DeRA-MOS: Optimizing Text-to-Music Evaluation via Decoupled Listwise Ranking and Modality Alignment

Chien-Chun Wang, *Member, IEEE,* Hung-Shin Lee, Hsin-Min Wang, *Senior Member, IEEE,* and Berlin Chen, *Member, IEEE*

***Abstract*—Evaluating text-to-music (TTM) systems remains expensive because music impression (MI) and text alignment (TA) scores rely on human mean opinion scores (MOS). Most automatic MOS estimators are trained with point-wise regression or distributional classification. These objectives do not directly optimize rank-based metrics and provide weak geometric constraints for cross-modal coherence. To address these gaps, we propose DeRA-MOS, a decoupled optimization framework for TTM evaluation. For MI, we introduce a batch-aware listwise ranking loss that models relative order within each mini-batch and better aligns with evaluation based on Spearman's rank correlation coefficient (SRCC). For TA, we introduce a score-anchored modality alignment loss that maps human scores to target audio-text similarity and regularizes the latent space before fusion. By effectively mitigating the point-wise training mismatch and modality drift, experiments on MusicEval demonstrate that our decoupled framework yields substantial improvements in both MI and TA ranking metrics, establishing a robust paradigm for large-scale TTM evaluation.**

***Index Terms*—Text-to-music evaluation, mean opinion score, listwise ranking, cross-modal alignment, metric learning.**

I. Introduction

THE proliferation of text-to-music (TTM) generative models [1]–[4] enables high-fidelity synthesis from natural language, yet evaluation lags behind. Quality assessment remains subjective, relying heavily on human mean opinion scores (MOS) for music impression (MI) and text alignment (TA) [5]. Given costly annotations, accurate automatic MOS estimation is essential for scalable TTM development [6], [7].

To automate assessment, previous efforts framed MOS prediction as regression. Traditional approaches use point-wise objectives (e.g., L1 or MSE) to predict single scalars, standard in speech synthesis (e.g., MOSNet [8], MBNet [9], UTMOS [7]). Recognizing human ratings as discrete and ordinal, recent methods shifted toward distributional [10] and listener-dependent modeling [11]. For instance, DeePMOS-$\mathcal{B}$ [12] uses Beta distributions capturing annotator variance, while CORAL [13] predicts cumulative probabilities. Recently, DORA-MOS [14] proposed Gaussian label softening over discrete bins to implicitly encode MOS ordinality. Similarly, objective TA evaluation frequently uses reference-free metrics like Fréchet audio distance (FAD) [15], [16]. Contemporary models often use pre-trained joint embeddings (e.g., CLAP [17], MuQ [18]) and cross-attention to implicitly predict coherence.

Despite these improvements, a mismatch persists between training objectives and evaluation metrics. First, metrics like Spearman's rank correlation coefficient (SRCC) evaluate relative order, not absolute scores. Even with ordinal smoothing [19], [20], point-wise training treats samples independently, ignoring batch-level ranking structure. Second, cross-attention-based TA estimation lacks explicit geometric constraints on unimodal embeddings. Unanchored to human judgment, these representations can drift, degrading semantic alignment.

To address these challenges, we propose **DeRA-MOS** (**De**coupled **R**anking and **A**lignment)[1], a framework decoupling training objectives into two specialized losses while keeping the backbone unchanged. Though inspired by established metric learning, our novelty lies in their targeted, decoupled combination to address TTM evaluation bottlenecks. For MI, the Batch-Aware Listwise Ranking (BALR) loss treats each mini-batch as a query list to optimize global ranking via learning-to-rank principles [21], [22]. For TA, the Score-Anchored Modality Alignment (SAMA) loss aligns pre-fusion cosine similarity of L2-normalized embeddings with human TA MOS [23]. This formulation directly targets ranking metrics while stabilizing cross-modal geometry.

In summary, our main contributions are:

1) **Batch-Aware Listwise Ranking:** We replace point-wise MI supervision with batch-aware listwise ranking, directly aligning with SRCC evaluation.
2) **Score-Anchored Modality Alignment:** We anchor latent audio-text similarity to TA MOS, mitigating representation drift and enhancing semantic coherence.
3) **Robust Performance with DeRA-MOS:** On MusicEval, the unified DeRA-MOS framework consistently outperforms a reproduced homogeneous point-wise baseline in terms of SRCC for both MI and TA tasks.

