Robust Generative Audio Quality Assessment: Disentangling Quality from Spurious Correlations

Kuan-Tang Huang*, Chien-Chun Wang* 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 Information Science, Academia Sinica, Taiwan
‡United Link Co., Ltd., Taiwan

***Abstract*—The rapid proliferation of AI-Generated Content (AIGC) has necessitated robust metrics for perceptual quality assessment. However, automatic Mean Opinion Score (MOS) prediction models are often compromised by data scarcity, predisposing them to learn spurious correlations—such as dataset-specific acoustic signatures—rather than generalized quality features. To address this, we leverage domain adversarial training (DAT) to disentangle true quality perception from these nuisance factors. Unlike prior works that rely on static domain priors, we systematically investigate domain definition strategies ranging from explicit metadata-driven labels to implicit data-driven clusters. Our findings reveal that there is no "one-size-fits-all" domain definition; instead, the optimal strategy is highly dependent on the specific MOS aspect being evaluated. Experimental results demonstrate that our aspect-specific domain strategy effectively mitigates acoustic biases, significantly improving correlation with human ratings and achieving superior generalization on unseen generative scenarios.**

***Index Terms*—audio quality assessment, mean opinion score, domain adversarial training, robust generalization.**

I. Introduction

The exponential growth of AI-Generated Content (AIGC) has revolutionized the multimedia landscape. Particularly within the audio modality, generative audio has established itself as a cornerstone of modern content creation. This domain encompasses a diverse array of tasks, ranging from text-to-speech (TTS) to text-to-music (TTM) and universal text-to-audio (TTA). These advanced technologies are now driving a wide range of applications, such as creating dynamic soundscapes for immersive streaming, generating background scores for personalized media, and enabling interactive audio for virtual environments. However, accurately evaluating the perceptual quality of these generated sounds remains a critical challenge. While subjective listening tests yielding mean opinion scores (MOS) represent the gold standard for assessment, they are notoriously expensive and time-consuming to conduct. Consequently, the development of automatic MOS prediction models has become indispensable, serving as a scalable and efficient proxy for human ratings [1]–[3].

However, the reliability of these data-driven models is often severely compromised by the lack of large-scale subjectively labeled data. In such low-resource regimes, models are predisposed to learn spurious correlations rather than generalized quality features. Within limited training sets, high subjective ratings may coincidentally align with specific, non-quality-related acoustic characteristics intrinsic to the source data. For instance, a model might erroneously learn to associate high quality with the specific timbre of a musical instrument, the background ambiance of an environmental recording, or a particular room reverberation pattern in speech, simply because these traits dominate the highly rated samples in the limited training corpus. Consequently, the model overfits to these nuisance factors—treating them as proxies for quality. When deployed to unseen data where these specific acoustic signatures are absent, the model's predictions become unreliable, highlighting the need to disentangle true quality perception from these domain-specific biases.

The necessity of disentangling quality from confounding factors is echoed in adjacent fields. In video quality assessment, [4] attempts to separate aesthetics from technical quality by engineering specific input views. Similarly, in speaker recognition, recent works [5] have proposed specialized Gaussian inference layers to disentangle static speaker traits from dynamic linguistic content. However, these approaches often rely on either intricate, hand-crafted heuristics or complex, task-specific architectural designs to define the separation boundaries. To avoid reliance on rigid heuristics or task-specific designs, we introduce a generalized domain adversarial training (DAT) [6] framework capable of automatically purging dataset-induced biases. By leveraging this strategy, we enforce the model to discard these nuisance factors within the latent space, retaining only the features pertinent to intrinsic perceptual quality.

While DAT has been widely adopted in speech and audio tasks—such as addressing variations arising from different noise conditions [7], accents [8], or recording corpora [9], [10]—these works typically treat domain definition as a static prior. In contrast, we argue that for MOS prediction under data scarcity, the optimal definition of a "domain" is not self-evident. A critical yet underexplored question arises: what constitutes the most effective "domain" for adversarial training? Crucially, our experiments suggest there is no "one-size-fits-all" answer; instead, we find that different MOS aspects necessitate specific domain definition strategies to maximize prediction generalization.

