QAMRO: Quality-aware Adaptive Margin Ranking Optimization for Human-aligned Assessment of Audio Generation Systems

Chien-Chun Wang*, Kuan-Tang Huang*, Cheng-Yeh Yang*, Hung-Shin Lee‡, Hsin-Min Wang†, and Berlin Chen*
*Dept. Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
†Institute of Computer Science, Academia Sinica, Taiwan
‡United Link Co., Ltd., Taiwan

***Abstract*—Evaluating audio generation systems, including text-to-music (TTM), text-to-speech (TTS), and text-to-audio (TTA), remains challenging due to the subjective and multi-dimensional nature of human perception. Existing methods treat mean opinion score (MOS) prediction as a regression problem, but standard regression losses overlook the relativity of perceptual judgments. To address this limitation, we introduce QAMRO, a novel Quality-aware Adaptive Margin Ranking Optimization framework that seamlessly integrates regression objectives from different perspectives, aiming to highlight perceptual differences and prioritize accurate ratings. Our framework leverages pre-trained audio-text models such as CLAP and Audiobox-Aesthetics, and is trained exclusively on the official AudioMOS Challenge 2025 dataset. It demonstrates superior alignment with human evaluations across all dimensions, significantly outperforming robust baseline models.**

***Index Terms*—Audio quality assessment, mean opinion score, ranking loss, quality-aware weighting, adaptive margins.**

I. Introduction

Audio generation systems, including text-to-music (TTM) [1], [2], text-to-speech (TTS) [3]–[5], and text-to-audio (TTA) [6], [7], have made rapid progress driven by breakthroughs such as diffusion models and large language models [8]–[15]. These systems can generate expressive audio from text, supporting applications in areas such as music production, content creation, and interactive media. A key challenge is to evaluate the perceptual quality of the generated audio. While subjective mean opinion scores (MOS) can provide reliable assessments, they are expensive, non-scalable, and difficult to reproduce. In contrast, objective metrics, such as Fréchet Audio Distance (FAD) [16] and Inception Score (IS) [17], are often inconsistent with human preferences, especially in terms of semantic alignment and perceptual quality.

As a result, MOS prediction has been formulated as a regression task, where the model estimates human perceptual quality scores based on audio and optionally text [18]. Most previous research efforts [19]–[23] employ training losses such as Mean Absolute Error (MAE) or Mean Squared Error (MSE) to minimize the prediction errors of the models. However, these methods do not account for the relative rankings among samples, a critical component of perceptual assessment. Absolute MOS values may vary across datasets and annotators, while relative preferences are often more consistent and meaningful in practical use cases such as system comparison and model selection. Although relatively underexplored in MOS prediction, the development of effective ranking loss functions has recently gained traction in several application domains [24]–[41] due to their ability to better capture relative rankings in subjective regression tasks. However, standard ranking loss functions typically adopt a fixed margin and treat all sample pairs uniformly, ignoring perceptual differences and the different importance of high-quality content.

In light of these limitations, we propose a novel modeling framework to introduce a ranking-based perspective to the MOS prediction task. Specifically, we present a **Q**uality-aware **A**daptive **M**argin **R**anking **O**ptimization (QAMRO) strategy that enhances the training of MOS prediction models by encouraging correct pairwise rankings. Unlike conventional ranking losses [42]–[44], QAMRO adjusts the margin according to perceptual score gaps and gives more weight to high-quality samples. We conducted experiments strictly on the official AudioMOS Challenge 2025 datasets (i.e., in a closed-set setting), and the results show that incorporating ranking-aware supervision can improve alignment with human judgment while maintaining accurate absolute score predictions. This demonstrates the effectiveness of our strategy in perceptual regression, thereby suggesting a promising direction for future audio assessment research.

II. Proposed Method

A. Quality-aware Adaptive Margin Ranking Optimization

To better align model predictions with human perceptual preferences, we design a quality-aware adaptive margin ranking loss, which puts focus on preserving the relative ranking of predicted scores, complementarily to standard regression losses that normally minimize absolute prediction errors.

