and finally explore a MOS-aware GRM design to address the limitations identified.

5.1 Main Results

Table 2 presents the main evaluation results, summarizing the performance of all models across individual datasets as well as overall.

Comparison of different modeling paradigms. As shown in the evaluation results, the classic scalar models achieve the highest overall accuracy (80.04% with BT loss), followed by the semi-scalar models (78.82% with BT loss), while the GRMs

Model	BVCC	NISQA	SingMOS	SOMOS	TMHINT-QI	VMC'23	Overall
MOS prediction models							
UTMOS	86.70	73.17	63.80	65.00	68.10	60.83	68.18
UTMOSv2	**89.80**	73.88	59.90	65.80	63.20	61.00	67.88
NISQA v2.0	73.50	**85.13**	60.70	51.50	60.80	61.60	67.32
LLM-as-a-judge							
Gemini-2.5-Pro	63.90	81.96	58.50	64.10	73.20	65.57	69.26
Qwen2-Audio-7B-Instruct	51.20	46.52	50.10	29.70	48.90	43.53	44.88
Qwen2.5-Omni-7B	54.00	63.97	57.70	56.30	69.90	58.43	60.23
Qwen3-Omni-30B-A3B-Instruct	63.00	77.19	60.80	64.80	77.10	66.73	69.13
Scalar reward models							
Classic$_{BT\ Loss}$	85.70	83.93	74.80	**76.80**	**81.10**	**77.73**	**80.04**
Classic$_{MSE\ Loss}$	82.80	79.33	69.80	74.30	79.40	72.23	75.83
Semi-scalar reward models							
CLoud$_{BT\ Loss}$	85.50	81.07	**78.20**	75.30	80.60	75.70	78.82
CLoud$_{MSE\ Loss}$	84.50	77.81	76.10	75.20	80.50	73.67	77.01
Generative reward models							
GRM$_{SFT}$	80.60	79.60	75.10	70.00	77.90	69.23	74.63
GRM$_{GRPO}$	82.50	80.31	76.40	74.60	78.10	73.13	76.94
GRM$_{DAPO}$	82.60	82.05	77.50	74.40	79.30	73.13	77.60

Table 2: Evaluation results of models from different paradigms on the MOS-Pref test set. The best results are shown in **bold** and the second best is with underline. Classic$_{BT/MSE\ Loss}$ denote scalar reward models trained with BT or MSE loss; CLoud$_{BT/MSE\ Loss}$ denote semi-scalar reward models with BT or MSE loss; GRM$_{SFT/GRPO/DAPO}$ denote generative reward models trained without RL, with GRPO, or with DAPO.

attain slightly lower overall performance. These findings indicate that, under the current evaluation setup, the classic scalar paradigm remains highly effective in distinguishing speech quality. Notably, while UTMOS and UTMOSv2 perform strongly on BVCC, and NISQA v2.0 performs well on NISQA, their accuracy exhibits a marked drop on the other datasets, suggesting that MOS prediction models may generalize poorly across diverse datasets. Furthermore, all models within the LLM-as-a-judge paradigm achieve accuracy below 70% on most datasets and in overall performance, indicating that applying current audio large language models to SQA remains challenging. We also conducted a study on preference reversal across different reward modeling paradigms, with detailed results provided in Appendix C.2 for further analysis.

Comparison of training objectives for scalar-based reward models. Within scalar-based paradigms, we examine the effect of different training objectives: BT loss and MSE loss. For the classic scalar models, BT loss consistently outperforms MSE loss, achieving over a 4% gain in overall accuracy; a similar trend is observed for the semi-scalar models. The advantage of BT loss over MSE loss is even more pronounced on the OOD VMC'23 dataset. These findings indicate that optimizing relative sample ordering with BT loss is generally more effective than direct regression to MOS scores. The advantage likely stems from BT loss explicitly encouraging the model to capture relative quality differences, whereas MSE is more sensitive to variations in absolute MOS scores across datasets, affecting cross-domain robustness.

Comparison of training strategies for GRMs. For GRMs, we evaluate two reinforcement learning strategies: DAPO and GRPO. Both methods yield comparable overall performance, with 77.60% for DAPO and 76.94% for GRPO. Although the differences are modest, DAPO appears to have a slight advantage in learning from paired audio comparisons overall. More importantly, both strategies yield a marked and consistent performance gain over SFT alone, further highlighting the benefit and necessity of reinforcement learning for GRMs. Furthermore, all GRM variants demonstrate competitive performance across both in-domain and OOD datasets, particularly on NISQA and SingMOS, despite their overall accuracy remaining slightly lower than that of the classic scalar models.