To this end, we systematically investigate three distinct domain definition strategies: 1) Source-based, which leverages

explicit metadata (e.g., dataset identity) as static priors; 2) K-means clustering, which discovers implicit, data-driven acoustic patterns in the latent space, where we further examine the impact of cluster granularity (K) on adaptation performance; and 3) Random assignment, serving as a control baseline to validate the necessity of meaningful domain structures. Our findings reveal that the choice of domain definition is pivotal. By identifying the optimal adversarial target among these strategies, we effectively mitigate specific acoustic biases and achieve superior generalization across diverse, unseen generative scenarios. The main contributions of this study[1] are summarized as follows:

- **Addressing Spurious Correlations:** We identify data scarcity causes overfitting to acoustic signatures, proposing a DAT framework to mitigate this without complex heuristics.
- **Systematic Investigation of Domain Definitions:** We systematically explore effective adversarial targets, ranging from explicit metadata to implicit data-driven clusters.
- **Generalizability:** We demonstrate that our findings on domain granularity are robust across various backbone models.

II. Proposed Method

We propose a robust MOS prediction framework that incorporates Domain Adversarial Training (DAT) to learn quality-aware representations invariant to domain shifts. In this section, we detail the model architecture and systematically investigate domain definition strategies.

A. Model Architecture

The overall framework is illustrated in Fig. 1. The model comprises three key components: a pre-trained SSL feature extractor, a quality prediction backbone (i.e., MultiGauss), and a domain adversarial branch.

SSL-Based Feature Extractor: Given the vast diversity of pre-trained SSL models, we specifically select the XLS-R 2B model [11] due to its exceptional model capacity and training scale. While it is primarily trained on speech, previous works [12], [13] show that speech-based SSL representations can accurately assess the quality of vinyl music collections and encode general audio events, such as environmental sounds, with high fidelity. By leveraging this broad acoustic knowledge, we utilize XLS-R as a general-purpose encoder to ensure stable quality assessment across the diverse speech, music, and audio conditions in our study.

MultiGauss Backbone (MOS Predictor): We leverage the state-of-the-art MultiGauss framework [14] as our backbone. It extracts a flattened latent representation $\mathbf{h}$ to predict a multivariate mean vector $\mathbf{m}$ (representing quality scores across multiple aspects) and a covariance matrix $\mathbf{\Lambda}$. The matrix $\mathbf{\Lambda}$ explicitly models predictive uncertainty and captures correlations between these quality dimensions. While $\mathbf{\Lambda}$ models predictive uncertainty, our analysis focuses on the latent representation $\mathbf{h}$ and the mean vector $\mathbf{m}$ to align directly with MOS-based evaluation metrics. This ensures that the domain adaptation process prioritizes the most salient features responsible for quality score prediction.

[1]Our code: https://github.com/610494/domainGRL.

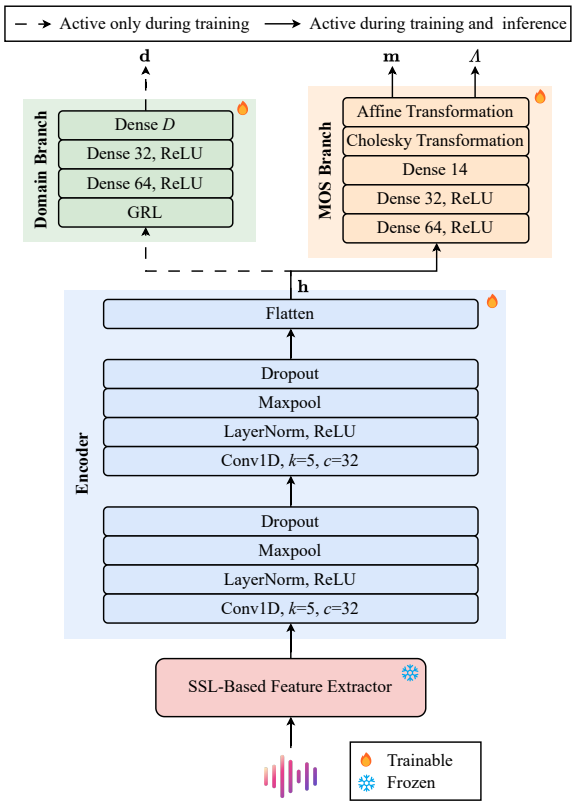


Fig. 1. The proposed model architecture with DAT.