Given a set of sample pairs $\mathcal{P} = \{(i, j) \mid y_i \neq y_j\}$ in a mini-batch, where y_i and y_j denote the ground-truth MOS scores of samples i and j, respectively, we define two ranking-based loss functions. The first is the *margin ranking loss*, which is a standard ranking loss that enforces a fixed margin m between the predicted scores of samples with different ratings, with all sample pairs being equally weighted:

$$\mathcal{L}_{\text{MR}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \max\Big(0,\ -\text{sign}(y_i - y_j) \cdot (\hat{y}_i - \hat{y}_j) + m\Big), \tag{1}$$

where $\hat{y}_i$ and $\hat{y}_j$ mark, respectively, the predicted scores for samples i and j, and $\text{sign}(x)$ equals $+1$ if $x > 0$, and -1 otherwise, indicating the correct ranking. The second is the proposed *quality-aware adaptive margin ranking loss*, which is defined by

$$\mathcal{L}_{\text{QAMRO}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} q_{ij} \cdot \max\Big(0,\ -\text{sign}(y_i - y_j) \cdot (\hat{y}_i - \hat{y}_j) + \alpha \cdot |y_i - y_j|\Big), \tag{2}$$

where q_{ij} is a quality-aware weight, and α denotes a margin scaling factor. This formulation remedies the deficiency of the conventional pairwise ranking loss through two novel countermeasures. First, it employs a data-dependent margin (*cf.* the term $\alpha \cdot |y_i - y_j|$ in Eq. 2) that varies based on the absolute difference of the ground-truth MOS values, thereby capturing perceptual discrepancies. Second, a quality-aware weighting mechanism is incorporated:

$$q_{ij} = 1 + (\beta - 1) \cdot \max(\tilde{y}_i, \tilde{y}_j), \tag{3}$$

where $\tilde{y}_i$ and $\tilde{y}_j \in [0, 1]$ are the normalized ground-truth MOS scores, and the preference factor $\beta > 1$ modulates the strength of the weights. This accentuates sample pairs that contain at least one high-quality utterance, thus encouraging the model to rank such cases more reliably.

In summary, $\mathcal{L}_{\text{MR}}$ serves as a baseline with a uniform margin and equal weighting, assuming a fixed perceptual threshold for distinguishing the score ranking. As a soft-margin formulation, it allows for subtle violations of the target ordering while penalizing larger discrepancies. Our $\mathcal{L}_{\text{QAMRO}}$ improves on this by adapting the margin according to the perceptual gap $|y_i - y_j|$ and giving higher weights to high-quality samples via the weighting term q_{ij}. This reflects the intuition that human perceptual differences vary with the score distance and that errors in high-quality regions are usually more critical.

B. MOS Prediction Tasks Across Audio Generation Systems

To demonstrate the generalizability and effectiveness of our proposed optimization strategy, we first apply it to a MOS prediction model for TTM systems. The model is built upon the pre-trained CLAP model [45], which captures audio-text joint semantics. The music clips are encoded using the CLAP audio encoder, while the textual prompts are embedded via the CLAP text encoder. Two separate three-layer multi-layer perceptrons (MLPs) are employed: one regresses the musical impression score from the audio-only embeddings, and the other estimates the text alignment score using the concatenated audio and text embeddings. The training objective combines our proposed QAMRO loss with the Huber loss [46]. The QAMRO loss enforces consistency with human perceptual rankings, especially in high-quality regions, while the Huber loss enhances robustness by reducing the impact of outliers.

Based on the success of this strategy for TTM, we extend our framework to a more general MOS prediction setting, including TTS, TTA, and TTM. We adopt the pre-trained Audiobox-Aesthetics model [47] as the backbone. Audio input streams are processed through a convolutional encoder and a Transformer encoder to extract deep audio representations. These embeddings are fed into four independent MLP prediction heads, each for one evaluation dimension: production quality, production complexity, content enjoyment, and content usefulness. The model is trained with a multi-objective loss function combining our QAMRO loss with traditional regression objectives, consequently balancing perceptual alignment and score accuracy for different content types.

