GT-PCQA: Geometry-Texture Decoupled Point Cloud Quality Assessment with MLLM

Guohua Zhang[1] Jian Jin[2] Meiqin Liu[1*] Chao Yao[3] Weisi Lin[2] Yao Zhao[1]
[1]Beijing Jiaotong University [2]Nanyang Technological University
[3]University of Science and Technology Beijing
24125207@bjtu.edu.cn, jian.jin@ntu.edu.sg, mqliu@bjtu.edu.cn,
yaochao@ustb.edu.cn, wslin@ntu.edu.sg, yzhao@bjtu.edu.cn

***Abstract*—With the rapid advancement of Multi-modal Large Language Models (MLLMs), MLLM-based Image Quality Assessment (IQA) methods have shown promising generalization. However, directly extending these MLLM-based IQA methods to PCQA remains challenging. On the one hand, existing PCQA datasets are limited in scale, which hinders stable and effective instruction tuning of MLLMs. On the other hand, due to large-scale image-text pretraining, MLLMs tend to rely on texture-dominant reasoning and are insufficiently sensitive to geometric structural degradations that are critical for PCQA. To address these gaps, we propose a novel MLLM-based no-reference PCQA framework, termed GT-PCQA, which is built upon two key strategies. First, to enable stable and effective instruction tuning under scarce PCQA supervision, a 2D–3D joint training strategy is proposed. This strategy formulates PCQA as a relative quality comparison problem to unify large-scale IQA datasets with limited PCQA datasets. It incorporates a parameter-efficient Low-Rank Adaptation (LoRA) scheme to support instruction tuning. Second, a geometry-texture decoupling strategy is presented, which integrates a dual-prompt mechanism with an alternating optimization scheme to mitigate the inherent texture-dominant bias of pre-trained MLLMs, while enhancing sensitivity to geometric structural degradations. Extensive experiments demonstrate that GT-PCQA achieves competitive performance and exhibits strong generalization.**

***Index Terms*—point cloud quality assessment, multi-modal large language models, geometry–texture decoupling**

I. Introduction

Point clouds, as 3D representations of objects or scenes, are widely used in applications such as Virtual Reality (VR) [1], Augmented Reality (AR) [2], 3D modeling [3], and autonomous driving [4]. However, point clouds are inevitably degraded during acquisition, processing, storage, and transmission [5]–[9]. Therefore, accurate Point Cloud Quality Assessment (PCQA) metrics are crucial.

PCQA metrics are commonly classified into Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) methods. Early efforts primarily focused on FR metrics, which calculate distortions based on complete reference information, ranging from geometric distances like MSE-p2pl [10] and HD-p2pl [10] to color-based PSNR-yuv [11] and structural similarity-based descriptors such as GraphSIM [12] and PointSSIM [13]. Since obtaining the original reference point cloud is often difficult in real-world scenarios, this also makes NR-PCQA more critical and challenging.

Recently, NR-PCQA has experienced significant improvement through the use of advanced Deep Neural Networks (DNNs) [5]. These approaches have evolved from traditional hand-crafted statistical models such as 3D-NSS [14] to more sophisticated deep architectures, including ResSCNN [15], PQA-net [16], and the MM-PCQA [17], which integrates multi-modal features. Despite their promising performance on individual benchmarks, these methods are typically optimized on specific datasets, causing the learned quality representations to be closely tied to dataset-specific distortion characteristics. As a result, their performance often degrades under distribution shifts between training and testing data, leading to limited cross-dataset generalization.

To improve the generalization of NR-PCQA, various advanced training techniques have been explored, like the domain adaptation [18]–[21]. For the traditional image quality assessment, Multi-modal Large Language Models (MLLMs) have demonstrated promising cross-dataset generalization for Image Quality Assessment (IQA) tasks. Among MLLM-based IQA methods [22], [23], Compare2Score [24], which formulates IQA as a relative quality comparison, enables MLLMs to leverage supervision from multiple datasets, providing a scalable training paradigm for robust quality assessment.

However, directly extending these MLLM-based IQA methods to PCQA remains challenging. On the one hand, existing PCQA datasets are limited in scale, which hinders stable and effective instruction tuning of MLLMs. On the other hand, due to large-scale image-text pretraining, MLLMs tend to rely on texture-dominant reasoning and are insufficiently sensitive to geometric structural degradations that are critical for PCQA.