5.2 Error Analysis

What constrains model performance? We observe that all models struggle with sample pairs

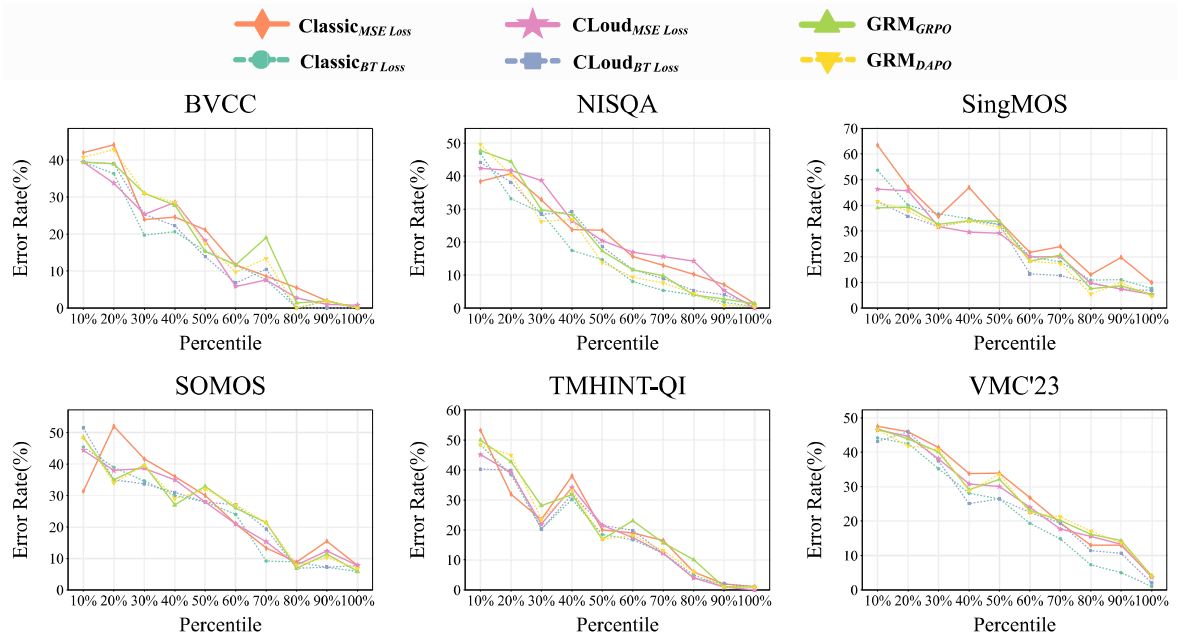


Figure 3: Percentile-based error analysis across datasets: error rates are highest for pairs with small MOS differences and decline markedly as the MOS gap widens.

having small MOS differences. To quantify this phenomenon, we conduct a percentile-based error analysis: within each dataset, pairs are first sorted by MOS difference in ascending order and then divided into percentile bins of equal size. For each bin, we compute the proportion of incorrectly ranked pairs for each reward model. Figure 3 presents the results. Across all models, error rates are consistently high in the lowest MOS difference percentiles, even the top-performing classic scalar model exhibits error rates of 40% or higher. As the MOS difference increases, however, the error rates for all models decrease markedly. These findings indicate that fine-grained speech quality discrimination remains a critical bottleneck for current reward models, highlighting the need for methods capable of capturing subtle quality differences.

5.3 MOS-aware GRM

Design of the MOS-aware reward function. GRMs offer a flexible and interpretable framework for reward modeling, allowing the reward signal to be decomposed and extended. During the original GRM reinforcement learning process, the R_{ACC} considers only whether the model correctly ranks the quality of an audio pair, treating all pairs as equally informative. This uniform treatment overlooks the inherent difficulty of pairs with small MOS differences, which may constrain the model's ability to learn fine-grained quality distinctions. To address this limitation, we introduce the $R_{\Delta MOS}$, which incorporates the MOS gap of each audio pair into the reward function:

$$R_{\Delta MOS} = \begin{cases} (cos(\Delta MOS \cdot \pi) + 1)/2, & if \ \ S_c > S_r, \\ (cos(\Delta MOS \cdot \pi) - 1)/2, & otherwise \end{cases} \quad (2)$$