III. Experimental Setups

A. Datasets

Our experiments were conducted solely on the AudioMOS Challenge 2025. For the TTM MOS prediction task, we used the MusicEval dataset [18], which comprises 2,748 music clips generated by 31 text-to-music systems, each conditioned on one of 384 unique text prompts. Each clip was rated by five annotators selected from a pool of 14 music experts, along two perceptual dimensions: musical impression and alignment with the given text. To further evaluate the generalizability of our framework across various types of audio content, we employed the AES-Natural dataset [47], which contains 2,950 audio samples covering speech, music, and general audio content. Specifically, it includes 950 speech samples (from EARS, LibriTTS, and Common Voice 13.0), 1,000 music samples (from MUSDB18-HQ and MusicCaps), and 1,000 general audio samples from AudioSet. Each sample was annotated by 10 experts with backgrounds in audio or music. To support multi-faceted evaluation, each clip was rated along four perceptual axes: production quality, production complexity, content enjoyment, and content usefulness.

B. Configurations

All models were trained with a batch size of 256 using the SGD optimizer [48] and a fixed learning rate of 0.0005. For the proposed ranking loss, we set $\alpha = 0.2$ and $\beta = 7.0$ in Eqs. 2 and 3, respectively. Early stopping was applied if the validation loss did not improve for 20 epochs. Our code is available at https://github.com/JethroWangSir/QAMRO.

C. Evaluation Metrics

We evaluated system performance using several objective system-level metrics: MSE, linear correlation coefficient (LCC), Spearman rank correlation coefficient (SRCC), and Kendall Tau rank correlation (KTAU) [49]. These metrics are computed by first averaging MOS predictions across all clips generated by each system, then comparing these system-level predictions with corresponding human-annotated system-level MOS scores. System-level evaluation was adopted because, in real-world scenarios, assessing and ranking complete speech systems offers greater practical relevance than evaluating individual clips. MSE ranges from 0 to ∞, with lower values indicating better performance. LCC, SRCC, and KTAU range

TABLE I

RESULTS AND ABLATIONS ON THE MUSICEVAL DATASET, EVALUATING MUSICAL IMPRESSION AND TEXTUAL ALIGNMENT. NOTE THAT "W/O ADAPTIVE MARGIN" DENOTES REPLACING THE PROPOSED MARGIN-BASED RANKING LOSS $\mathcal{L}_{\text{QAMRO}}$ WITH A FIXED-MARGIN RANKING LOSS $\mathcal{L}_{\text{MR}}$.

Model	**Musical Impression**				**Textual Alignment**			
	MSE ↓	**LCC** ↑	**SRCC** ↑	**KTAU** ↑	**MSE** ↓	**LCC** ↑	**SRCC** ↑	**KTAU** ↑
MusicEval-baseline [18]	0.247	0.851	0.845	0.655	0.143	0.819	0.779	0.578
QAMRO	**0.139**	**0.961**	**0.972**	**0.876**	0.109	**0.918**	**0.916**	**0.763**
w/o Quality-aware Weighting	0.172	0.910	0.924	0.779	**0.101**	0.894	0.885	0.707
w/o Adaptive Margin	0.214	0.886	0.875	0.687	0.134	0.875	0.838	0.643

TABLE II

RESULTS AND ABLATIONS ON THE AES-NATURAL DATASET, REPORTING SYSTEM-LEVEL SRCC FOR PQ, PC, CE, AND CU.

Model	**PQ**	**PC**	**CE**	**CU**
Audiobox-Aesthetics [47]	0.864	0.938	0.845	0.809
QAMRO	**0.883**	**0.942**	**0.869**	**0.852**
w/o Quality-aware Weighting	0.874	0.940	0.861	0.843
w/o Adaptive Margin	0.871	0.939	0.853	0.829

from -1 to 1, where higher values indicate stronger correlation and better alignment with human judgments. SRCC, in particular, reflects ranking consistency across systems.