To address these challenges, we propose an MLLM-based GT-PCQA that includes two key strategies. First, to enable stable and effective instruction tuning of MLLMs under scarce PCQA supervision, a 2D–3D joint training strategy is proposed. Specifically, PCQA is formulated as a relative quality comparison problem, which not only serves as a unified bridge for jointly leveraging large-scale IQA datasets

This work is supported by the National Natural Science Foundation of China (62372036, 62120106009, U22A2022, 62332017), and the Ministry of Education, Singapore (Tier 1 RG103/24).

Corresponding author: mqliu@bjtu.edu.cn.

and limited PCQA datasets, but also naturally aligns with instruction–response learning. This comparative formulation facilitates the construction of scalable and reusable supervision signals, allowing the model to learn from both image- and point cloud-based quality annotations in a unified framework. To further support stable and efficient instruction tuning under limited PCQA supervision, we integrate a parameter-efficient Low-Rank Adaptation (LoRA) [25] scheme. It constrains task-specific updates to a low-rank subspace, preserving the general visual–language capabilities of the pre-trained MLLM while enabling the effective learning of distortion-aware representations. Second, to mitigate the inherent texture-dominant bias of pre-trained MLLMs, while enhancing sensitivity to geometric structural degradations, a geometry–texture decoupling strategy is presented. This strategy employs a dual-prompt mechanism combined with an alternating optimization scheme, explicitly separating the learning of geometry-aware and texture-aware representations. Specifically, the dual-prompt mechanism applies geometry-aware and texture-aware prompts to multi-view point clouds and images, respectively, guiding the model to attend to geometric structural degradations without suppressing texture cues, thereby preserving their texture awareness and enhancing sensitivity to geometric structural degradations. Furthermore, under this alternating optimization scheme, geometry-focused and texture-focused optimization steps are strictly alternated during fine-tuning, thereby mitigating the inherent texture-dominant bias of pre-trained MLLMs.

The contributions can be summarized as follows:

- We propose the GT-PCQA, which formulates PCQA as a unified comparative evaluation task, effectively unifying image- and point cloud-based quality assessment under a relative comparison paradigm.
- We propose a 2D-3D joint training strategy that leverages relative quality comparison to integrate large-scale IQA datasets with limited PCQA datasets and uses a parameter-efficient LoRA scheme, enabling stable and effective instruction tuning.
- We propose a geometry-texture decoupling strategy, consisting of a dual-prompt mechanism and an alternating optimization scheme, which mitigates the inherent texture-dominant bias of pre-trained MLLMs, while enhancing sensitivity to geometric structural degradations.

II. Proposed Method

An overview of the proposed GT-PCQA framework is illustrated in Fig. 1, and the details are introduced in the following subsections.

A. 2D–3D Joint Training Strategy

To enable stable and effective instruction tuning of Multimodal Large Language Models (MLLMs) under scarce PCQA supervision, a 2D–3D joint training strategy is proposed that formulates PCQA as a relative quality comparison problem, serving as a unified bridge to leverage large-scale IQA datasets and limited PCQA datasets jointly. This comparative formulation naturally aligns with instruction–response learning and enables the construction of scalable and reusable supervision signals for instruction tuning.

Since subjective evaluation protocols vary across datasets and lead to inconsistent perceptual scales, we adopt a pairwise quality comparison mechanism for images and multi-view point clouds. This design enables the construction of large-scale and reusable instruction–response pairs. Following empirical criteria [26], quality differences are discretized into five levels: inferior, worse, similar, better, and superior. Following common practice in subjective quality modeling [24], we assume that the rating uncertainties of two samples are independent, which allows the quality difference between them to be normalized by their aggregated uncertainty.

Formally, we randomly sample n_k image or multi-view point clouds pairs $\{(x_k^{(i)}, x_k^{(j)})\}_{i,j=1}^{n_k}$ from each dataset. For each sampled pair $\left(x^{(i)}, x^{(j)}\right)$, the corresponding normalized quality level L_{ij}, which represents the relative quality of sample j with respect to sample i, is formulated as:

$$L_{ij} = \begin{cases} \text{inferior}, & \text{if } z_{ij} > 2 \\ \text{worse}, & \text{if } 1 < z_{ij} \leq 2 \\ \text{similar}, & \text{if } -1 < z_{ij} \leq 1 \\ \text{better}, & \text{if } -2 < z_{ij} \leq -1 \\ \text{superior}, & \text{if } z_{ij} < -2 \end{cases} \tag{1}$$

where z_{ij} denotes the standardized quality difference between samples i and j, formulated as:

$$z_{ij} = \frac{q^{(i)} - q^{(j)}}{\sqrt{(\sigma^{(i)})^2 + (\sigma^{(j)})^2}}, \tag{2}$$

where $q^{(i)}$ and $q^{(j)}$ represent the Mean Opinion Score (MOS) of samples i and j, respectively, while $\sigma^{(i)}$ and $\sigma^{(j)}$ denote their corresponding rating standard deviations.

