PIT-QMM: A LARGE MULTIMODAL MODEL FOR NO-REFERENCE POINT CLOUD QUALITY ASSESSMENT

Shashank Gupta[⋆†]   *Gregoire Phillips*[†]   *Alan C Bovik*[⋆]

⋆ The University of Texas at Austin
†Ericsson Research

ABSTRACT

Large Multimodal Models (LMMs) have recently enabled considerable advances in the realm of image and video quality assessment, but this progress has yet to be fully explored in the domain of 3D assets. We are interested in using these models to conduct No-Reference Point Cloud Quality Assessment (NR-PCQA), where the aim is to automatically evaluate the perceptual quality of a point cloud in absence of a reference. We begin with the observation that different modalities of data – text descriptions, 2D projections, and 3D point cloud views – provide complementary information about point cloud quality. We then construct PIT-QMM, a novel LMM for NR-PCQA that is capable of consuming text, images and point clouds end-to-end to predict quality scores. Extensive experimentation shows that our proposed method outperforms the state-of-the-art by significant margins on popular benchmarks with fewer training iterations. We also demonstrate that our framework enables distortion localization and identification, which paves a new way forward for model explainability and interactivity. Code and datasets are available at https://www.github.com/shngt/pit-qmm.

Index Terms— No-reference quality assessment, point clouds, large multimodal models, distortion localization

1. INTRODUCTION

Point clouds, collections of 3D points with attributes like color and opacity, are fundamental to applications such as autonomous driving, immersive gaming, and digital twins [1]. Their flexibility allows detailed spatial analysis with minimal geometric assumptions but makes them susceptible to distortions from sensor inaccuracies, compression, and transmission errors, which degrade perceptual quality and impair downstream tasks.

To address this, automated point cloud quality assessment (PCQA) has become a critical research focus. Traditional metrics like PSNR and SSIM [2], adapted from image/video quality assessment, fail to capture the complexities of 3D data. Learning-based methods are also not that effective, as most PCQA datasets contain only a few hundred samples.

Recently, large multimodal models (LMMs) trained on vast datasets have set benchmarks in 2D quality assessment. However, they are not easily extendable to the 3D case. Point-text multimodal models have been developed for semantic tasks such as object classification. However, due to computational constraints, they are restricted to smaller point clouds, for which the quality problem is no longer meaningful. Thus, while image-text models excel in quality assessment and point-text models in 3D comprehension, neither fully captures both aspects needed for PCQA.

To bridge this gap, we propose the Point-Image-Text Quality Multimodal Model (PIT-QMM), the first end-to-end point-image-text LMM for PCQA. PIT-QMM leverages complementary strengths of multiple modalities: PIT-QMM leverages the complementary strengths of different modalities: point cloud patches capture local variations often lost in 2D projections, image projections provide a global perspective, and text inputs add psychometric context and priming for the quality task. LMMs also excel in visual localization – linking specific regions with textual cues – which PIT-QMM leverages to accurately localize and categorize quality issues.

Our main contributions may be summarized as follows:

- We propose PIT-QMM, the first end-to-end point-image-text multimodal model tailored for PCQA. We also introduce task-aware prompts, efficient encoder-aware point cloud sampling, and a two-stage training strategy for effective multimodal fusion as an enhancement over prior work.

- We perform thorough benchmarking, and show that our model beats state-of-the-art (SOTA) methods by a large margin with fewer training iterations. We validate the importance of each modality with thorough ablations.

- We show that PIT-QMM can identify specific distortions and their locations when prompted. Not only does this enhance interpretability and overall utility, it hints at potential reasoning capabilities about quality. To our knowledge, this is the first exploration of quality localization in the point cloud domain.
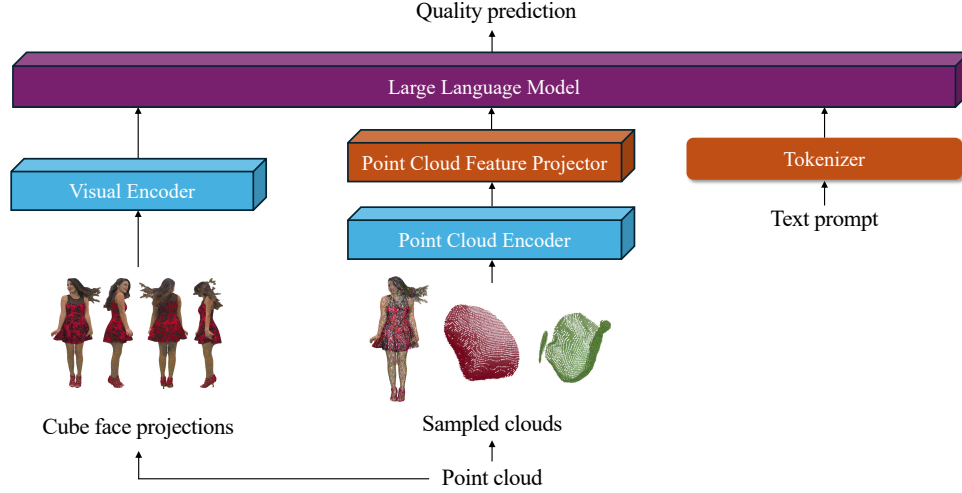
Supplementary material is available at https://dx.doi.org/10.60864/6kge-6c07.