Domain Discriminator: To disentangle domain-specific information, we introduce a parallel "Domain Branch" connected to the shared representation $\mathbf{h}$ via a Gradient Reversal Layer (GRL) [15]. This branch comprises stacked dense layers culminating in an output layer of size D (where D denotes the number of domains), mapping the features to a domain prediction vector $\mathbf{d}$. During training, the GRL reverses the gradients flowing from this discriminator to the encoder, effectively forcing $\mathbf{h}$ to become invariant to the domain distinctions.

Optimization Objective: The entire framework is trained end-to-end using a multi-task objective:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \mathcal{L}_{adv} \tag{1}$$

Following MultiGauss [14], we employ the Gaussian Negative Log-Likelihood (GNLL) loss as $\mathcal{L}_{task}$ to estimate the multivariate parameters ($\mathbf{m}$ and $\mathbf{\Lambda}$). $\mathcal{L}_{adv}$ denotes the standard cross-entropy loss for domain classification. Through the GRL, the gradients from $\mathcal{L}_{adv}$ are reversed during backpropagation, thereby enforcing the encoder to learn domain-invariant representations while minimizing the prediction error.

B. Domain Definition Strategies

The effectiveness of the GRL hinges on the definition of the adversarial target. Unlike prior works that rely on fixed domain labels, we systematically explore three strategies covering explicit, latent, and stochastic definitions:

- **DAT-Source (Dataset Origin):** This explicit strategy utilizes the inherent dataset identifiers (e.g., AudioSet vs. LibriTTS) as domain labels ($N = 6$). It aims to capture macro-level variations in production environments, such as differences in recording equipment, codec standards, and post-processing pipelines specific to each data source.
- **DAT-Kmeans (Latent Acoustic):** To capture acoustic variations that transcend dataset boundaries, we employ unsupervised K-means clustering on the pre-trained acoustic embeddings extracted from the training set. Specifically, we utilize the last-layer representations from the same frozen SSL backbone used for MOS prediction, applying temporal mean pooling to obtain global utterance-level embeddings for standard K-means clustering (using Euclidean distance). We treat the number of clusters K as a dynamic hyperparameter representing the granularity of the domain definition. We explore a range of granularities (e.g., $K \in \{2, \dots, 10\}$), selected to encompass the number of explicit source datasets ($N = 6$), to identify the optimal resolution for capturing fine-grained, implicit acoustic textures—such as reverberation patterns or background noise profiles—that are often not annotated but significantly impact the domain distribution.
- **DAT-Random (Perturbation):** This strategy assigns random labels to samples. It serves as a baseline to verify whether performance gains stem from meaningful domain disentanglement or merely from the stochastic regularization effect of the gradient reversal mechanism.

III. Experimental Setup

A. Dataset

We evaluate our proposed method on the AES-Natural dataset [16], utilizing a rigorous split protocol to benchmark generalization from natural acoustic priors to unseen generative scenarios. The Training and Validation sets consist of natural recordings stratified into three categories: Speech (EARS, LibriTTS, and Common Voice), Music (MUSDB18 and MusicCaps), and General Audio (AudioSet). Due to source availability, the final training set comprises 2,544 clips (approx. 31.6 hours), and the validation set consists of 232 clips (approx. 3.2 hours). In contrast, the Evaluation set is strictly disjoint, containing 3,060 machine-generated audio samples (approx. 7.9 hours) synthesized by various generative models. To ensure reliable ground truth, all samples were annotated by a panel of 10 expert listeners possessing professional backgrounds in audio engineering or music theory.