Specifically, the visual encoder and visual abstractor are fully fine-tuned to capture distortion-sensitive representations from both images and multi-view point clouds, enabling the model to perceive fine-grained geometric artifacts and bridge the substantial domain gap between natural images and point cloud projections. In contrast, the large language model is adapted via lightweight LoRA modules inserted into the query and value projection matrices, while all other parameters remain frozen.

From a mechanism perspective, quality assessment within MLLMs is primarily governed by attention-based comparison and evidence aggregation [27]. Updating the query projections modulates how visual and linguistic tokens attend to distortion-related cues, while adapting the value projections controls how quality-relevant evidence is propagated through attention layers. Therefore, restricting adaptation to the query–value subspace enables effective recalibration of quality-aware reasoning while preserving the general semantic and syntactic representations of the language model, which is consistent with the design principle of LoRA [25].

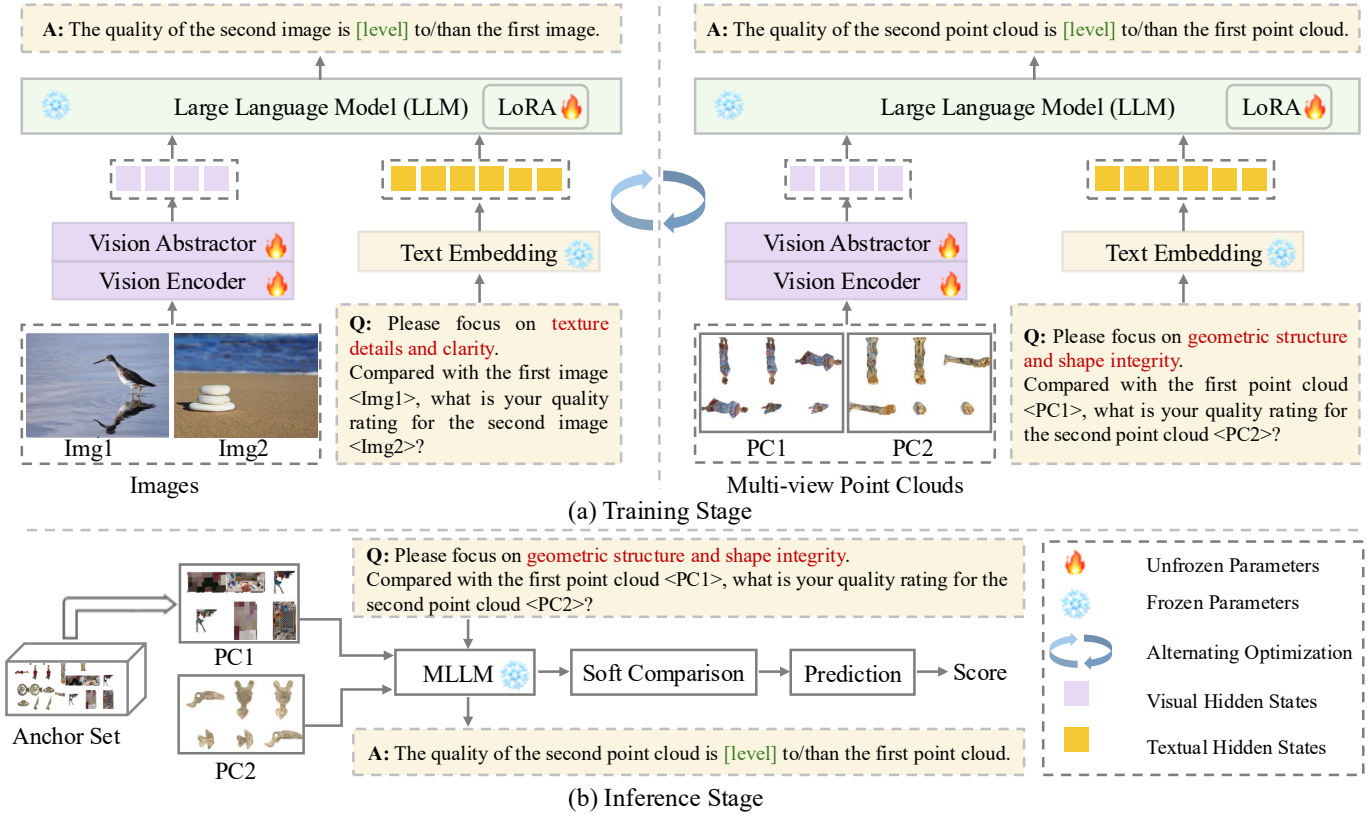

Fig. 1. Architecture of the proposed GT-PCQA. (a) During the training stage, the model alternates between image pairs and multi-view point cloud pairs. Visual inputs are encoded by the vision encoder and abstracted into compact representations, while attribute-specific text prompts (*e.g.,* texture-aware or geometry-aware) are embedded into textual representations. The aligned multimodal features are then fed into a LoRA-adapted LLM to perform relative quality comparison, enabling stable and effective instruction tuning under heterogeneous 2D–3D supervision. (b) During the inference stage, the trained MLLM is fully frozen, and each test point cloud is evaluated against an anchor set via soft comparison, ultimately predicting the final quality score.

Overall, the proposed 2D–3D joint training strategy reformulates PCQA as a relative quality comparison task, enabling large-scale IQA data to be reused as unified and reusable instruction-level supervision. Together with parameter-efficient LoRA-based adaptation, it enables stable and effective instruction tuning under scarce PCQA supervision.

B. Geometry–Texture Decoupling Strategy

To mitigate the inherent texture-dominant bias of pre-trained MLLMs, while enhancing sensitivity to geometric structural degradations, a geometry–texture decoupling strategy is presented that integrates a dual-prompt mechanism with an alternating optimization scheme to decouple geometry-aware and texture-aware learning explicitly.