Work done during an internship at Ericsson Research Silicon Valley.

Fig. 1. **An overview of the proposed Point-Image-Text Quality Multimodal Model (PIT-QMM).** PIT-QMM takes a raw point cloud and extracts both 2D and 3D views. Rich feature representations of these views are encoded by pretrained foundation models. These representations are then passed into a large multimodal model along with a textual description of the task and experimental setup, which is trained to predict quality scores.

2. RELATED WORK

Traditional NR-PCQA models involve handcrafted features and regressors, which have limited expressiveness. Deep learning-based methods, such as ResSCNN [3] and MM-PCQA [4], leverage 3D neural architectures for PCQA but struggle to generalize. CoPA [5] leverages the large-scale LS-PCQA [3] to employ contrastive learning to obtain robust quality features, but has high pre-training costs.

LMMs like Q-Align [6] achieve state-of-the-art results in 2D domains, but extending them to 3D by using 2D projections of the content does not work well, due to loss of local variations, occluded sections, and depth ambiguity. While methods like LMM-PCQA [7] integrate 2D LMM predictions with handcrafted cloud features, they do not leverage expressive learned representations or multimodal interactions. Moreover, extracting the handcrafted features involves expensive k-NN-based preprocessing on the entire cloud, making it impractical for large clouds.

PIT-QMM overcomes these issues by integrating deep point cloud encoders with existing image-text LMMs, enabling end-to-end multimodal training that fully exploits the complementary strengths of point cloud, image, and text modalities for superior PCQA performance.

3. METHOD

This section outlines the construction of our instruction-following dataset, the architecture of PIT-QMM, and our experimental setup and model modifications for distortion localization and identification. Section 3.1 addresses the dataset construction, while Section 3.2 details how the modalities are encoded and processed to produce the desired output.

3.1. Point-Image-Text Instruction-Following Quality Data

3.1.1. Point Clouds

A key challenge in including point clouds is their large size in quality assessment datasets. Popular encoders like Point-BERT [8] are pre-trained on point clouds containing thousands of points, whereas quality assessment datasets feature millions, making direct input infeasible. Therefore, the point cloud is subsampled to capture multiple levels of information.

First, furthest point sampling is applied to create a sparse global view, capturing overall shape and content-level attributes. Next, small local patches are randomly sampled to detect high-frequency local distortions. To construct each patch, we randomly select an anchor point and take its k-nearest neighbours. We also explored a two-scale variant which combines patches from the original and a downsampled cloud for multi-level granularity.

It is important to note that altogether these samples comprise only 3-5% of the total cloud. While this allows high sample efficiency and inference speed, they are not a holistic representation, necessitating a complementary global view.

3.1.2. Image Projections

To address the limitations of point cloud sampling, we incorporate multi-view image projections. For a point cloud P, we normalize it to zero-mean and unit-maximum distance with $\mathcal{N}(\cdot)$, then render $\mathcal{N}(P)$ into multi-view im-

Table 1. **Instruction following prompt.** {Experimental Setup} describes the psychometric setup. {im_tokens} are image tokens and {p_tokens} are point tokens.

{System Prompt}	
USER:	This is a point cloud rated for quality. It was displayed to a human in a single stimulus setup with absolute category ratings. {Experimental Setup}{im_tokens}<p_start>{p_tokens}<p_end> Can you rate the quality of the point cloud?
ASSISTANT:	The quality of the point cloud is excellent.

ages $\{x_i \in \mathbb{R}^{H \times W \times C}|_{i=1}^{6}\}$ from six perpendicular viewpoints (i.e., along the positive and negative directions of the x, y, z-axes) with fixed viewing distances. Where point cloud features provide local quality perspectives, these image projections provide a global quality perspective and allow leveraging pretrained image quality models.