II. Proposed Methodology

While DORA-MOS effectively models TTM evaluation via a dual-branch structure, its optimization objective is sub-optimal. Specifically, it uses a point-wise Gaussian-softened classification loss. Though capturing MOS's ordinal nature, it treats samples independently, ignoring batch-level ranking dynamics. Furthermore, for TA, it learns modality coherence implicitly via cross-attention, lacking a direct geometric constraint. We introduce two complementary objectives that

Chien-Chun Wang is with E.SUN Financial Holding Co., Ltd., Taipei 10546, Taiwan (e-mail: jethrowang0531@gmail.com).

Hung-Shin Lee is with United Link Co., Ltd., Taipei 11493, Taiwan (Corresponding author; e-mail: hungshinlee@gmail.com).

Hsin-Min Wang is with the Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan (e-mail: whm@iis.sinica.edu.tw).

Berlin Chen is with the Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei 11677, Taiwan (Corresponding author; e-mail: berlin@ntnu.edu.tw).

[1]Our code is available at: https://github.com/JethroWangSir/DeRA-MOS.

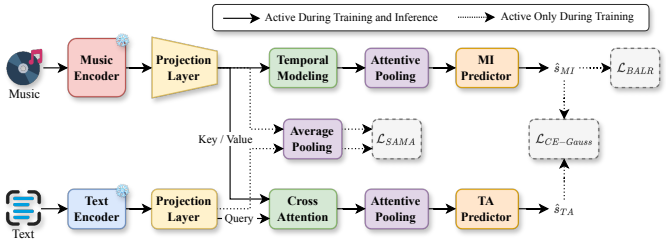


Fig. 1. **Overview of the proposed DeRA-MOS framework.** The architecture explicitly decouples the TTM evaluation process using task-specific objectives (dotted paths denote training-only operations). To overcome the limitations of homogeneous point-wise learning ($\mathcal{L}_{CE-Gauss}$), we introduce $\mathcal{L}_{BALR}$ to enforce batch-aware global ranking for the MI branch. Crucially, for the TA branch, $\mathcal{L}_{SAMA}$ is applied *before* cross-attention fusion to geometrically anchor the unimodal representations, preventing latent drift and ensuring robust cross-modal coherence.

explicitly enforce global ranking and cross-modal geometric consistency, as illustrated in Fig. 1.

A. Batch-Aware Listwise Ranking for Music Impression

The evaluation of music impression is inherently a relative ranking task. A model predicting shifted absolute values can still achieve a perfect SRCC if the rank order is preserved. Point-wise objectives fail to optimize this metric directly. To reduce the mismatch between training and SRCC-oriented evaluation, we propose a batch-aware listwise ranking loss. Inspired by foundational learning-to-rank paradigms such as ListNet [21], [22] and LambdaRank [24], we conceptualize the entire mini-batch as a single query list. We transform both ground-truth and predicted MI scores into batch-wise probability distributions using temperature-scaled softmax. Let $\mathbf{s}_{MI} = [s_{MI}^{(1)}, s_{MI}^{(2)}, \ldots, s_{MI}^{(B)}]^T$ be the ground-truth MI scores for a batch of size B, and $\hat{\mathbf{s}}_{MI} = [\hat{s}_{MI}^{(1)}, \hat{s}_{MI}^{(2)}, \ldots, \hat{s}_{MI}^{(B)}]^T$ be the corresponding model predictions. For sample i, the target and predicted listwise probabilities are defined as

$$p_i = \frac{\exp(s_{MI}^{(i)}/\tau)}{\sum_{j=1}^{B} \exp(s_{MI}^{(j)}/\tau)}, \quad \hat{p}_i = \frac{\exp(\hat{s}_{MI}^{(i)}/\tau)}{\sum_{j=1}^{B} \exp(\hat{s}_{MI}^{(j)}/\tau)}, \tag{1}$$