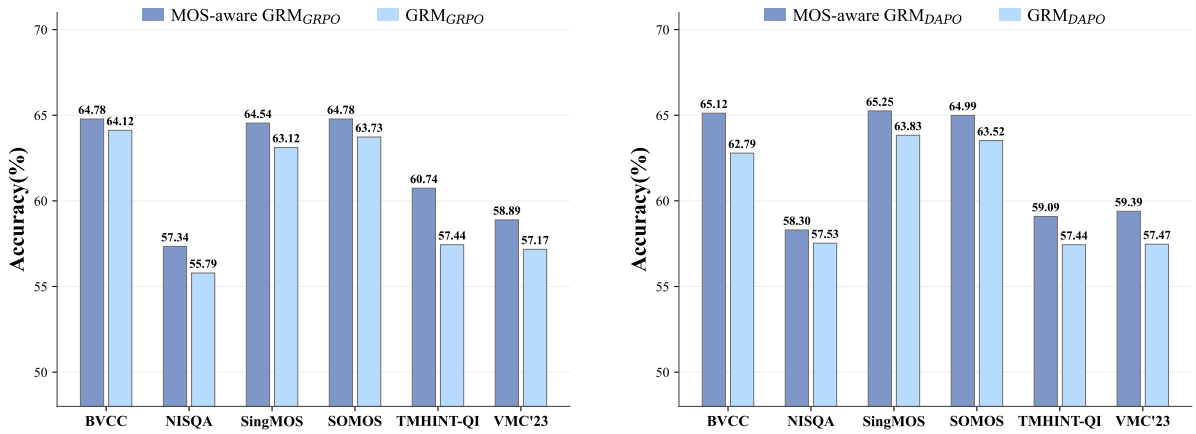
Specifically, the MOS gap refers to the difference between the original MOS scores of the paired samples in the dataset. ΔMOS is obtained by normalizing this gap by the 90th percentile of MOS differences in the dataset, and then clamping the resulting value to the [0,1] range to mitigate the influence of extreme outliers. Based on the R_{ACC} and the $R_{\Delta MOS}$, the final MOS-aware reward function is defined as follows:

$$R_{MOS-aware} = R_{ACC} + R_{\Delta MOS} \quad (3)$$

This formulation offers two key advantages. First, it enables adaptive scaling of the reward based on the relative difficulty of each sample pair: for pairs with small MOS differences, correct predictions are assigned relatively larger rewards while incorrect predictions incur relatively smaller penalties; for pairs with large MOS differences, the rewards and penalties are modulated accordingly. Second, the cosine-based shaping ensures smooth transitions at both ends of the scale. By adjusting the reward according to the implied difficulty of each pair, the MOS-aware design provides a more informative learning signal, encouraging the model to attend to subtle perceptual distinctions while still penalizing clearly incorrect predictions.

Models	BVCC	NISQA	SingMOS	SOMOS	TMHINT-QI	VMC'23	Overall
MOS-aware GRM$_{GRPO}$	83.10 (+0.60)	81.47 (+1.16)	76.50 (+0.10)	75.80 (+1.20)	80.40 (+2.30)	74.13 (+1.00)	78.00 (+1.06)
MOS-aware GRM$_{DAPO}$	83.40 (+0.80)	82.14 (+0.09)	78.40 (+0.90)	75.10 (+0.70)	79.90 (+0.60)	73.57 (+0.44)	78.08 (+0.48)

Table 3: Evaluation results of MOS-aware GRMs with GRPO and DAPO on MOS-Pref test set. Numbers in parentheses denote absolute improvements over baseline GRMs.



(a) Training strategy: GRPO (b) Training strategy: DAPO

Figure 4: Performance comparison of MOS-aware GRMs and baseline GRMs trained with different reinforcement learning methods on samples with MOS difference ≤ 0.5.

Impact of MOS-aware reward on GRM performance. To evaluate the effectiveness of the proposed $R_{MOS-aware}$, we incorporate it into the GRM training process under both GRPO and DAPO strategies, while keeping all other training configurations unchanged. This design enables a direct comparison with baseline GRMs that rely solely on the R_{ACC}. Table 3 summarizes the overall results, showing that MOS-aware GRMs achieve consistent improvements over their baseline counterparts across all evaluation scenarios.

To further examine performance on perceptually challenging pairs, we evaluate models on sample pairs with a MOS difference threshold of 0.5. Figure 4 presents the corresponding results. Across all datasets, MOS-aware GRMs consistently outperform the baselines on these fine-grained comparisons. For illustration, with the GRPO strategy, accuracy increases from 57.44% to 60.74% on THMINT-QI and from 57.17% to 58.89% on VMC'23; under the DAPO strategy, the accuracy improves from 62.79% to 65.12% on BVCC and from 57.47% to 59.39% on VMC'23.