3.1.3. Text

The textual component of the dataset primes the model for no-reference quality assessment against single-stimulus absolute category ratings, conveying psychometric context and driving it to draw on relevant world knowledge. Point cloud rendering parameters like point size and viewing distance, which influence quality, are also encoded in the prompt.

3.1.4. Final Instruction-Following Prompt

The final input format, as in Table 1, combines point cloud data, image projections, and text into a multi-modal question-answer structure. Special tokens `<p_start>` and `<p_end>` mark the start and end of the point cloud input. The model predicts discrete quality levels, as detailed in Section 3.3.1.

3.2. Model Architecture

As shown in Figure 1, our PIT-QMM is a generative model that aims to complete multi-modal sentences containing point clouds, images and text. The model consists of four main components - an image encoder f_{im}, a point cloud encoder f_{point}, a point cloud embedding projector f_{point_proj}, and a large language model (LLM) backbone f_{llm}.

The point cloud encoder f_{point} takes in a point cloud $P \in \mathbb{R}^{s \times n \times d}$, where s is the number of patches, n is the patch size and d is the feature dimension. The output is a sequence of patched point features $X \in \mathbb{R}^{s \times m \times c}$, where m is the number of patch features and c is the feature dimension. The projector f_{proj} is a multi-layer perceptron (MLP) that maps the point features X to point tokens $Y \in \mathbb{R}^{s \times m \times c'}$, where c' matches the dimension of the text and image tokens. Finally this is flattened to $Z \in \mathbb{R}^{sm \times c'}$, which we feed into f_{llm}.

The LLM f_{llm} takes in a sequence in $\mathbb{R}^{n' \times c'}$, where n' is the length of the total input token sequence. As a decoder-only LLM, it produces a probability distribution for the next token of size $\mathbb{R}^V$, where V is the vocabulary size.

3.3. Training and Inference

3.3.1. Label Smoothing and Discretization

As observed in Q-Align, LMMs optimized for quality prediction perform better when they are asked to produce discrete text labels, largely due to their bias to produce text as opposed to numeric values. We follow a similar discretization strategy during training and convert continuous quality scores to five-point Likert levels. During inference, discrete outputs are mapped to continuous scores by taking a weighted average of numeric label levels based on output token probabilities.

3.3.2. Two-stage Training

We employ a two stage training strategy. In the first feature alignment stage, the parameters of the point cloud projector are trained while others remain frozen. This stage uses small point clouds from the Cap3D [9] dataset. The point cloud sampling strategy is not applied here, as the input size is small. In the second instruction-tuning stage, we unfreeze the image abstractor and add LoRA [10] adapters to the LLM and the point cloud encoder. The model is fine-tuned end-to-end using the constructed quality dataset. During this stage, the image abstractor adapts to the domain of 2D projections, and the point cloud encoder adjusts to the domain of local patches with high-frequency variations.

3.4. Distortion Identification and Localization

In order to demonstrate the quality representation abilities of our model, we constructed a synthetic distortion identification and localization task. Specifically, we took pristine clouds, isolated a specific octant of each and applied a distortion from a predefined bank on it, and merged it back to the original cloud. The model is now fine-tuned to predict the octant and type of distortion from the distorted cloud.

Performing well on this task requires two key modifications. First, since random patches may not cover all octants, we deterministically sample patches to cover each octant. Since this may exceed the context length, we average pool the point cloud features within each patch before passing on to the LLM. Next, since the visual tokens inherently contain no information about which projection they belong to, we add learnable position embeddings shared across tokens originating from the same view. This allows the model to discriminate features from different views, which aids in localization.

4. EXPERIMENTS

4.1. Datasets

Our experiments are based on three popular PCQA datasets, namely LS-PCQA [3], SJTU-PCQA [11], and WPC [12]. LS-PCQA is a large-scale PCQA dataset with 104 pristine and 24,024 distorted point clouds. Each pristine point

cloud is impaired by 33 types of distortions at 7 levels of severity. The labels in LS-PCQA are mostly synthetically geenerated pseudo-MOSs, with only 930 samples having psychometrically-collected true MOSs. We term this subset LSPCQA-small and report results of ablations on it, along with WPC. SJTU-PCQA contains 9 reference and 378 distorted samples impaired by 7 types of distortions at 6 levels, while WPC contains 20 reference point clouds and 740 distorted samples disturbed by 5 types of distortions.