1) Dual-Prompt Mechanism: The dual-prompt mechanism applies geometry-aware and texture-aware prompts to multi-view point clouds and images, respectively, guiding the model to attend to geometric structures without suppressing texture cues, thereby preserving their texture awareness and enhancing sensitivity to geometric structural degradations.

Specifically, we use two prompts: a texture-aware prompt for images and a geometry-aware prompt for multi-view point clouds. The former guides the model to focus on fine-grained surface details and color, while the latter emphasizes spatial structures and shapes. During training, visual features are concatenated with the corresponding prompt and fed into the LLM, steering it to focus on quality-relevant cues.

The instruction-response formats for image and point cloud pairs are defined as follows:

> **Texture prompt:** Please focus on texture details and clarity. Compared with the first image <Img1>, what is your quality rating for the second image <Img2>?
> **Response:** The quality of the second image is [level] to/than the first image.
>
> **Geometry prompt:** Please focus on geometric structure and shape integrity. Compared with the first point cloud <PC1>, what is your quality rating for the second point cloud <PC2>?
> **Response:** The quality of the second point cloud is [level] to/than the first point cloud.

2) Alternating Optimization Scheme: To mitigate the inherent texture-dominant bias of pre-trained MLLMs, an alternating optimization scheme is proposed.

Specifically, the optimization process alternates between geometry-aware and texture-aware updates during fine-tuning. At each optimization step t, the model is guided by either geometry-oriented or texture-oriented supervision, ensuring that geometric cues are periodically reinforced rather than

being suppressed by the stronger texture bias inherited from pre-trained MLLMs.

The training loss $\mathcal{L}_{ce}^{(t)}$ at step t is formulated as:

$$\mathcal{L}_{ce}^{(t)} = \begin{cases} \mathbb{E}_{x\sim\mathcal{D}_T}\big[\mathrm{CE}(x)\big], & \text{if } t \text{ is even} \\ \mathbb{E}_{x\sim\mathcal{D}_G}\big[\mathrm{CE}(x)\big], & \text{if } t \text{ is odd} \end{cases}, \tag{3}$$

where $\mathcal{D}_T$ and $\mathcal{D}_G$ denote the texture-oriented and geometry-oriented data distributions, respectively. The cross-entropy loss $\mathrm{CE}(x)$ is formulated as:

$$\mathrm{CE}(x) = -\sum_{c=1}^{C} y_c \log p_c(x), \tag{4}$$

where $p_c(x)$ is the predicted probability of the c-th relative quality level, C is the total number of levels, and y_c is the corresponding ground-truth level label.

By separating geometry- and texture-driven updates, our mechanism prevents bias toward texture cues and ensures stable learning of geometry-aware quality features, which is crucial for robust PCQA.