These results indicate that incorporating MOS gap information into the reward function produces stable and significant gains. The MOS-aware design provides a difficulty-sensitive learning signal, enabling GRMs to better capture fine-grained perceptual quality differences beyond the conventional R_{ACC}. A detailed evaluation of the $R_{MOS-aware}$'s effectiveness in more challenging scenarios is provided in Appendix C.3.

6 Conclusion

In this work, we reformulate SQA from score regression to a unified preference-based paradigm. We construct MOS-Pref, a large-scale MOS-derived preference dataset that supports training and evaluation of SQA models and enables consistent analysis across diverse speech domains.

Based on MOS-Pref, we implement a range of reward modeling paradigms to develop stronger SQA models, and systematically evaluate them alongside existing SQA methods. Our results reveal clear differences across paradigms and identify challenges in fine-grained speech quality discrimination.

Motivated by these findings, we propose MOS-aware GRM that incorporates MOS gap information into the reward function. By adapting the reward signal to the relative difficulty of preference pairs, the proposed approach improves fine-grained quality discrimination and consistently enhances performance on challenging cases.

We hope that this work can serve as a useful foundation for advancing research on speech quality reward modeling and preference alignment in speech generation systems.

Limitations

Despite the contributions of this work, several limitations remain and warrant further investigation.

First, MOS-Pref relies on subjective human judgments of speech quality, which are inherently affected by listener variability and individual perception differences. Although preference reformulation helps mitigate some inconsistencies, such subjectivity cannot be fully eliminated.

Second, the current study focuses on perceptual quality and does not model paralinguistic attributes such as prosody, emotion, or speaking style. Incorporating these factors could enable a more comprehensive assessment of speech quality, especially for expressive speech generation scenarios.

Third, the covered languages and speech scenarios remain limited. Expanding the dataset to additional languages and real-world acoustic conditions would be important for evaluating generalization and robustness in broader application settings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D Chang, and Prithviraj Ammanabrolu. 2024. Critique-out-loud reward models. arXiv preprint arXiv:2408.11791.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2024. 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, pages 818–824. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33:12449–12460.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 39(3/4):324–345.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Eng Siong Chng. 2025. Audio large language models can be descriptive speech quality evaluators. arXiv preprint arXiv:2501.17202.

Yu-Wen Chen and Yu Tsao. 2021. Inqss: a speech intelligibility and quality assessment model using a multi-task learning network. arXiv preprint arXiv:2111.02585.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024a. Qwen2-audio technical report. CoRR, abs/2407.10759.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024b. Qwen2-audio technical report. arXiv preprint arXiv:2407.10759.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261.

Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2023. The voicemos challenge 2023: Zero-shot subjective speech quality prediction for multiple domains. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 1–7. IEEE.

Erica Cooper and Junichi Yamagishi. 2021. How do voices from past speech synthesis challenges compare today? arXiv preprint arXiv:2105.02373.

Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, and Huaming Wang. 2024. Pam: Prompting audio-language models for audio quality assessment. arXiv preprint arXiv:2402.00282.

Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang. 2018. Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm. arXiv preprint arXiv:1808.05344.

Tomoki Hayashi, Ryuichi Yamamoto, Katsuki Inoue, Takenori Yoshimura, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Yu Zhang, and Xu Tan. 2020. Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 7654–7658. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing, 29:3451–3460.

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2022. The voicemos challenge 2022. arXiv preprint arXiv:2203.11389.

Wen-Chin Huang, Lester Phillip Violeta, Songxiang Liu, Jiatong Shi, and Tomoki Toda. 2023. The singing voice conversion challenge 2023. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 1–8. IEEE.

Keith Ito and Linda Johnson. 2017. The lj speech dataset.

Yuto Kondo, Hirokazu Kameoka, Kou Tanaka, and Takuhiro Kaneko. 2025. Rethinking mean opinion scores in speech quality assessment: Score aggregation through quantized distribution fitting. In 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025, pages 1–5. IEEE.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling. arXiv preprint arXiv:2504.02495.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. Mosnet: Deep learning based objective assessment for voice conversion. arXiv preprint arXiv:1904.08352.

Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis. 2022. Somos: The samsung open mos dataset for the evaluation of neural text-to-speech synthesis. arXiv preprint arXiv:2204.03040.