4.2. Evaluation Protocol

We tested PIT-QMM against other SOTA models on all datasets in Section 4.1. We first constructed instruction-tuning data from the raw datasets, as in Section 3.1. Each sample is thus a set of point cloud samples, cubic image projections and instruction text. We split each dataset into content-separated train-test sets in a 4:1 ratio. We minimized loss on the training set and obtained metrics on the test set. Due to the randomness involved in sampling from the point cloud, we computed metrics on the test set with 10 different seeds and took the mean. Finally, the test metrics were averaged over 5 different train-test splits to obtain the final reported metrics. Two popular evaluation metrics were used to quantify the agreement between predicted quality scores and MOSs: Spearman rank order correlation coefficient (SROCC) and Pearson linear correlation coefficient (PLCC).

4.3. Implementation Details

Our experiments were performed with PyTorch using 3×40 GB NVIDIA A100 GPUs. For the point cloud encoder, we used Point-BERT pretrained with ULIP-2 [13]. We sampled three patches in total, including the furthest point sample of the cloud. The point cloud projector is a randomly initialized MLP. The image encoder is a Vit-L/14 and the LLM is taken from mPLUG-Owl2.

For the alignment stage, we pretrained on the instruction-following variant of Cap3D from Point-LLM [14] for 3 epochs with a batch size of 12. We used a learning rate of 2×10^{-3} with cosine annealing and a warmup of 0.3. During finetuning, we trained on LS-PCQA for 5 epochs, SJTU-PCQA for 90 epochs and WPC for 30 epochs. We used a learning rate of 2×10^{-4} with cosine annealing and a warmup of 0.3. For LoRA, we used $r = 128$, $\alpha = 256$, and $p = 0.05$ on the multiway V_{proj} and Q_{proj} layers in mPLUG-Owl2 and the V and Q matrices in Point-BERT.

4.4. Comparison with State-of-the-Art Methods

We selected 15 state-of-the-art PCQA methods for comparison, including 9 FR-PCQA and 5 NR-PCQA methods. The FR-PCQA methods are MSE-p2point [15], HD-p2point [15], MSE-p2plane [16], HD-p2plane [16], PSNR-yuv [17], PointSSIM [18], PCQM [19], MS-GraphSIM [20],

and MPED [21]. The NR-PCQA methods are IT-PCQA [22], ResSCNN [3], MM-PCQA [4], CoPA+FT [5] and LMM-PCQA [7]. As only one split for LMM-PCQA is available, we reproduce the code and test on our splits. The other results are reported verbatim from the CoPA+FT paper.

4.4.1. Within-Dataset Performance

The within dataset performance on LS-PCQA, SJTU-PCQA and WPC is reported in Table 2. From the table, we observed that our model outperformed all NR-PCQA and FR-PCQA methods on all three datasets. Moreover, our model delivered robust performance across all datasets, despite variations in dataset scale, content, and distortion types.

4.4.2. Cross-Dataset Performance

The cross-dataset performance is reported in Table 3. Since LSPCQA is the largest dataset, followed by WPC, then SJTU, we trained on the full LSPCQA and tested on WPC and SJTU. We also trained on WPC and tested on SJTU. From the Table, it may be observed that PIT-QMM outperforms the other NR-PCQA models, thus demonstrating superior generalizability.

4.4.3. Training and Inference Cost

As demonstrated in Table 4, PIT-QMM converges to the best results when tuning for quality with fewer epochs compared to other SOTA learning-based methods. The savings were most significant on the large LS-PCQA dataset, where merely 5 epochs were sufficient to obtain SOTA performance. On the other hand, on the much smaller SJTU-PCQA dataset, we need more epochs, likely as more parameters have to tuned.

PIT-QMM is also efficient for inference, requiring $\sim$0.9s per sample of which $\sim$0.3s is for preprocessing. This is over 30x faster than LMM-PCQA for a cloud of 1 million points, which involves expensive handcrafted feature extraction.

4.5. Ablation Study

We conducted an ablation study to evaluate the contributions of different components in our proposed dataset construction strategy. Table 5 summarizes the results of this study. We used WPC and LSPCQA-small databases in these ablations.

First, using only 2D image projections to predict quality (row ①) yielded strong performance on both datasets, validating the use of pretrained vision models. However, performance improved when point cloud data was incorporated.