Departing from traditional MOS datasets that strictly evaluate technical degradation, AES-Natural characterizes audio perception along four distinct dimensions. This multi-dimensional schema allows us to disentangle low-level signal fidelity from inherent content attributes:

- **Production Quality (PQ):** Reflects the *low-level technical fidelity* of the signal. This metric focuses on physical signal degradations, such as noise floor, distortion, and bandwidth limitations, which are typically dependent on the recording equipment and channel characteristics.
- **Production Complexity (PC):** Quantifies the structural richness and density of the audio content (e.g., the number of active stems in a mix or the layering of sound effects). While strongly correlated with content type, simple models may risk forming spurious correlations between this metric and specific dataset signatures rather than the content itself.
- **Content Enjoyment (CE):** Represents the *intrinsic aesthetic appeal* and engagement value of the audio. As an intrinsic aesthetic attribute, CE abstracts away from simple signal fidelity but can be susceptible to bias arising from listener preferences for specific genres or recording styles.
- **Content Usefulness (CU):** Assesses the functional utility of the audio for its intended application (e.g., speech intelligibility or atmospheric immersion for environmental sounds).

B. Training Setup

To verify generalizability, we integrate the DAT strategy into two distinct backbone architectures. First, for **MultiGauss [14]**, we follow the original implementation, training for 30 epochs with a batch size of 64 and a learning rate of 1×10^{-4}. The best checkpoint is selected based on the lowest validation loss. Second, we evaluate **Audiobox-Aesthetics [16]**, an architecture that directly predicts quality from multi-layer WavLM features without an additional encoder. Unlike the frozen backbone in MultiGauss, we fully fine-tune this encoder for 200 epochs (batch size 16, learning rate 1×10^{-5}) to enable adversarial gradient propagation. Both models are optimized using AdamW with 0 weight decay.

The adversarial loss weight $\mathbf{\Lambda}$ acts as a hyperparameter controlling the trade-off between task performance and domain invariance. Through empirical validation, we set $\lambda = 0.5$ for the DAT-Source strategy to effectively bridge the significant distribution gaps between distinct datasets. Conversely, for the DAT-Kmeans and DAT-Random strategies, we set $\lambda = 0.1$. This reduced weight is crucial for the DAT-Kmeans strategy to prevent over-regularization since the implicit clusters may partially encode quality-related information that should not be aggressively suppressed. For the DAT-Kmeans strategy, we specifically set $K = 8$ as the default granularity, which will be further analyzed in Sec. IV-C.

C. Evaluation Metrics

To comprehensively assess our model, we report system-level Mean Squared Error (MSE) and Spearman's Rank Correlation Coefficient (SRCC). Following the evaluation protocols of prominent MOS prediction challenges [21], all metrics are calculated by first averaging the predictions and ground-truth labels for all utterances belonging to the same generative system. SRCC is prioritized as the primary metric to reflect the model's capability to reliably rank diverse generative systems. By combining it with system-level MSE, which assesses absolute precision and model calibration, we ensure a robust evaluation that disentangles ranking consistency from absolute scale errors in cross-domain scenarios.

TABLE I

PERFORMANCE COMPARISON WITH EXISTING BASELINES ACROSS FOUR ASPECTS: PRODUCTION QUALITY (PQ), PRODUCTION COMPLEXITY (PC), CONTENT ENJOYMENT (CE), AND CONTENT USEFULNESS (CU). THE SYMBOL † INDICATES RESULTS CITED FROM ORIGINAL PAPERS. BEST PERFORMANCE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**.

System / Strategy	PQ (Technical)		PC (Content)		CE (Content)		CU (Functional)	
	MSE↓	SRCC↑	MSE↓	SRCC↑	MSE↓	SRCC↑	MSE↓	SRCC↑
Existing Baselines (SOTA)								
QAMRO† [17]	-	0.883	-	0.942	-	0.869	-	0.852
DRASP† [18]	-	0.900	-	0.936	-	0.890	-	0.911
AESA-Net† [19]	0.635	0.896	**0.198**	0.928	3.991	0.904	**0.533**	0.894
MultiGauss [14]	0.557	0.942	1.093	0.947	1.841	0.938	0.945	0.961
+ L2 Regularization [20]	0.472	0.941	0.962	0.944	1.634	0.945	0.874	0.962
+ High Dropout	0.649	0.944	1.894	0.945	2.182	0.965	1.060	0.947
+ **DAT-Source** (Ours)	0.413	0.940	0.747	**0.969**	**1.581**	**0.967**	0.855	0.959
+ **DAT-Kmeans** (Ours)	0.479	**0.953**	0.928	0.945	1.605	0.952	0.835	**0.963**
+ **DAT-Random** (Ours)	**0.390**	0.941	0.945	0.958	1.689	0.961	0.789	0.959