where $\tau > 0$ is the temperature. When MOS values are densely concentrated (e.g., between 3.0 and 4.0), τ controls the sharpness of the induced distribution and therefore the sensitivity to small rank differences. The BALR loss is defined as the cross-entropy between $\{p_i\}_{i=1}^{B}$ and $\{\hat{p}_i\}_{i=1}^{B}$:

$$\mathcal{L}_{BALR} = -\sum_{i=1}^{B} p_i \log \hat{p}_i. \tag{2}$$

Minimizing $\mathcal{L}_{BALR}$ encourages correct global ordering within each batch, rather than only absolute score matching, and thus provides gradients more consistent with SRCC. Theoretically, this continuous approximation translates discrete rank metrics into differentiable objectives. Unlike pairwise methods (e.g., QAMRO [25]) that evaluate $\mathcal{O}(B^2)$ isolated local pairs without full list context, BALR elegantly captures global batch dynamics while reducing complexity to $\mathcal{O}(B)$.

B. Score-Anchored Modality Alignment for Text Alignment

For predicting the TA score, previous methods fuse audio and text representations using cross-attention, assuming the network will implicitly discover semantic correspondence. However, without an explicit constraint on the individual unimodal spaces before fusion, the embeddings can drift, leading to representations that memorize the training set rather than learning true modality coherence. To regularize the latent space, we propose the score-anchored modality alignment loss. The key idea is to align latent audio-text similarity with human TA MOS before cross-attention fusion [26]. Let $\mathbf{h}_a^{(i)}$ and $\mathbf{h}_t^{(i)}$ denote the pre-fusion audio and text embeddings for sample i, which are obtained by applying average pooling over their respective temporal features. First, we compute cosine similarity as

$$\text{sim}(\mathbf{h}_a^{(i)}, \mathbf{h}_t^{(i)}) = \frac{\mathbf{h}_a^{(i)} \cdot \mathbf{h}_t^{(i)}}{\|\mathbf{h}_a^{(i)}\|_2 \|\mathbf{h}_t^{(i)}\|_2}. \tag{3}$$

Since cosine similarity ranges from -1 to 1, we linearly transform it to the range $[0, 1]$ using $\tilde{c}^{(i)} = (\text{sim}(\mathbf{h}_a^{(i)}, \mathbf{h}_t^{(i)}) + 1)/2$. Correspondingly, we linearly map the human-annotated TA MOS $s_{TA}^{(i)}$ (bounded within $[1, 5]$) to $y_{sim}^{(i)} \in [0, 1]$:

$$y_{sim}^{(i)} = \frac{s_{TA}^{(i)} - 1.0}{4.0}. \tag{4}$$

The SAMA loss is then defined as

$$\mathcal{L}_{SAMA} = \frac{1}{B} \sum_{i=1}^{B} \left(\tilde{c}^{(i)} - y_{sim}^{(i)} \right)^2. \tag{5}$$

This parameter-free mapping empirically outperforms nonlinear alternatives. Empirical sensitivity analysis further confirmed that adopting L2 normalization and explicitly anchoring to the $[0, 1]$ range prevented the instability and divergence observed in unnormalized mappings. It bounds the space and mitigates drift without overfitting.

C. Joint Optimization

The overall training objective combines our proposed losses with the baseline task-specific losses. Let $\mathcal{L}_{CE-Gauss}$ denote the sum of the baseline Gaussian-softened classification losses for both the MI and TA branches. The final loss function is defined as

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{CE-Gauss} + \alpha \mathcal{L}_{BALR} + \beta \mathcal{L}_{SAMA}, \tag{6}$$

where $\alpha \geq 0$ and $\beta \geq 0$ are weighting coefficients balancing the contribution of the global ranking constraint and the geometric alignment constraint, respectively. During training, $\mathcal{L}_{BALR}$ is applied to MI predictions, while $\mathcal{L}_{SAMA}$ is applied to pre-fusion TA embeddings. During inference, both auxiliary losses are removed, and the prediction pipeline remains identical to the baseline.