IV. RESULTS AND DISCUSSION

A. Performance Analysis on MusicEval

Table I reports the results and ablations of our proposed framework, QAMRO, on the MusicEval dataset. QAMRO achieves the highest overall performance, with SRCC scores of 0.972 for musical impression and 0.916 for textual alignment, outperforming the MusicEval-baseline [18]. For a fair comparison, both our framework and the baseline used the Huber loss for score regression. These results demonstrate the effectiveness of our quality-aware adaptive margin ranking optimization strategy. Although SRCC is most directly influenced by our ranking loss, our framework also improves other metrics like MSE, LCC, and KTAU. This shows that it improves regression quality without sacrificing it, strengthening overall learning rather than focusing solely on ranking.

Ablation studies further validate the contribution of each component. Removing the quality-aware weighting leads to a noticeable drop in performance for both musical impression and textual alignment. Further excluding the adaptive margin mechanism results in an even larger degradation. These results demonstrate the importance of both the quality-aware weighting and adaptive margin for optimal ranking performance.

Furthermore, we analyzed the influence of the preference factor β in Fig. 1, which controls the emphasis on high-quality samples via the weighting term q_{ij} in Eq. 3. Performance peaks at $\beta = 7.0$, suggesting that moderate preference weighting improves ranking precision in perceptually sensitive regions. Larger values slightly degrade performance due to overemphasis on high scores. These findings highlight the importance of tuning β to balance fine-grained discrimination and global consistency. Notably, all tested β values consistently outperform the baseline without quality-aware weighting, demonstrating the effectiveness of our framework.

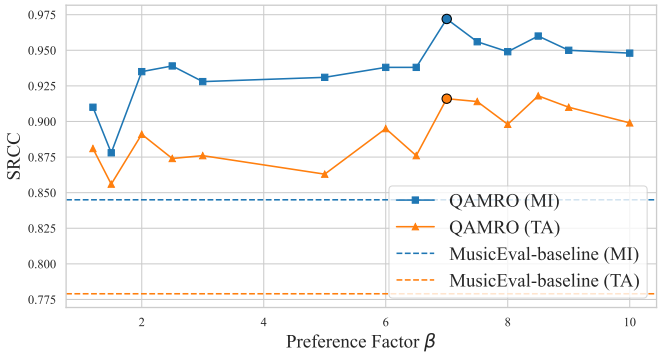


Fig. 1. Effect of the preference factor β on the performance of QAMRO in terms of musical impression (MI) and textual alignment (TA) on the MusicEval dataset. Baseline scores are shown for reference.

B. Evaluation Results on AES-Natural

Table II presents the results on the AES-Natural dataset, evaluating production quality (PQ), production complexity (PC), content enjoyment (CE), and content usefulness (CU) using SRCC. QAMRO achieves the best performance across all dimensions, with notable improvements over the Audiobox-Aesthetics baseline [47], particularly on content usefulness. Moreover, our framework consistently shows effectiveness across different model architectures and various MOS score aspects, demonstrating its robustness and generalizability.

The ablation studies further validate the effectiveness of each component, using the same ablation settings as in Table I. The removal of the quality-aware weighting consistently degrades performance across evaluation aspects, with a more pronounced impact observed when the adaptive margin mechanism is also excluded, especially for content usefulness. These results highlight the importance of both the quality-aware weighting and adaptive margin in capturing nuanced perceptual and semantic aspects of music content.

V. CONCLUSION AND FUTURE WORK

This study presents QAMRO, a ranking loss with quality-aware adaptive margins for human-aligned audio quality assessment. Evaluations on the AudioMOS Challenge 2025 benchmarks reveal that QAMRO significantly outperforms strong baselines in correlating with human MOS scores, particularly for high-quality samples, demonstrating the first effective use of ranking loss in this context. As to future work, we envisage to explore listwise ranking methods apart from our current pairwise strategy, which consider global ranking structures and are anticipated to improve alignment with human judgments across diverse generated audio.