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

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021b. NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021, pages 2127–2131. ISCA.

Gabriel Mittag, Saman Zadtootaghaj, Thilo Michael, Babak Naderi, and Sebastian Möller. 2021c. Bias-aware loss for training image and speech quality prediction models from multiple datasets. In 2021 13th International Conference on Quality of Multimedia Experience (QoMEX), pages 97–102. IEEE.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744.

Olivier Perrotin, Brooke Stephenson, Silvain Gerber, and Gérard Bailly. 2023. The blizzard challenge 2023. In 18th Blizzard Challenge Workshop, pages 1–27. ISCA.

Jaden Pieper and Stephen D Voran. 2024. Alignnet: Learning dataset score alignment functions to enable better training of speech quality estimators. arXiv preprint arXiv:2406.10205.

Zili Qi, Xinhui Hu, Wangjin Zhou, Sheng Li, Hao Wu, Jian Lu, and Xinkang Xu. 2023. Le-ssl-mos: Self-supervised learning mos prediction with listener enhancement. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 1–6. IEEE.

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

International Telecommunication Union. Telecommunication Standardization Sector. 1996. Methods for subjective determination of transmission quality. International Telecommunication Union.

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

Yuxun Tang, Jiatong Shi, Yuning Wu, and Qin Jin. 2024. Singmos: An extensive open-source singing voice dataset for mos prediction. arXiv preprint arXiv:2406.10911.

Cassia Valentini-Botinhao and Junichi Yamagishi. 2018. Speech enhancement of noisy and reverberant speech for text-to-speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(8):1420–1433.

Jean-Marc Valin and Jan Skoglund. 2019. Lpcnet: Improving neural speech synthesis through linear prediction. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5891–5895. IEEE.

Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, and 1 others. 2025a. Enabling auditory large language models for automatic speech quality evaluation. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025b. Spark-tts: An efficient llm-based text-to-speech model

with single-stream decoupled speech tokens. arXiv preprint arXiv:2503.01710.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 13618–13628.

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. 2025a. Qwen2.5-omni technical report. CoRR, abs/2503.20215.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025b. Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025c. Qwen3-omni technical report. CoRR, abs/2509.17765.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476.

Ryandhimas E Zezario, Yu-Wen Chen, Szu-Wei Fu, Yu Tsao, Hsin-Min Wang, and Chiou-Shann Fuh. 2024. A study on incorporating whisper for robust speech assessment. In 2024 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE.

Ryandhimas E Zezario, Sabato M Siniscalchi, Hsin-Min Wang, and Yu Tsao. 2025. A study on zero-shot non-intrusive speech assessment using large language models. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36:46595–46623.

A Dataset Source

BVCC: The BVCC dataset originates from large-scale listening tests and comprises 7,106 samples, including natural speech and synthetic speech generated by 187 systems spanning diverse TTS and voice conversion (VC) methods. The data sources include Blizzard Challenge, Voice Conversion Challenge, and publicly available samples from ESPnet-TTS (Hayashi et al., 2020).

NISQA: NISQA is designed for speech quality assessment in communication networks, covering both simulated distortions and real call recordings. The training and validation sets comprise 11,020 and 2,700 samples, respectively, and consist of simulated and live subsets. The test set contains four subsets, with a total of 952 samples.

SingMOS: SingMOS is an open-source, high-quality singing voice MOS dataset in Chinese and Japanese, containing 3,421 segments from both natural singing and 33 systems. It covers various synthesis techniques including singing voice synthesis (SVS), singing voice conversion (SVC), and vocoder-based re-synthesis.

SOMOS: SOMOS is a large-scale MOS dataset focusing on neural TTS. It contains 20,100 samples synthesized from LJ Speech (Ito and Johnson, 2017) by 200 neural TTS systems and natural speech, all generated using the same LPCNet vocoder (Valin and Skoglund, 2019) to isolate acoustic-model differences. MOS-Pref adopts the SOMOS-clean subset to ensure annotation consistency and reliability.

TMHINT-QI: TMHINT-QI is a MOS dataset focused on Mandarin speech, mainly for evaluating speech enhancement (SE) systems. It contains 24,408 samples generated by adding four types of noise (babble, street, pink, white) at four SNR levels (-2, 0, 2, 5 dB) to clean speech, then processed by five SE systems.