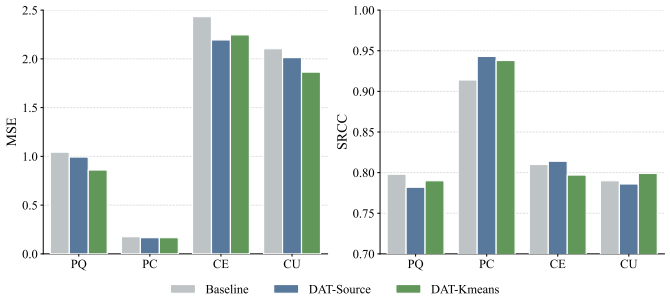


Fig. 2. Performance comparison on Audiobox-Aesthetics across MSE and SRCC. The results are reported for four aspects: PQ, PC, CE, and CU.

IV. RESULTS AND DISCUSSION

A. Main results

We evaluate the effectiveness of the proposed DAT framework on the MultiGauss backbone. Table I presents the performance comparison. The results demonstrate that explicitly disentangling domain information consistently improves model robustness. Following [22], a two-sided t-test ($p \leq 0.05$) confirms that the performance gains of our proposed DAT strategies are statistically significant compared to the baseline. Our experiments reveal that the optimal definition of a "domain" is inherently dimension-dependent, reflecting the distinct nature of different perceptual attributes.

For inherent content attributes, specifically Production Complexity (PC) and Content Enjoyment (CE), the DAT-Source strategy yields the most significant improvements. Since these attributes exhibit systematic biases—for instance, music datasets inherently yield much higher complexity scores than speech datasets—the baseline model is prone to "shortcut learning" based on dataset signatures. DAT-Source penalizes this reliance, significantly reducing PC MSE from 1.093 to 0.747 while achieving the highest SRCC of 0.969. This improvement stems from forcing the encoder to learn intrinsic structural representations rather than relying on source identity.

In contrast, technical and functional attributes, such as Production Quality (PQ) and Content Usefulness (CU), achieve optimal performance under the DAT-Kmeans strategy. Unlike content-related biases, technical degradations (e.g., background noise, reverberation) frequently transcend dataset boundaries and overlap across sources. We observe an interesting trade-off here: while explicit source labels help calibrate the absolute score scale (reducing PQ MSE to 0.413), the latent acoustic clusters discovered by K-means better capture fine-grained texture variations essential for preserving relative rankings. This is evidenced by DAT-Kmeans achieving the highest SRCC of 0.953 for PQ. Thus, for technical metrics where domain distributions overlap, unsupervised acoustic clustering offers a superior adversarial target for refining ranking capabilities.

To verify whether performance gains stem from blind regularization rather than principled domain invariance, we compared our strategies against L2 regularization, High Dropout, and DAT-Random. While L2 regularization and DAT-Random provide improvements in absolute error (MSE) for certain aspects, they consistently fail to match the superior ranking performance of our aspect-specific DAT strategies in terms of SRCC. Crucially, SRCC serves as our primary evaluation metric as it directly reflects the model's capability to reliably rank generative systems, a task where traditional and stochastic regularization prove inadequate compared to our targeted disentanglement approach. Furthermore, increasing stochasticity via High Dropout leads to statistically significant performance degradation. Notably, our framework consistently outperforms these traditional and stochastic regularization techniques across the majority of dimensions. These results confirm that targeted domain disentanglement is fundamentally superior to blind generic regularization. By explicitly addressing the specific nature of each quality dimension, our framework effectively purges spurious correlations and empowers state-of-the-art backbones to capture more intrinsic and generalized quality features.