After training, the MLLM is frozen and used in an inference stage, where the learned multimodal feature embeddings are utilized within an MLLM-based soft comparison [24] to evaluate each test point cloud by comparing it against an anchor set for quality prediction. Following this framework, an anchor set $\mathcal{A}$ is constructed from the SJTU-PCQA dataset to ensure reliable pairwise comparisons. Specifically, the dataset is partitioned into β quality intervals, and one anchor object is selected from each interval based on rating consistency. The anchor set $\mathcal{A}$ is formulated as:

$$\mathcal{A} = \bigcup_{k=1}^{\beta} a_k, \tag{5}$$

where a_k denotes the anchor object corresponding to the k-th quality interval, which is selected by minimizing the variance of subjective scores, formulated as:

$$a_k = \arg\min_{x\in\mathcal{D}_k} \sigma^2(x), \tag{6}$$

where $\mathcal{D}_k$ represents the k-th quality interval of the SJTU-PCQA dataset, and $\sigma^2(x)$ denotes the variance of MOS scores for multi-view point cloud x.

Then, following the soft comparison protocol, the pairwise comparison outcomes are aggregated into a probability matrix P, which captures the relative quality ordering between the test sample and the anchor set. Based on this matrix, the quality prediction score $\hat{q}$ of the test sample is inferred by solving a posterior-driven optimization problem:

$$\hat{q} = \arg\max_{q} \ \mathcal{F}(q; P), \tag{7}$$

where q denotes the latent continuous quality variable, and $\mathcal{F}(q; P)$ denotes an objective function that encourages consistency between the estimated quality score and the observed soft comparison relations encoded in P, following common practices in probabilistic ranking and quality assessment [28].

III. Experiments

A. Experimental Setup

1) Dataset and Evaluation Metrics: Our GT-PCQA is trained under a joint 2D–3D setting using seven datasets in total, including the SJTU-PCQA [29] dataset for PCQA supervision and six large-scale Image Quality Assessment (IQA) datasets (LIVE [30], KADID-10k [31], CLIVE [32], KonIQ-10k [33], BID [34], and CSIQ [35]) to support stable and effective instruction tuning with LoRA. **Since SJTU-PCQA is the only dataset providing standard deviation annotations required for our training objective**, it is used for both training and in-dataset evaluation. To assess generalization, the trained model is directly evaluated on the unseen PCQA dataset WPC [36] without any fine-tuning. We adopt Pearson's Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank Order Correlation Coefficient (KROCC), and Root Mean Square Error (RMSE) as metrics. Better performance is indicated by higher PLCC, SROCC, and KROCC, and lower RMSE.

2) Implementation Details: Our GT-PCQA is built upon the mPLUG-Owl2 framework [37]. equipped with a CLIP-ViT-L vision encoder [38], a six-layer Q-Former visual abstractor, and an LLaMA-2-7B [39] as the LLM. We train the model on 210k images and multi-view point clouds pairs using the cross-entropy over predicted logits, with a batch size of 64 for 3 epochs. We adopt AdamW [40] as the optimizer. The initial learning rate is set to 2e-5 and decays gradually using the cosine decay strategy. To enable parameter-efficient adaptation, LoRA adapters are inserted into the attention projection layers of the LLM. The rank is set to $r = 128$, which is selected based on ablation studies, while the scaling factor is empirically set to $\alpha = 256$ to match the chosen rank. All other training settings are kept consistent across different datasets. Training is conducted on three RTX 3090 GPUs, and each epoch takes approximately 5.1 hours to complete, while inference can be performed on a single RTX 3090 GPU. Furthermore, to obtain the anchor set, we divide the training set of the SJTU-PCQA [29] into five ($\beta = 5$) quality intervals based on their MOSs and Standard deviations, from which we select one representative anchor multi-view point cloud.

B. Performance Evaluation

As shown in Tab. I, we compare GT-PCQA with representative full-reference (FR) and no-reference (NR) PCQA methods on the SJTU-PCQA dataset. Among NR approaches, GT-PCQA achieves competitive performance, with an SROCC of 0.8953 and a PLCC of 0.8883, substantially outperforming several established deep learning baselines such as Compare2Score [24] (by 29.1% in SROCC) and IT-PCQA [18] (by 40.7% in SROCC). In addition, GT-PCQA significantly reduces the RMSE to 0.8629, corresponding to a 53.2% error reduction compared to Compare2Score [24].

It is worth noting that MM-PCQA [17] achieves higher in-domain accuracy on SJTU-PCQA, largely due to its fully supervised training paradigm and dataset-specific optimization, which can lead to overfitting on the training distribution.

In contrast, GT-PCQA emphasizes cross-dataset generalization by leveraging relative quality comparison and reusable instruction-level supervision, along with parameter-efficient instruction tuning. As shown in the cross-dataset evaluation (Sec. III-D), GT-PCQA demonstrates substantially better generalization when applied to unseen PCQA datasets, reflecting a more favorable trade-off between in-domain performance and robustness across diverse point cloud distributions.