Next, we examined three point cloud sampling schemes: local patches (row ②), adding furthest point samples (row ③), and multi-scale sampling with half-scale patches (row ④). Sampling local patches alone showed limited improvement due to the pretrained encoder's domain gap, which focuses on semantic understanding of object-like point clouds. Adding furthest point samples improved results by introducing content-oriented features. However, incorporating multi-scale information had minimal effect. Likely, the patches

Table 2. Performance results on the LS-PCQA [3], SJTU-PCQA [11] and WPC [12] databases. "P" and "I" stand for the the point cloud and image modality, respectively. ↑ indicates that larger is better. The best performance results are marked in **RED** and the second best results are marked in **BLUE** for both FR-PCQA and NR-PCQA methods. "FT" indicates fine-tuning.

Ref	Modal	Methods	LS-PCQA		SJTU-PCQA		WPC	
			SROCC ↑	PLCC ↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
FR	P	MSE-p2po	0.325	0.528	0.783	0.845	0.564	0.557
	P	HD-p2po	0.291	0.488	0.681	0.748	0.106	0.166
	P	MSE-p2pl	0.311	0.498	0.703	0.779	0.445	0.491
	P	HD-p2pl	0.291	0.478	0.617	0.661	0.344	0.380
	P	PSNR-yuv	**0.548**	**0.547**	0.704	0.715	0.563	0.579
	P	PointSSIM	0.180	0.178	0.735	0.747	0.453	0.481
	P	PCQM	0.439	0.510	0.864	0.883	**0.750**	**0.754**
	P	MS-GraphSIM	0.389	0.348	**0.888**	**0.914**	**0.704**	**0.718**
	P	MPED	**0.659**	**0.671**	**0.898**	**0.915**	0.656	0.670
NR	I	IT-PCQA	0.326	0.347	0.539	0.629	0.422	0.468
	P	ResSCNN	0.594	0.624	0.834	0.863	0.735	0.752
	P+I	MM-PCQA	0.581	0.597	0.876	0.898	0.761	0.774
	P	CoPA+FT	0.613	0.636	**0.897**	**0.913**	0.779	0.785
	P	LMM-PCQA	**0.684**	**0.691**	0.730	0.724	**0.854**	**0.825**
	P+I	**PIT-QMM**	**0.751**	**0.766**	**0.906**	**0.916**	**0.872**	**0.844**

Table 3. Cross-dataset evaluation of NR-PCQA methods. Training and testing were both conducted on complete datasets. Results of PLCC are reported.

Train	Test	ResSCNN	MM-PCQA	CoPA+FT	LMM-PCQA	**PIT-QMM**
LS	SJTU	0.546	0.581	0.644	**0.656**	**0.682**
LS	WPC	0.466	0.454	0.516	**0.603**	**0.648**
WPC	SJTU	0.572	0.612	**0.643**	0.597	**0.671**

Table 4. Epochs required to converge to best results across all databases. Bold denotes the best performing model.

Method	Batch size	LS-PCQA	SJTU-PCQA	WPC
MM-PCQA	8	50	**50**	50
CoPA + FT	16	20	150	150
PIT-QMM	10	**5**	90	**30**

Table 5. Ablation study on the LSPCQA-small [3] and WPC [12] databases. ↑ indicates that larger is better.

Methods	LSPCQA-small		WPC	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑
①	0.684	0.664	0.837	0.804
②	0.722	0.681	0.866	0.835
③	0.734	0.699	0.872	0.844
④	0.730	0.694	0.865	0.839
⑤	0.343	0.322	0.447	0.405
⑥	0.733	0.704	0.870	0.832
⑦	0.737	0.706	0.869	0.838

need to be matched before processing, so that the encoders would become receptive to the fine details.

Using only point cloud features (row ⑤) significantly decreased performance, highlighting the domain gap in pre-trained encoders. Lastly, varying text prompts with additional task, psychometric (row ⑥), and rendering contexts (row ⑦) slightly improved performance.

4.6. Distortion Identification and Localization

We report the result of our localization experiments in Table 6. Since there are no existing baselines for this task, we compared against a ViT and a Q-Align model trained to predict the category and the octant of distortion from cubic projections. Synthetic data was generated from LSPCQA-small. First, we observe that the ViT baseline performed poorly on this task, likely due to a significant domain shift. Next, we observe that Q-Align also demonstrated strong localization abilities, which is expected for an LMM-based method. Finally, PIT-QMM outperformed both baselines with the help of the view-based positional embeddings and point cloud features.

Table 6. Accuracy on distortion identification and localization tasks. Bold denotes the best performing model.