To verify whether the effectiveness of our aspect-specific strategies generalizes across different frameworks, we evaluate the proposed method on the Audiobox-Aesthetics [16] architecture. Unlike the MultiGauss backbone, which uses frozen XLS-R features, this model utilizes fine-tuned WavLM representations, providing a distinct feature space for validation. As illustrated in Fig. 2, the performance trends are highly consistent with previous observations: the DAT-

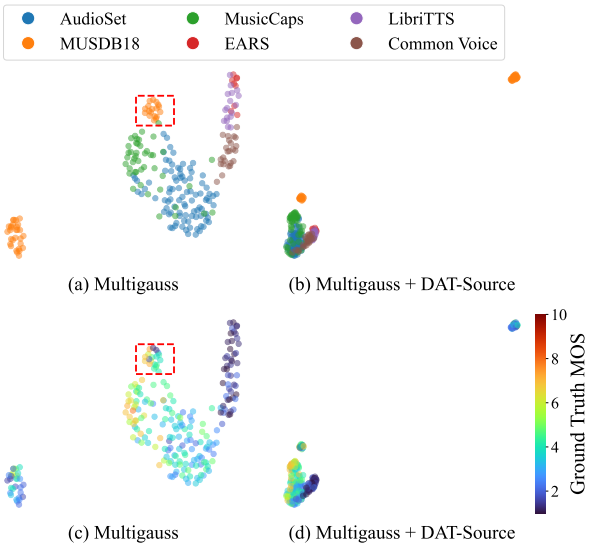


Fig. 3. Visualization of the $\mathbf{h}$ on the development set using UMAP. The top row is colored by source domain labels, and the bottom by PC scores.

Source strategy remains superior for inherent content attributes (PC, CE), while DAT-Kmeans consistently excels in technical and functional dimensions (PQ, CU). This alignment across different backbone architectures and SSL feature extractors validates the robustness of our domain definition strategies and confirms their benefits are independent of the underlying model configuration.

B. Latent Space Analysis

To analyze the manifold structure and verify whether the DAT framework effectively removes domain-specific information from the latent space, we project the bottleneck features h obtained from the model encoder into a two-dimensional space using UMAP. As illustrated in Fig. 3 (top row), the baseline model exhibits severe domain bias. Specifically, the region highlighted by the red dashed box in Fig. 3(a) shows a tight cluster formed solely by dataset identity. However, referencing Fig. 3(c), it becomes evident that samples within this domain-driven cluster possess vastly different quality scores. This clustering fragments the semantic space: high-quality samples are isolated within their respective domain "islands" rather than forming a cohesive high-quality region. This confirms that the baseline learns spurious correlations, grouping samples by domain signatures rather than their actual perceptual quality. In contrast, our proposed method successfully merges these heterogeneous domains into a unified manifold, indicating the removal of non-causal signatures. Crucially, this alignment preserves intrinsic quality information, as Fig. 3(d) reveals a continuous quality gradient transitioning from low to high quality across the aligned manifold.

To further investigate the relationship between domain invariance and quality prediction, we extend the UMAP projection to three dimensions by incorporating the predicted MOS as the vertical axis, creating a "Quality Terrain" visualization. This allows us to inspect whether the latent manifold maintains a consistent structural hierarchy across different score ranges.

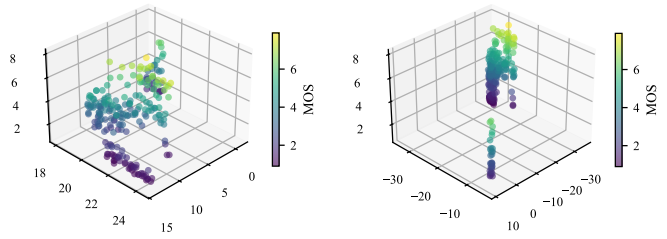


Fig. 4. 3D "Quality Terrain" generated by combining 2D UMAP projections of encoder features $\mathbf{h}$ with the predicted MOS as the z-axis.