TABLE I
PERFORMANCE COMPARISON ON SJTU-PCQA DATASET. THE BEST AND SECOND-BEST NR-PCQA RESULTS ARE HIGHLIGHTED IN **BOLD** AND *italic*, RESPECTIVELY.

Type	Methods	SROCC↑	PLCC↑	KROCC↑	RMSE↓
FR	MSE-p2pl [10]	0.6277	0.5940	0.4825	2.2815
	HD-p2pl [10]	0.6441	0.6874	0.4565	2.1255
	PSNR-yuv [11]	0.7950	0.8170	0.6196	1.3151
	GraphSIM [12]	0.8783	0.8449	0.6947	1.0321
	PointSSIM [13]	0.6867	0.7136	0.4964	1.7001
NR	ResSCNN [15]	0.8600	0.8100	-	-
	PQA-net [16]	0.8372	0.8586	0.6304	1.0719
	3D-NSS [14]	0.7144	0.7382	0.5174	1.7686
	IT-PCQA [18]	0.6361	0.6934	0.4932	1.6240
	MM-PCQA [17]	**0.9103**	**0.9226**	**0.7838**	**0.7716**
	Compare2Score [24]	0.6933	0.7926	0.5516	1.8436
	GT-PCQA (Ours)	*0.8953*	*0.8883*	*0.7382*	*0.8629*

C. Ablation Study

We conduct comprehensive ablation studies to evaluate the contributions of key components in GT-PCQA and the impact of the LoRA rank. Component-level results for the Dual-Prompt (DP), Alternating Optimization (AO), and LoRA are reported in Tab. II, while the effect of different LoRA ranks is presented in Tab. III.

Impact of the Dual-Prompting mechanism (DP). Removing DP leads to a significant drop in SROCC and PLCC, indicating that DP helps the model balance geometric and texture learning, thereby preserving texture awareness while enhancing sensitivity to geometric structural degradations. As geometric structural degradations are a key factor in PCQA, DP plays a crucial role in quality discrimination.

Impact of the Alternating Optimization scheme (AO). Removing AO caused slight but consistent drops across all metrics, indicating that the alternating optimization mechanism is crucial for training stability. It prevents over-reliance on texture cues and consistently enhances sensitivity to geometric structures across objectives, improving overall generalization.

Impact of the LoRA. Removing LoRA results in the largest performance degradation, with consistently lower PLCC and KROCC and higher RMSE. This demonstrates that directly fine-tuning the language model under limited PCQA supervision leads to unstable instruction optimization. In contrast, the lightweight LoRA adaptation enables stable and effective instruction tuning of the MLLM by constraining language-side updates, while allowing the visual branch to be fully optimized for learning distortion-sensitive representations, thereby improving overall prediction accuracy.

TABLE II
ABLATION STUDIES ON DP, AO, AND LORA, WHERE "W/O" DENOTES REMOVING THE CORRESPONDING COMPONENT; IN PARTICULAR, "W/O LORA" INDICATES FULL FINE-TUNING OF ALL LLM PARAMETERS.

Model	SROCC↑	PLCC↑	KROCC↑	RMSE↓
w/o DP	0.8045	0.8321	0.6887	1.2415
w/o LoRA	0.8031	0.7845	0.5818	1.7629
w/o AO	0.8345	0.8451	0.6921	1.2314
GT-PCQA(Ours)	**0.8953**	**0.8883**	**0.7382**	**0.8629**

Impact of LoRA Rank. We further analyze the effect of the LoRA rank on performance. As shown in Tab. III, the results reveal a clear trade-off between model capacity and generalization. Increasing the rank from $r = 32$ to $r = 128$ consistently improves performance, suggesting that a moderate rank is sufficient to support the dual-prompt mechanism and learn decoupled geometry–texture representations. However, further increasing it to $r = 256$ degrades performance, likely due to overfitting under limited PCQA supervision. Consequently, $r = 128$ achieves the best balance between capacity and generalization and is used in all experiments.

TABLE III
ABLATION STUDY ON LORA RANK r USING THE SJTU-PCQA DATASET. BOLD INDICATES THE BEST PERFORMANCE.

Rank (r)	SROCC↑	PLCC↑	KROCC↑	RMSE↓
32	0.8375	0.8411	0.6515	1.9118
64	0.8562	0.8556	0.6840	1.5394
128	**0.8953**	**0.8883**	**0.7382**	**0.8629**
256	0.8601	0.8241	0.6638	1.6082

TABLE IV
CROSS-DATASET EVALUATION FROM SJTU-PCQA TO WPC. GT-PCQA IS TRAINED ON SJTU-PCQA, AND NO FINE-TUNING IS PERFORMED ON WPC. BOLD INDICATES THE BEST PERFORMANCE.