VMC'23: The VMC'23 dataset originates from The VoiceMOS Challenge 2023 (Cooper et al., 2023). It includes three tracks: (1) French TTS, based on Blizzard Challenge 2023 (Perrotin et al., 2023) listening tests, with 1,460 samples; (2) singing voice conversion, based on the Voice Conversion Challenge 2023 (Huang et al., 2023), containing 4,040 samples; (3) noisy and enhanced speech, based on TMHINT-QI(S) (Zezario et al., 2024), consisting of 1,960 samples.

B Experimental Details

B.1 Model Configurations and Hyperparameters

All models are trained in a fully parameterized manner.

Classic scalar reward models: Both BT-loss and MSE-loss variants are trained for one epoch

with a batch size of 32, using the AdamW optimizer (initial learning rate 5e-6, weight decay 5e-6).

Semi-scalar reward models: Both BT-loss and MSE-loss variants are trained for one epoch with a batch size of 32, using the AdamW optimizer (initial learning rate 1e-6, weight decay 1e-6). The total loss is a weighted sum of the reward loss and the LM loss, where the LM loss is assigned a weight of 1.25.

Generative reward models: Both GRPO and DAPO are trained for one epoch with a batch size of 64 using the AdamW optimizer (learning rate 1e-6), generating 4 completions per prompt with temperature 1.0, and employing ZeRO-2 optimization.

For inference, the temperature was set to 0.0 for all models. Training and inference are conducted on 8× H20 GPUs for all models.

B.2 Prompt Templates

This section details the prompt templates used both for data annotation and for model inference during evaluation.

Prompts for data annotation. To construct the training sets for the reward models, we used Gemini-2.5-Pro as the annotator and designed two kinds of prompts. For single-sample critic annotation, as illustrated in Figure 5, Gemini-2.5-Pro is guided to listen to a single speech sample and provide a detailed critique of its perceptual quality across multiple dimensions. For paired-sample critic annotation, as shown in Figure 6, Gemini-2.5-Pro compares two speech samples, offers a detailed assessment of each along key perceptual dimensions, and then assigns overall quality scores from 1 to 10 to both samples. This paired-comparison protocol yields both textual rationales and numerical ratings that serve as the foundation for preference-based generative reward modeling.

Prompts for model inference. During evaluation, different reward-modeling paradigms adopt distinct prompt formats. Figure 7 presents the prompt structure for both the Classic scalar reward models and the Semi-scalar reward models: the Classic scalar models assess a single speech sample and output a scalar score representing its overall perceptual quality, whereas the Semi-scalar models produce a detailed critique of the sample before deriving a scalar quality score. Figure 8 shows the prompt for the GRMs, which requires the model to compare two speech samples across four perceptual dimensions and then assign an overall quality score from 1 to 10 to each sample. Finally, Figure 9 illustrates the prompt for the LLM-as-a-judge paradigm, where the model is asked to decide which of two input speech samples has higher overall perceptual quality and to output the identifier of the higher-quality sample.

C Additional Experimental Results

C.1 Human Verification

Validation of the preference annotation. To assess the reliability of the MOS-derived labels, we conducted a human validation study. Four human evaluators annotated "chosen/rejected" labels for 100 randomly sampled preference pairs. The annotations yielded an 82% agreement with the MOS-derived preference labels. This level of agreement indicates that the MOS-Pref closely reflects human perceptual preferences.

Validation of LLM annotation reliability. We evaluated the consistency between natural-language critiques generated by Gemini-2.5-Pro and human perceptual judgments. Four human evaluators assessed whether the model's critiques aligned with their own judgments on 100 randomly sampled items. We observed that 70% of the critiques were consistent with human perception. Unlike binary preference judgments, natural-language critiques require describing fine-grained perceptual attributes (e.g., noise, distortion, continuity, naturalness), where subjective variability is inevitable. Under these conditions, a 70% alignment rate is considered reasonable and expected.

C.2 Study of Preference Reversal

Preference reversal primarily refers to the presence of preference cycles, where a model's pairwise judgments violate transitivity (e.g., A>B, B>C, but C>A). Scalar and semi-scalar reward models assign deterministic scalar scores independently for each sample (with temperature = 0), which inherently prevents the formation of cycles. In contrast, GRMs score each pair directly, making them susceptible to exhibiting such cycles.