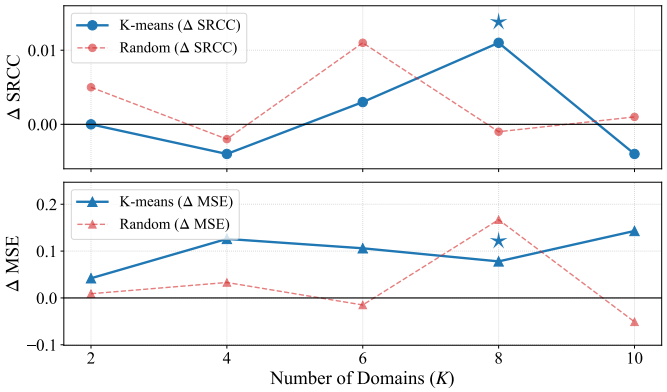


Fig. 5. Ablation study on domain granularity K for the PQ dimension. The top panel shows the absolute improvement in SRCC (Δ SRCC), with the bottom for MSE (Δ MSE) relative to the baseline. The star ($\star$) denotes the optimal configuration at $K = 8$, yielding the most balanced gains.

As shown in Fig. 4(a), the baseline model's features remain scattered into fragmented clusters across the 3D space. Even at identical quality levels, samples from different domains are horizontally segregated, forcing the model to navigate a disjointed latent space. In contrast, our DAT strategy (Fig. 4(b)) collapses these horizontal domain variances into a cohesive "Quality Pillar." In this structure, the manifold aligns vertically according to the quality gradient, where samples from all heterogeneous domains are successfully mapped onto a shared, continuous trajectory. This vertical alignment confirms that our domain-adversarial objective does not compromise the ranking capability; instead, it enforces a principled representation where domain-invariant features and quality-relevant information are effectively disentangled and organized.

Linear probing on $\mathbf{h}$ (Table II) reveals the baseline's severe dataset entanglement (90.9% Domain Acc.) artificially inflates PC SRCC (0.891) via identity shortcuts. By contrast, the DAT-Source strategy effectively purges these spurious signatures (87.5% Acc.), intentionally dismantling the shortcut to prioritize intrinsic quality features for superior zero-shot generalization (Table I). Conversely, DAT-Kmeans increases domain predictability (92.2%) by clustering acoustic textures; this effectively organizes the latent manifold, yielding the optimal predictor for technical attributes such as PQ (0.800).

C. Impact of Domain Granularity and Grouping Strategy

To investigate the impact of domain granularity K, we evaluated model performance across $K \in \{2, 4, 6, 8, 10\}$,

TABLE II
LINEAR PROBING ANALYSIS ON LATENT FEATURES $\mathbf{h}$.

Strategy	Domain Acc. (%, ↓)	PQ SRCC (↑)	PC SRCC (↑)
Baseline	90.9	0.795	**0.891**
DAT-Source	**87.5**	0.798	0.879
DAT-Kmeans	92.2	**0.800**	0.886

centered around $K = 6$ to provide a direct comparison with the DAT-Source strategy. This analysis aims to determine whether latent acoustic clusters offer more precise domain disentanglement than explicit dataset identities. As illustrated in Fig. 5, the proposed DAT-Kmeans strategy demonstrates a more structured and superior performance trend compared to the random assignment baseline. The DAT-Kmeans strategy reaches its performance peak at $K = 8$ (marked with $\star$ in Fig. 5), achieving the highest gain in ranking consistency (ΔSRCC ≈ 0.011) and a significant reduction in error (ΔMSE ≈ 0.08). While $K = 10$ yields a slightly higher MSE gain, its ΔSRCC drops below the baseline, suggesting that over-partitioning the acoustic space introduces noise that hinders the model's ranking ability.

In contrast, the random strategy exhibits high instability. Although it shows sporadic gains in MSE (e.g., at $K = 8$), its impact on SRCC is erratic and often falls into the negative range, indicating a degradation in ranking capability compared to the baseline. This disparity confirms that domain definitions must be anchored in meaningful acoustic sub-structures to provide reliable adversarial gradients. The superior trajectory of K-means validates our hypothesis that data-driven clustering effectively captures the underlying domain bias essential for robust audio quality assessment.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced Domain Adversarial Training (DAT) to the task of MOS prediction to address the critical issue of shortcut learning. Specifically, we proposed an *aspect-specific* DAT framework, demonstrating that by forcing the encoder to be invariant to domain factors, we can significantly improve the robustness and generalization of quality assessment models. Our analysis reveals that the optimal definition of a "domain" is inherently dimension-dependent: explicit source labels are superior for disentangling content-related biases, while latent acoustic clusters are more effective for refining technical quality rankings.