III. Experimental Setups

A. Dataset and Evaluation Metrics

Our DeRA-MOS framework was evaluated on the MusicEval dataset [5], the official benchmark for the AudioMOS 2025 Challenge [28]. The dataset comprised outputs from 31 distinct TTM systems generated from 384 fixed prompts, each annotated by experts for MI and TA. Unlike previous work that relied on custom stratified splits for internal validation [14],

TABLE I
PERFORMANCE EVALUATION ON MUSICEVAL. MODELS MARKED WITH AN ASTERISK (*) DENOTE OFFICIAL RESULTS REPORTED BY THE CHALLENGE ORGANIZERS OR RESPECTIVE AUTHORS, WHILE UNMARKED MODELS INDICATE OUR LOCALLY REPRODUCED BASELINES AND PROPOSED VARIANTS. FOR MSE, LOWER IS BETTER (↓); FOR ALL OTHER METRICS, HIGHER IS BETTER (↑). BEST OVERALL RESULTS ARE HIGHLIGHTED IN **BOLD**.

Model	Music Impression				Text Alignment			
	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑
MusicEval-Baseline* [5]	0.378	0.821	0.818	0.623	0.199	0.744	0.724	0.532
DRASP* [27]	0.076	0.949	0.957	0.858	0.058	0.897	0.890	0.726
QAMRO* [25]	0.139	0.961	0.972	0.876	0.109	0.918	0.916	0.763
DORA-MOS* [14]	**0.017**	0.986	0.988	0.913	0.033	0.946	0.944	0.809
DORA-MOS (Reproduced)	0.018	0.985	0.981	0.890	0.060	0.956	0.952	**0.835**
+ Ranking ($\mathcal{L}_{BALR}$)	0.030	0.985	0.985	0.908	0.031	0.946	0.940	0.789
+ Alignment ($\mathcal{L}_{SAMA}$)	0.031	0.985	0.983	0.908	0.030	0.956	0.954	0.832
DeRA-MOS (Full)	0.018	**0.989**	**0.989**	**0.940**	**0.028**	**0.958**	**0.956**	**0.835**

we strictly adhered to the official train, development, and test data partitions to ensure standardized benchmarking.

We evaluated performance using four metrics: SRCC, Kendall's Tau (KTAU) [29], linear correlation coefficient (LCC), and MSE [30]. While SRCC and KTAU were our primary metrics for validating the global ranking capability introduced by BALR, LCC and MSE were included to assess linear tracking and absolute prediction accuracy. This evaluation-metric suite ensured that our decoupled framework not only ranked systems correctly but also maintained absolute score fidelity through SAMA's geometric anchoring.

B. Implementation Details

To rigorously isolate the effectiveness of our task-decoupled optimization strategy, we adopted the optimal dual-branch architecture from the DORA-MOS baseline [14] as our fixed backbone. Specifically, audio features were extracted using the frozen pre-trained MuQ model [18], and text embeddings were encoded via the frozen RoBERTa model [31]. The temporal dependencies were modeled using a Transformer encoder paired with attention pooling, serving as a shared backbone before splitting into task-specific MI and TA prediction heads. By fixing the architectural components, we minimized confounding factors so that observed gains could be attributed primarily to BALR and SAMA. Applied immediately after the initial linear projections, $\mathcal{L}_{SAMA}$ imposes a pre-fusion geometric constraint. It acts as a score-conditioned regularizer, pulling high-TA pairs closer while separating low-TA pairs, to preserve cross-modal structure and suppress representation drift. Since $\mathcal{L}_{BALR}$ and $\mathcal{L}_{SAMA}$ serve purely as training regularizers, DeRA-MOS maintains the exact same inference complexity as the baseline (zero additional parameters and FLOPs), ensuring highly efficient deployment.

C. Training Configuration

The network was optimized using an AdamW optimizer [32] with a learning rate of 5×10^{-5}. Crucially, since the BALR loss conceptualizes the mini-batch as a query list for global sorting, we utilized a batch size of 32. Under standard random shuffling, the selected batch size provided a sufficiently diverse distribution of MOS scores within each step to compute a meaningful ranking gradient. For the hyperparameter settings, the temperature in the BALR softmax distribution was empirically set to $\tau = 1.0$. The overall loss function integrated the task-decoupled objectives with weighting coefficients $\alpha = 0.2$ and $\beta = 0.3$, selected by development-set tuning. The training process spanned a maximum of 100 epochs, with early stopping (patience: 15 epochs) based on validation loss to prevent overfitting. Evaluating with five random seeds yielded highly stable MI and TA SRCCs (0.989 ± 0.002 and 0.956 ± 0.003, respectively). We also conducted significance testing on paired SRCC outcomes between DeRA-MOS and the reproduced baseline.