To directly assess the occurrence of preference cycles in GRMs, we conducted an explicit triplet-based cycle detection study. For each dataset, we followed the original grouping rules and sampled triplets (A,B,C) with distinct MOS scores such that MOS(A)>MOS(B)>MOS(C), yielding three preference pairs: (A,B), (B,C), and (A,C). Across all datasets, we constructed a total of 565 triplets

(1,695 preference pairs), including 65 from SOMOS and 100 from each of the remaining datasets. We then evaluated GRMs trained with SFT, GRPO, and DAPO to quantify both the pairwise accuracy and the preference-cycle rate, as reported in Table 4 and Table 5.

The results indicate that the SFT-trained GRM exhibits a small yet non-negligible preference-cycle rate (2.83% overall), accompanied by lower pairwise accuracy. In contrast, the RL-trained GRMs (GRPO and DAPO) achieve substantially higher accuracy and reduce preference cycles to almost zero (0.18% overall for both methods), with only a single cycle observed across all datasets. These findings demonstrate that RL-based training not only enhances alignment with human preferences but also effectively suppresses preference reversals.

C.3 Effectiveness of the MOS-aware Reward in Challenging Scenarios

To further evaluate the performance of MOS-aware GRMs on more challenging samples, we specifically tested pairs with extremely small MOS gaps ($\Delta MOS < 0.2$). Both GRPO-trained and DAPO-trained GRMs were evaluated, and the MOS-aware GRMs were compared against the corresponding vanilla GRMs. The evaluation results are presented in Table 6 and Table 7. Across both training algorithms, the MOS-aware GRMs consistently achieve improved or comparable accuracy on pairs with extremely small MOS differences.

GRPO setting: Compared with the vanilla GRM, the MOS-aware GRM increases overall accuracy from 54.41% to 55.33%, with gains observed on BVCC, NISQA, SingMOS, TMHINT-QI, and VMC'23. A slight decrease is noted on SOMOS, but the gap is minor.

DAPO setting: The MOS-aware GRM improves overall accuracy from 54.75% to 55.78%, showing equal or better performance across all datasets.

These results demonstrate that incorporating MOS gap information into the reward function yields measurable benefits, particularly in the fine-grained regime where perceptual differences are extremely small and the task is most challenging.

D Usage of Large Language Models

In this study, large language models were used only for polishing parts of the manuscript's text to improve fluency and readability. They did not participate in the research design, the development or execution of the methodology, the collection or analysis of data, or the creation and validation of the core scientific content. All core research content and findings are entirely the work and responsibility of the authors.

Model	BVCC	NISQA	SingMOS	SOMOS	TMHINT-QI	VMC'23	Overall
GRM$_{SFT}$	82.67	85.00	73.67	64.62	80.00	66.33	76.05
GRM$_{GRPO}$	84.67	85.33	74.67	72.31	77.33	71.33	77.94
GRM$_{DAPO}$	84.00	87.00	75.67	70.77	81.00	72.33	78.94

Table 4: Pairwise accuracy (%) of GRMs trained with SFT, GRPO, and DAPO on the preference-cycle evaluation set constructed from sampled triplets.

Model	BVCC	NISQA	SingMOS	SOMOS	TMHINT-QI	VMC'23	Overall
GRM$_{SFT}$	2.00	2.00	2.00	1.54	0.00	9.00	2.83
GRM$_{GRPO}$	0.00	0.00	0.00	0.00	0.00	1.00	0.18
GRM$_{DAPO}$	0.00	0.00	0.00	0.00	1.00	0.00	0.18

Table 5: Preference-cycle rate (%) of GRMs trained with SFT, GRPO, and DAPO on the preference-cycle evaluation set constructed from sampled triplets.

Model	BVCC	NISQA	SingMOS	SOMOS	TMHINT-QI	VMC'23	Overall
GRM$_{GRPO}$	56.79	52.85	60.98	**58.19**	43.24	53.20	54.41
MOS-aware GRM$_{GRPO}$	**58.02**	**54.40**	**63.41**	57.63	**48.65**	**53.78**	**55.33**

Table 6: Evaluation results of the MOS-aware GRM with GRPO on the subset with $\Delta MOS < 0.2$. The better results in each column are shown in **bold**.

Model	BVCC	NISQA	SingMOS	SOMOS	TMHINT-QI	VMC'23	Overall
GRM$_{DAPO}$	58.02	54.40	63.41	57.06	48.65	52.62	54.75
MOS-aware GRM$_{DAPO}$	58.02	**55.44**	63.41	**59.32**	48.65	**53.49**	**55.78**

Table 7: Evaluation results of the MOS-aware GRM with DAPO on the subset with $\Delta MOS < 0.2$. The better results in each column are shown in **bold**.