These strategies consistently empower state-of-the-art backbones to capture intrinsic quality features rather than spurious correlations. Future work will focus on developing a unified multi-branch architecture that simultaneously integrates both explicit source constraints and latent acoustic clustering. By leveraging these complementary domain definitions, we aim to build a robust, universal model that achieves optimal performance across all perceptual dimensions of audio quality.

REFERENCES

[1] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022," in *Proc. Interspeech*, 2022.

[2] Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, and Huaming Wang, "PAM: Prompting audio-language models for audio quality assessment," in *Proc. Interspeech*, 2024.

[3] Ryandhimas E. Zezario, Szu-Wei Fu, Fei Chen, Chiou-Shann Fuh, Hsin-Min Wang, and Yu Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2023.

[4] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proc. ICCV*, 2023.

[5] Tianchi Liu, Kong Aik Lee, Qiongqiong Wang, and Haizhou Li, "Disentangling voice and content with self-supervision for speaker recognition," in *Proc. NeurIPS*, 2023.

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[7] Yusuke Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," in *Proc. Interspeech*, 2016.

[8] Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie, "Domain adversarial training for accented speech recognition," in *Proc. ICASSP*, 2018.

[9] Mohammed Abdelwahab and Carlos Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 2423–2435, 2018.

[10] Tomohiro Tanaka, Ryo Masumura, Hiroshi Sato, Mana Ihori, Kohei Matsuura, Takanori Ashihara, and Takafumi Moriya, "Domain adversarial self-supervised speech representation learning for improving unknown domain downstream tasks," in *Proc. Interspeech*, 2022.

[11] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022.

[12] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines, "Audio quality assessment of vinyl music collections using self-supervised learning," in *Proc. ICASSP*, 2023.

[13] Tung-Yu Wu, Tsu-Yuan Hsu, Chen-An Li, Tzu-Han Lin, and Hung-yi Lee, "The efficacy of self-supervised speech models for audio representations," in *Proc. HEAR*, 2022.

[14] Fredrik Cumlin, Xinyu Liang, Victor Ungureanu, Chandan K. A. Reddy, Christian Schüldt, and Saikat Chatterjee, "Multivariate probabilistic assessment of speech quality," in *Proc. Interspeech*, 2025.

[15] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. ICML*, 2015.

[16] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," in *Arxiv preprint arXiv:2502.05139*, 2025.

[17] Chien-Chun Wang, Kuan-Tang Huang, Cheng-Yeh Yang, Hung-Shin Lee, Hsin-Min Wang, and Berlin Chen, "QAMRO: Quality-aware adaptive margin ranking optimization for human-aligned assessment of audio generation systems," in *Proc. ASRU*, 2025.

[18] Cheng-Yeh Yang, Kuan-Tang Huang, Chien-Chun Wang, Hung-Shin Lee, Hsin-Min Wang, and Berlin Chen, "DRASP: A dual-resolution attentive statistics pooling framework for automatic MOS prediction," in *Proc. APSIPA ASC*, 2025.

[19] Dyah A. M. G. Wisnu, Ryandhimas E. Zezario, Stefano Rini, Hsin-Min Wang, and Yu Tsao, "Improving perceptual audio aesthetic assessment via triplet loss and self-supervised embeddings," in *Proc. ASRU*, 2025.

[20] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[21] Wen-Chin Huang, Hui Wang, Cheng Liu, Yi-Chiao Wu, Andros Tjandra, Wei-Ning Hsu, Erica Cooper, Yong Qin, and Tomoki Toda, "The audioMOS challenge 2025," in *Proc. ASRU*, 2025.

[22] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. ICASSP*, 2022.