IV. RESULTS AND DISCUSSION

A. Comparisons with State-of-the-Art Systems

Table I presents a comprehensive comparison between our proposed framework and state-of-the-art systems from the AudioMOS 2025 Challenge [28]. To ensure a rigorous and transparent evaluation, we included both the officially reported numbers (denoted with *) and our reproduced DORA-MOS baseline. The minor performance drop of our reproduced baseline compared to the official report is attributed to our strict adherence to standardized data splits and unified hyperparameter tuning, necessary for a fair ablation analysis. Although the MI SRCC improvement over the official DORA-MOS* is numerically small, DeRA-MOS matches this level of performance without increasing architectural complexity. Crucially, compared with the reproduced baseline, it improves KTAU from 0.890 to 0.940 ($+0.050$), indicating substantially fewer rank inversions while preserving absolute score fidelity (MSE: 0.018). For TA, DeRA-MOS improves SRCC from 0.952 to 0.956 and reduces MSE from 0.060 to 0.028. A system-level paired Wilcoxon signed-rank test [33] shows that the MI/TA SRCC gains over the reproduced baseline are statistically significant ($p < 0.01$).

Beyond system-level evaluation, we also investigated the utterance-level performance to assess fine-grained quality ranking. For music impression, DeRA-MOS achieves an utterance-level MI SRCC of 0.854, outperforming the reproduced baseline (0.845). This utterance-level MI gain indicates that BALR improves local relative quality ordering. However, utterance-level TA SRCC decreases slightly ($0.632 \rightarrow 0.605$). Although $\mathcal{L}_{SAMA}$ stabilizes global geometry, the batch-level pressure induced by $\mathcal{L}_{BALR}$ can oversmooth prompt-specific nuances. Importantly, adjusting the α weight provides a controllable trade-off between global ordering and fine-grained

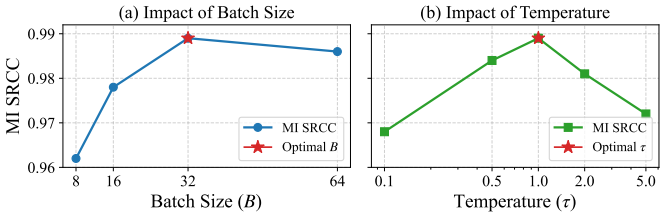


Fig. 2. Hyperparameter sensitivity analysis of DeRA-MOS on MusicEval. (a) The impact of batch size (B) on MI SRCC. (b) The impact of the temperature scaling factor (τ) in the BALR loss on MI SRCC.

local accuracy, which we plan to further optimize using multi-task gradient surgery [34].

B. Ablation Study: The Synergy of Decoupling

To unpack the contributions of our proposed objectives, we systematically ablated DeRA-MOS from our reproduced baseline (Table I, bottom section). This ablation analysis isolates the role of each decoupled component.

The cost of isolated ranking: Using only +Ranking ($\mathcal{L}_{BALR}$) improved MI ordering (KTAU: $0.890 \rightarrow 0.908$), confirming that listwise supervision was effective for rank consistency. However, TA SRCC dropped ($0.952 \rightarrow 0.940$), indicating that ranking pressure alone could distort shared cross-modal representations.

The role of geometric anchoring: Using only +Alignment ($\mathcal{L}_{SAMA}$) sharply improved TA calibration (MSE: $0.060 \rightarrow 0.030$) while preserving strong ranking metrics. This ablation result confirmed that geometric anchoring enforced meaningful audio-text correspondence before fusion.

Synergy in DeRA-MOS: Combining both objectives yielded the best overall profile. The full model recovered MI calibration (MSE 0.018), reached the highest MI KTAU (0.940), and improved TA robustness (MSE 0.028, SRCC 0.956). These results showed that $\mathcal{L}_{BALR}$ and $\mathcal{L}_{SAMA}$ were complementary: BALR strengthened global ordering, while SAMA constrained geometry to prevent representation drift.