Method	Identification Acc.	Localization Acc.
ViT	53.8%	28.1%
Q-Align	79.1%	72.7%
PIT-QMM	**84.3%**	**75.2%**

5. CONCLUSION

In this paper, we presented a novel end-to-end LMM-based NR-PCQA algorithm. By leveraging complementary information from different modalities and large pretrained encoders, our proposed PIT-QMM model predicts quality scores across a wide variety of distortion and content types. Extensive experiments show that PIT-QMM achieves competitive performance across varied benchmarks with fewer training iterations than other SOTA models. Preliminary experiments show that PIT-QMM can also pinpoint the nature and location of distortions with high accuracy, which indicates an exciting new path towards interactive and explainable quality agents.

6. REFERENCES

[1] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, et al., "Pointclip: Point cloud understanding by clip," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8552–8562.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] Y. Liu, Q. Yang, Y. Xu, and L. Yang, "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.

[4] Z. Zhang, W. Sun, X. Min, Q. Zhou, J. He, Q. Wang, and G. Zhai, "Mm-pcqa: Multi-modal learning for no-reference point cloud quality assessment," *arXiv preprint arXiv:2209.00244*, 2022.

[5] Z. Shan, Y. Zhang, Q. Yang, H. Yang, Y. Xu, J. Hwang, X. Xu, and S. Liu, "Contrastive pre-training with multi-view fusion for no-reference point cloud quality assessment," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25942–25951.

[6] H. Wu, Z. Zhang, W. Zhang, et al., "Q-align: Teaching lmms for visual scoring via discrete text-defined levels," *arXiv preprint arXiv:2312.17090*, 2023.

[7] Z. Zhang, H. Wu, Y. Zhou, C. Li, W. Sun, et al., "Lmm-pcqa: Assisting point cloud quality assessment with lmm," *arXiv preprint arXiv:2404.18203*, 2024.

[8] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *CVPR*, 2022.

[9] T. Luo, C. Rockwell, H. Lee, and J. Johnson, "Scalable 3d captioning with pretrained models," *arXiv:2306.07279*, 2023.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, et al., "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[11] Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, and J. Sun, "Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration," *IEEE Transactions on Multimedia*, vol. 23, pp. 3877–3891, 2021.

[12] Q. Liu, H. Su, Z. Duanmu, W. Liu, and Z. Wang, "Perceptual quality assessment of colored 3d point clouds," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 8, pp. 3642–3655, 2022.

[13] L. Xue, N. Yu, S. Zhang, A. Panagopoulou, J. Li, R. Martín-Martín, J. Wu, C. Xiong, et al., "Ulip-2: Towards scalable multimodal pre-training for 3d understanding," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27091–27101.

[14] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, "Pointllm: Empowering large language models to understand point clouds," *arXiv preprint arXiv:2308.16911*, 2023.

[15] R. Mekuria, Z. Li, C. Tulvan, and P. Chou, "Evaluation criteria for point cloud compression," *ISO/IEC MPEG*, 2016.

[16] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, "Geometric distortion metrics for point cloud compression," in *IEEE International Conference on Image Processing*, 2017, pp. 3460–3464.

[17] E. M. Torlig, E. Alexiou, T. A. Fonseca, R. L. de Queiroz, and T. Ebrahimi, "A novel methodology for quality assessment of voxelized point clouds," in *Applications of Digital Image Processing XLI*, 2018, vol. 10752, pp. 174–190.

[18] E. Alexiou and T. Ebrahimi, "Towards a point cloud structural similarity metric," in *IEEE International Conference on Multimedia and Expo Workshops*, 2020, pp. 1–6.

[19] G. Meynet, Y. Nehmé, J. Digne, and G. Lavoué, "Pcqm: A full-reference quality metric for colored 3d point clouds," in *International Conference on Quality of Multimedia Experience*, 2020, pp. 1–6.

[20] Y. Zhang, Q. Yang, and Y. Xu, "Ms-graphsim: Inferring point cloud quality via multiscale graph similarity," in *ACM International Conference on Multimedia*, 2021, pp. 1230–1238.

[21] Q. Yang, Y. Zhang, S. Chen, Y. Xu, J. Sun, and Z. Ma, "Mped: Quantifying point cloud distortion based on multiscale potential energy discrepancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6037–6054, 2022.

[22] Q. Yang, Y. Liu, S. Chen, Y. Xu, and J. Sun, "No-reference point cloud quality assessment via domain adaptation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21179–21188.