Prompt for Gemini-2.5-Pro to annotate a single-audio critic

System:
You are a helpful assistant.

Prompt:
[Instruction starts]

You will be given an audio clip.
Please act as a professional speech quality evaluation expert and provide objective description for the audio clip based on the following four dimensions:
1. Noise: Whether there is background noise, and whether it interferes with understanding.
2. Distortion: Whether there are compression artifacts, electrical noise, or other distortions.
3. Naturalness: Whether the speech sounds natural and resembles real human speech.
4. Continuity: Whether the speech is fluent and continuous, or if there are any dropouts or interruptions.

[Instruction ends]

[Audio starts]
<audio>
[Audio ends]

Figure 5: Prompt structure for Gemini-2.5-Pro to annotate a single-audio critic.

Prompt for Gemini-2.5-Pro to annotate a paired-audio critic with scores

System:
You are a helpful assistant.

Prompt:
[Instruction starts]

You will be given two audio clips.
Please act as a professional speech quality evaluation expert and provide the scoring rationale based on the following four dimensions:
1. Noise: Whether there is background noise, and whether it interferes with understanding.
2. Distortion: Whether there are compression artifacts, electrical noise, or other distortions.
3. Naturalness: Whether the speech sounds natural and resembles real human speech.
4. Continuity: Whether the speech is fluent and continuous, or if there are any dropouts or interruptions.

After providing the above scoring rationale, give each audio an overall quality score from 1 to 10 (integer only, 10 = best quality).
The score should be consistent with the justification above.

Finally, output the scores in the exact format:
Final Scores: [[audio1_score, audio2_score]]

[Instruction ends]

[Audio1 starts]
<audio>
[Audio1 ends]

[Audio2 starts]
<audio>
[Audio2 ends]

Figure 6: Prompt structure for Gemini-2.5-Pro to annotate a paired-audio critic with scores.

Prompt for scalar and semi-scalar reward models inference

System:
You are a helpful assistant.

Prompt:
[Instruction starts]

You will be given an audio clip.
Please act as a professional speech quality evaluation expert and provide an objective description of the input audio based on the following four dimensions:
1. Noise: Whether there is background noise, and whether it interferes with understanding.
2. Distortion: Whether there are compression artifacts, electrical noise, or other distortions.
3. Naturalness: Whether the speech sounds natural and resembles real human speech.
4. Continuity: Whether the speech is fluent and continuous, or if there are any dropouts or interruptions.

[Instruction ends]

[Audio starts]
<audio>
[Audio ends]

Figure 7: Prompt structure for scalar and semi-scalar reward models inference.

Prompt for generative reward models inference

System:
You are a helpful assistant.

Prompt:
[Instruction starts]

You will be given two audio clips.
Please act as a professional speech quality evaluation expert and provide the scoring rationale based on the following four dimensions:
1. Noise: Whether there is background noise, and whether it interferes with understanding.
2. Distortion: Whether there are compression artifacts, electrical noise, or other distortions.
3. Naturalness: Whether the speech sounds natural and resembles real human speech.
4. Continuity: Whether the speech is fluent and continuous, or if there are any dropouts or interruptions.

After providing the above scoring rationale, give each audio an overall quality score from 1 to 10 (integer only, 10 = best quality).The score should be consistent with the justification above.

Finally, output the scores in the exact format:
Final Scores: [[audio1_score, audio2_score]]

[Instruction ends]

[Audio1 starts]
<audio>
[Audio1 ends]

[Audio2 starts]
<audio>
[Audio2 ends]

Figure 8: Prompt structure for generative reward models inference.

Prompt for LLM-as-a-judge

System:
You are a helpful assistant.

Prompt:
[Instruction starts]

You will be given two audio clips.
Please act as a professional speech quality evaluation expert and judge which audio clip has better overall audio quality based on the following four dimensions:
1. Noise: Whether there is background noise, and whether it interferes with understanding.
2. Distortion: Whether there are compression artifacts, electrical noise, or other distortions.
3. Naturalness: Whether the speech sounds natural and resembles real human speech.
4. Continuity: Whether the speech is fluent and continuous, or if there are any dropouts or interruptions.

Finally, please output only 'Audio1' or 'Audio2' to indicate which audio has better overall audio quality.

[Instruction ends]

[Audio1 starts]
<audio>
[Audio1 ends]

[Audio2 starts]
<audio>
[Audio2 ends]

Figure 9: Prompt structure for LLM-as-a-judge.