Methods	SROCC↑	PLCC↑	KROCC↑	RMSE↓
MM-PCQA [17]	0.4988	0.4354	0.2929	2.2103
Compare2Score [24]	0.5157	0.4589	0.3829	2.0103
GT-PCQA(Ours)	**0.6708**	**0.6339**	**0.4579**	**1.3672**

D. Cross-Dataset Validation

Cross-dataset generalization of different methods. Models are trained on SJTU-PCQA, with auxiliary IQA datasets used during training to support stable instruction tuning with LoRA. Evaluation is performed directly on the unseen WPC dataset without any fine-tuning. The performance of competitive methods, including MM-PCQA [17] and Compare2Score [24], is reported for comparison. As shown in Tab. IV, several observations can be made: i) GT-PCQA achieves the best performance across all metrics, demonstrating strong cross-dataset generalization. ii) Despite substantial differences in distortion types, point cloud density, and content characteristics between SJTU-PCQA and WPC, GT-PCQA maintains robust performance.

This indicates that the model captures intrinsic, distortion-aware geometric quality cues, and highlights the advantage of our geometry- and texture-aware dual-prompt mechanism combined with the proposed training strategy.

IV. Conclusion

In summary, we propose GT-PCQA, a novel MLLM-based NR-PCQA framework. To achieve stable and effective instruction tuning with limited PCQA supervision, we introduce a 2D–3D joint training strategy that formulates PCQA as a relative quality comparison task, unifies large-scale IQA data with scarce PCQA data, and employs a parameter-efficient LoRA scheme. To mitigate the texture-dominant bias of pre-trained MLLMs while preserving texture sensitivity, we further design a geometry–texture decoupling strategy with dual prompts and alternating optimization for explicit geometry- and texture-aware learning. Extensive experiments on multiple PCQA datasets demonstrate competitive performance and strong cross-dataset generalization, enhancing the perceptual capability of MLLMs for PCQA.

References

[1] Alex H Lang, Sourabh Vora, Holger Caesar, et al., "Pointpillars: Fast encoders for object detection from point clouds," in *CVPR*, 2019, pp. 12697–12705.

[2] Meiqin Liu, Chenming Xu, Chao Yao, et al., "JNMR: Joint non-linear motion regression for video frame interpolation," *IEEE Transactions on Image Processing*, vol. 32, pp. 5283–5295, 2023.

[3] Rufael Mekuria, Kees Blom, and Pablo Cesar, "Design, implementation, and evaluation of a point cloud codec for tele-immersive video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 828–842, 2016.

[4] Chunyi Li, Yuan Tian, Xiaoyue Ling, et al., "Image quality assessment: From human to machine preference," in *CVPR*, 2025, pp. 7570–7581.

[5] Wei Gao, Shangkun Sun, Huiming Zheng, et al., "Opendmc: an open-source library and performance evaluation for deep-learning-based multi-frame compression," in *ACM MM*, 2023, pp. 9685–9688.

[6] Lili Meng, Jian Jin, Yingnan Wang, et al., "Boosting the no-reference image quality assessment via low-quality pseudo references," *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[7] Renyu Yang, Jian Jin, Lili Meng, et al., "Scaling audio-visual quality assessment dataset via crowdsourcing," *arXiv preprint arXiv:2602.22659*, 2026.

[8] Tianang Chen, Jian Jin, Shilv Cai, et al., "MUGSQA: Novel multi-uncertainty-based gaussian splatting quality assessment method, dataset, and benchmarks," *arXiv preprint arXiv:2511.06830*, 2025.

[9] Zhuangzi Li, Jian Jin, Shilv Cai, and Weisi Lin, "R4-CGQA: Retrieval-based vision language models for computer graphics image quality assessment," *arXiv preprint arXiv:2603.10578*, 2026.

[10] Dong Tian, Hideaki Ochimizu, Chen Feng, et al., "Geometric distortion metrics for point cloud compression," *In ICIP*, pp. 3460–3464, 2017.

[11] Eric M Torlig, Evangelos Alexiou, Tiago A Fonseca, et al., "A novel methodology for quality assessment of voxelized point clouds," in *Applications of Digital Image Processing XLI*. SPIE, 2018, vol. 10752, pp. 174–190.