REFERENCES

[1] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "MusicLM: Generating music from text," in *Arxiv preprint arXiv:2301.11325*, 2023.

[2] H.-W. Dong, K. Chen, S. Dubnov, J. McAuley, and T. Berg-Kirkpatrick, "Multitrack music transformer," in *Proc. ICASSP*, 2023.

[3] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X.-Y. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and S. Zhao, "NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in *Proc. ICML*, 2024.

[4] S. Chen, S. Liu, L. Zhou, Y. Liu, X. Tan, J. Li, S. Zhao, Y. Qian, and F. Wei, "VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers," in *Arxiv preprint arXiv:2406.05370*, 2024.

[5] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," in *Arxiv preprint arXiv:2410.06885*, 2025.

[6] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *Proc. ICML*, 2023.

[7] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "SoundStorm: Efficient parallel audio generation," in *Arxiv preprint arXiv:2305.09636*, 2023.

[8] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. ICML*, 2023.

[9] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Defossez, "Simple and controllable music generation," in *Proc. NeurIPS*, 2023.

[10] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank, J. Engel, Q. V. Le, W. Chan, Z. Chen, and W. Han, "Noise2Music: Text-conditioned music generation with diffusion models," in *Arxiv preprint arXiv:2302.03917*, 2023.

[11] S. Ji, X. Yang, and J. Luo, "A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges," *ACM Comput. Surv.*, vol. 56, no. 1, pp. 7:1–7:39, 2023.

[12] P. Zhu, C. Pang, Y. Chai, L. Li, S. Wang, Y. Sun, H. Tian, and H. Wu, "ERNIE-music: Text-to-waveform music generation with diffusion models," in *Proc. AACL Demo*, 2023.

[13] P. Lu, X. Xu, C. Kang, B. Yu, C. Xing, X. Tan, and J. Bian, "MuseCoco: Generating symbolic music from text," in *Arxiv preprint arXiv:2306.00110*, 2023.

[14] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, "Moûsai: Text-to-music generation with long-context latent diffusion," in *Arxiv preprint arXiv:2301.11757*, 2023.

[15] A. Ziv, I. Gat, G. L. Lan, T. Remez, F. Kreuk, A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "Masked audio generation using a single non-autoregressive transformer," in *Proc. ICLR*, 2024.

[16] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," in *Arxiv preprint arXiv:1812.08466*, 2019.

[17] S. Barratt and R. Sharma, "A note on the inception score," in *Proc. ICML*, 2018.

[18] C. Liu, H. Wang, J. Zhao, S. Zhao, H. Bu, X. Xu, J. Zhou, H. Sun, and Y. Qin, "MusicEval: A generative music dataset with expert ratings for automatic text-to-music evaluation," in *Proc. ICASSP*, 2025.

[19] W.-C. Tseng, C.-y. Huang, W.-T. Kao, Y. Y. Lin, and H.-y. Lee, "Utilizing self-supervised representations for MOS prediction," in *Proc. Interspeech*, 2021.

[20] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-saruLab system for voiceMOS challenge 2022," in *Proc. Interspeech*, 2022.

[21] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2022.

[22] S.-W. Fu, K.-H. Hung, Y. Tsao, and Y.-C. F. Wang, "Self-supervised speech quality estimation and enhancement wsing only clean speech," in *Proc. ICLR*, 2023.

[23] W. Ariyanti, K.-Y. Chen, S. M. Siniscalchi, H.-M. Wang, and Y. Tsao, "Towards robust assessment of pathological voices via combined low-level descriptors and foundation model representations," in *Arxiv preprint arXiv: 2505.21356*, 2025.

[24] A. Menon, X. Jiang, S. Vembu, C. Elkan, and L. Ohno-Machado, "Predicting accurate probabilities with a ranking loss," in *Proc. ICML*, 2012.