C. Hyperparameter Sensitivity Analysis

To demonstrate the robustness of our framework and validate the mathematical intuition behind the BALR loss, we analyzed the impact of its two core hyperparameters: batch size (B) and temperature (τ).

Impact of Batch Size: As shown in Fig. 2 (a), MI SRCC degrades noticeably when $B \leq 16$, because a small mini-batch lacks sufficient distribution diversity to construct a meaningful query list, weakening the ranking gradient. The performance peaks at $B = 32$, providing adequate inter-sample dynamics, and saturates at $B = 64$, confirming that a moderately large batch is essential for effective global ranking.

Impact of Temperature Scaling: An excessively low temperature ($\tau = 0.1$) sharpens the softmax distribution toward a one-hot vector, impeding gradient flow (Fig. 2 (b)). Conversely, a high temperature ($\tau \geq 2.0$) overly smooths the probabilities, washing out subtle relative ranking signals among densely clustered MOS values. The empirical optimum at $\tau = 1.0$ strikes the best balance, providing theoretical and empirical stability by shaping a smooth yet discriminative probability distribution for the ranking landscape.

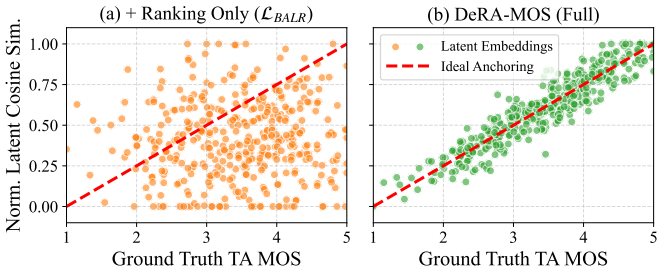


Fig. 3. Visualization of cross-modal representation drift and geometric anchoring. (a) Ranking alone ($\mathcal{L}_{BALR}$) leads to unconstrained representation drift. (b) Joint optimization with $\mathcal{L}_{SAMA}$ geometrically anchors the embeddings to the ideal trajectory (red dashed line).

Loss-Weighting Robustness: Beyond B and τ, we also examined the loss weights (α, β). Performance remained highly stable (MI SRCC variance < 0.002) across settings with $\alpha \in [0.1, 0.3]$ and $\beta \in [0.2, 0.4]$, confirming that DeRA-MOS maintains robust decoupling without exhaustive weight tuning.

D. Latent Space Analysis: Visualizing Representation Drift

To investigate the task interference observed in our ablation study, we visualized the latent cross-modal embedding space prior to cross-attention fusion. Fig. 3 plots the normalized latent cosine similarity $\tilde{c}^{(i)}$ (as described in Section II-B) against the ground truth TA MOS $s_{TA}^{(i)}$. The red dashed line represents the ideal anchoring trajectory defined in Eq. 4, which linearly maps the $[1, 5]$ subjective scores to the $[0, 1]$ geometric similarity space. As clearly shown in Fig. 3 (a), applying $\mathcal{L}_{BALR}$ without geometric constraints causes severe representation drift prior to cross-attention fusion. The latent similarities scatter widely with virtually no correlation to the ground truth, confirming that isolated ranking optimization disrupts the shared cross-modal semantic space. Conversely, Fig. 3 (b) demonstrates the efficacy of the $\mathcal{L}_{SAMA}$ objective. By explicitly penalizing deviations from the ideal trajectory, DeRA-MOS successfully anchors the previously drifting representations, forcing them to tightly converge along the red dashed line. This drift visualization confirms that explicit geometric anchoring is essential to prevent modality collapse during listwise optimization.

V. Conclusion and Future Work

In this letter, we proposed DeRA-MOS, a task-decoupled framework for TTM MOS evaluation. BALR aligns optimization with rank-based metrics, while SAMA enforces cross-modal geometric consistency. On MusicEval, DeRA-MOS significantly outperformed the baseline in SRCC/KTAU and improved absolute TA accuracy. Since both proposed losses are training-only, inference efficiency is preserved, making the framework ideal for large-scale model development.