[12] Qi Yang, Zhan Ma, Yiling Xu, et al., "Inferring point cloud quality via graph similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3015–3029, 2020.

[13] Evangelos Alexiou and Touradj Ebrahimi, "Towards a point cloud structural similarity metric," in *ICMEW*. IEEE, 2020, pp. 1–6.

[14] Zicheng Zhang, Wei Sun, Xiongkuo Min, et al., "No-reference quality assessment for 3D colored point cloud and mesh models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7618–7631, 2022.

[15] Yipeng Liu, Qi Yang, Yiling Xu, et al., "Point cloud quality assessment: Dataset construction and learning-based no-reference metric," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2s, pp. 1–26, 2023.

[16] Qi Liu, Hui Yuan, Honglei Su, et al., "PQA-Net: Deep no reference point cloud quality assessment via multi-view projection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4645–4660, 2021.

[17] Zicheng Zhang, Wei Sun, et al., "MM-PCQA: Multi-modal learning for no-reference point cloud quality assessment," *in IJCAI*, 2023.

[18] Qi Yang, Yipeng Liu, Siheng Chen, et al., "No-reference point cloud quality assessment via domain adaptation," *In CVPR*, pp. 21179–21188, 2022.

[19] Yiting Lu, Xin Li, Jianzhao Liu, et al., "StyleAM: Perception-oriented unsupervised domain adaptation for no-reference image quality assessment," *IEEE Transactions on Multimedia*, 2024.

[20] Nu Sun, Jian Jin, Lili Meng, Weisi Lin, et al., "MFCQA: Multi-range feature cross-attention mechanism for no-reference image quality assessment," *Knowledge-Based Systems*, vol. 310, pp. 113027, 2025.

[21] Guohua Zhang, Jian Jin, Meiqin Liu, et al., "QD-PCQA: Quality-aware domain adaptation for point cloud quality assessment," *arXiv preprint arXiv:2603.03726*, 2026.

[22] Chunyi Li, Jiaohao Xiao, Jianbo Zhang, et al., "Perceptual quality assessment for embodied AI," *arXiv e-prints*, pp. arXiv–2505, 2025.

[23] Haoning Wu, Zicheng Zhang, Weixia Zhang, et al., "Q-align: Teaching LMMs for visual scoring via discrete text-defined levels," *arXiv preprint arXiv:2312.17090*, 2023.

[24] Hanwei Zhu, Haoning Wu, Yixuan Li, et al., "Adaptive image quality assessment via teaching large multimodal model to compare," in *NeurIPS*, 2024, vol. 37, pp. 32611–32629.

[25] Edward J Hu, Yelong Shen, Phillip Wallis, et al., "LoRA: Low-rank adaptation of large language models," *in ICLR*, vol. 1, no. 2, pp. 3, 2022.

[26] Handbook Of Parametric, "Handbook of parametric and nonparametric statistical procedures," .

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention is all you need," *in NeurIPS*, 2017.

[28] Kristi Tsukida and Maya R Gupta, "How to analyze paired comparison data," Tech. Rep., 2011.

[29] Qi Yang, Hao Chen, Zhan Ma, et al., "Predicting the perceptual quality of point cloud: A 3D-to-2D projection-based exploration," *IEEE Transactions on Multimedia*, vol. 23, pp. 3877–3891, 2020.

[30] Hamid R Sheikh, Muhammad F Sabir, Alan C Bovik, et al., "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[31] Hanhe Lin, Vlad Hosu, and Dietmar Saupe, "KADID-10k: A large-scale artificially distorted iqa database," in *QoMEX*, 2019, pp. 1–3.

[32] Deepti Ghadiyaram and Alan C Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015.

[33] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, et al., "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[34] Alexandre Ciancio, Eduardo AB Da Silva, Amir Said, et al., "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 64–75, 2010.

[35] Eric C Larson and Damon M Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of electronic imaging*, vol. 19, no. 1, pp. 011006–011006, 2010.

[36] Honglei Su, Zhengfang Duanmu, Wentao Liu, et al., "Perceptual quality assessment of 3D point clouds," *in ICIP*, pp. 3182–3186, 2019.

[37] Qinghao Ye, Haiyang Xu, Jiabo Ye, et al., "mPLUG-Owl2: Revolutionising multi-modal large language model with modality collaboration," *in CVPR*, pp. 13040–13051, 2024.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[39] Hugo Touvron, Louis Martin, Kevin Stone, et al., "Llama 2: open foundation and fine-tuned chat models. arxiv," *arXiv preprint arXiv:2307.09288*, vol. 10, 2023.

[40] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.