[25] M. Engilberge, L. Chevallier, P. Pérez, and M. Cord, "SoDeep: A sorting deep net to learn ranking loss surrogates," in *Proc. CVPR*, 2019.

[26] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," in *Proc. CVPR*, 2019.

[27] Q. Qian, L. Chen, H. Li, and R. Jin, "DR loss: Improving object detection by distributional ranking," in *Proc. CVPR*, 2020.

[28] K. Oksuz, B. C. Cam, E. Akbas, and S. Kalkan, "A ranking-based, balanced loss function unifying classification and localisation in object detection," in *Proc. NeurIPS*, 2020.

[29] R. Swezey, A. Grover, B. Charron, and S. Ermon, "PiRank: Scalable learning to rank via differentiable sorting," in *Proc. NeurIPS*, 2021.

[30] S. Bruch, "An alternative cross entropy loss for learning-to-rank," in *Proc. WWW*, 2021.

[31] K. Oksuz, B. C. Cam, E. Akbas, and S. Kalkan, "Rank & sort loss for object detection and instance segmentation," in *Proc. ICCV*, 2021.

[32] G. Wu, C. LI, K. Xu, and J. Zhu, "Rethinking and reweighting the univariate losses for multi-label ranking: Consistency and generalization," in *Proc. NeurIPS*, 2021.

[33] Z. Liu, Y. Wang, K. Han, W. Zhang, S. Ma, and W. Gao, "Post-training quantization for vision transformer," in *Proc. NeurIPS*, 2021.

[34] D. Zhu, X. Wu, and T. Yang, "Benchmarking deep AUROC optimization: Loss functions and algorithmic choices," in *Arxiv preprint arXiv:2203.14177*, 2022.

[35] X.-R. Sheng, J. Gao, Y. Cheng, S. Yang, S. Han, H. Deng, Y. Jiang, J. Xu, and B. Zheng, "Joint optimization of ranking and calibration with contextualized hybrid model," in *Proc. KDD*, 2023.

[36] S. Hu, X. Wang, and S. Lyu, "Rank-based decomposable losses in machine learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 599–13 620, 2023.

[37] R. Xiao, Y. Ge, R. Jiang, and Y. Yan, "A unified framework for rank-based loss minimization," in *Proc. NeurIPS*, 2023.

[38] F. Yavuz, B. C. Cam, A. H. Dogan, K. Oksuz, E. Akbas, and S. Kalkan, "Bucketed ranking-based losses for efficient training of object detectors," in *Proc. ECCV*, 2024.

[39] F. Durmus, H. Saribas, S. Aldemir, J. Yang, and H. Cevikalp, "Pairwise ranking loss for multi-task learning in recommender systems," in *Arxiv preprint arXiv:2406.02163*, 2024.

[40] R.-X. Tan, K. Xue, S.-H. Lyu, H. Shang, Y. Wang, Y. Wang, S. Fu, and C. Qian, "Offline model-based optimization by learning to rank," in *Proc. ICLR*, 2025.

[41] T. Peng and J. Zhang, "Enhancing knowledge distillation of large language models through efficient multi-modal distribution alignment," in *Proc. COLING*, 2025.

[42] C. Burges, R. Ragno, and Q. Le, "Learning to rank with nonsmooth cost functions," in *Proc. NeurIPS*, 2006.

[43] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. ICML*, 2007.

[44] C. J. C. Burges, "From RankNet to LambdaRank to LambdaMART: An overview," *Learning*, vol. 11, p. 81, 2010.

[45] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. ICASSP*, 2023.

[46] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in Statistics: Methodology and Distribution*, S. Kotz and N. L. Johnson, Eds., 1992, pp. 492–518.

[47] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W.-N. Hsu, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," in *Arxiv preprint arXiv:2502.05139*, 2025.

[48] S. Ruder, "An overview of gradient descent optimization algorithms," in *Arxiv preprint arXiv:1609.04747*, 2017.

[49] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. ICASSP*, 2022.