Currently, our evaluation is limited to the MusicEval dataset due to a lack of other standardized TTM benchmarks. Future work will aim to validate this framework on broader, diverse datasets as they become available. Furthermore, we plan to extend this paradigm to broader audio domains, including text-to-speech [35], [36] and zero-shot sound effects evaluation, and integrate probabilistic modeling to better account for annotator error rates and subjective uncertainty [37].

REFERENCES

[1] 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, pp. 21 450–21 474.

[2] 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," 2023, arXiv:2301.11325. [Online]. Available: https://arxiv.org/abs/2301.11325

[3] 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, pp. 47 704–47 720.

[4] 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, pp. 13 916–13 932.

[5] 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, pp. 1–5.

[6] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Proc. Interspeech*, 2021, pp. 2127–2131.

[7] 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, pp. 4521–4525.

[8] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.

[9] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNET: MOS prediction for synthesized speech with mean-bias network," in *Proc. ICASSP*, 2021, pp. 391–395.

[10] J. H. M. Wong, H. Zhang, and N. F. Chen, "Modelling inter-rater uncertainty in spoken language assessment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2886–2898, 2023.

[11] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech," in *Proc. ICASSP*, 2022, pp. 896–900.

[12] X. Liang, F. Cumlin, V. Ungureanu, C. K. A. Reddy, C. Schüldt, and S. Chatterjee, "DeePMOS-B: Deep posterior mean-opinion-score Using beta distribution," in *Proc. EUSIPCO*, 2024, pp. 416–420.

[13] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognition Letters*, vol. 140, pp. 325–331, 2020.

[14] F. Ritter-Gutierrez, Y.-C. Lin, J.-C. Wei, J. H. M. Wong, N. F. Chen, and H.-y. Lee, "ASTAR-NTU solution to AudioMOS challenge 2025 track1," in *Proc. IEEE ASRU*, 2025.

[15] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *Proc. Interspeech*, 2019, pp. 2350–2354.

[16] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting frechet audio distance for generative music evaluation," in *Proc. ICASSP*, 2024, pp. 1331–1335.

[17] 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, pp. 1–5.

[18] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, "MuQ: Self-supervised music representation learning with mel residual vector quantization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3653–3664, 2025.

[19] X. Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[20] R. Diaz and A. Marathe, "Soft labels for ordinal regression," in *Proc. CVPR*, 2019, pp. 4738–4747.

[21] 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, pp. 129–136.

[22] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li, "Listwise approach to learning to rank: Theory and algorithm," in *Proc. ICML*, 2008, pp. 1192–1199.

[23] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *Proc. ICASSP*, 2022, pp. 976–980.

[24] C. Burges, R. Ragno, and Q. Le, "Learning to rank with nonsmooth cost functions," in *Proc. NeurIPS*, vol. 19, 2006.

[25] C.-C. Wang, K.-T. Huang, C.-Y. Yang, H.-S. Lee, H.-M. Wang, and B. Chen, "QAMRO: Quality-aware adaptive margin ranking optimization for human-aligned assessment of audio generation systems," in *Proc. IEEE ASRU*, 2025.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607.

[27] C.-Y. Yang, K.-T. Huang, C.-C. Wang, H.-S. Lee, H.-M. Wang, and B. Chen, "DRASP: A dual-resolution attentive statistics pooling framework for automatic MOS prediction," in *Proc. APSIPA ASC*, 2025, pp. 2038–2043.

[28] W.-C. Huang, H. Wang, C. Liu, Y.-C. Wu, A. Tjandra, W.-N. Hsu, E. Cooper, Y. Qin, and T. Toda, "The AudioMOS challenge 2025," in *Proc. IEEE ASRU*, 2025.

[29] M. G. Kendall, "A New Measure of Rank Correlation," *Biometrika*, 1938.

[30] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. ICASSP*, 2022, pp. 8442–8446.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, arXiv:1907.11692. [Online]. Available: http://arxiv.org/abs/1907.11692

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[33] R. F. Woolson, "Wilcoxon signed-rank test," in *Wiley Encyclopedia of Clinical Trials*, 2008, pp. 1–3.

[34] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in *Proc. NeurIPS*, vol. 33, 2020, pp. 5824–5836.

[35] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, 2021, pp. 5530–5540.

[36] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.

[37] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.