Points Meet Pixels: Bridging 2D Vision-Language Model and 3D Perception Gaps for Point Cloud Quality Assessment

Mingxuan Li, Zihao Huang, Xiaohui Chu, Fazhan Zhang, Bohan Fu, Runze Hu*

School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, China
{limx1630, huangzihhhh, zfz63622, fubohan809, hrzlpk2015}@gmail.com, 3120225380@bit.edu.cn

Abstract

Vision-Language Models (VLMs) have demonstrated significant progress in quality assessment tasks. However, a fundamental paradox arises when their application to Point Cloud Quality Assessment (PCQA). Existing VLMs, designed for image-text pairs, are inherently incompatible with 3D point cloud data due to the modality gap. While some PCQA research attempts to adapt point clouds to VLMs by 2D projection, this approach inevitably sacrifices crucial spatial structure information essential for accurate quality assessment. Conversely, directly integrating a dedicated 3D branch into a VLM-based PCQA framework introduces feature space misalignment and an influx of quality-insensitive information. To bridge these fundamental conflicts hindering VLMs' adaptation to PCQA, we propose the **PMP-PCQA** framework, which leverages the inherent mapping relationship between points and pixels to seamlessly apply VLMs to PCQA. Our approach introduces three key innovations: a **Spatial Awareness Enhancer(SAE)** module that enriches the image features with spatial coordinate clues to reinforce geometric awareness in 2D visual representations; a **Fine-to-coarse Consistency Alignment(FCA)** module that bridges the gap between 2D and 3D modalities by leveraging point-pixel correspondences to construct bridging features; and a **Text-Guided Adaptive Miner(TAM)** module that dynamically suppresses quality-insensitive features to mine discriminative visual clues for PCQA. Extensive evaluations demonstrate that PMP-PCQA consistently outperforms state-of-the-art methods across multiple benchmarks.

Code — https://github.com/Limx1630/PMP-PCQA

Introduction

As a convenient format for representing the 3D world, point cloud primarily consists of spatial coordinates, often supplemented with additional attributes such as color, surface normal, and reflectance (Guo et al. 2021). Propelled by applications in autonomous driving (Cui et al. 2022), robotics (Selvaratnam and Bazazian 2025), and immersive environments (Tliba et al. 2024), the rapid proliferation of 3D point cloud data has driven an urgent demand for robust PCQA metrics. Such metrics are essential for defining the point cloud quality and ultimately enhancing the Quality of Experience in 3D applications (Liu et al. 2021b). However, unlike traditional 2D media, point clouds pose unique challenges for PCQA due to their inherent data characteristics. Consequently, developing robust PCQA metrics remains a persistent and significant challenge for both academia and industry (Wang et al. 2021; Hu et al. 2025; Chu et al. 2025a).

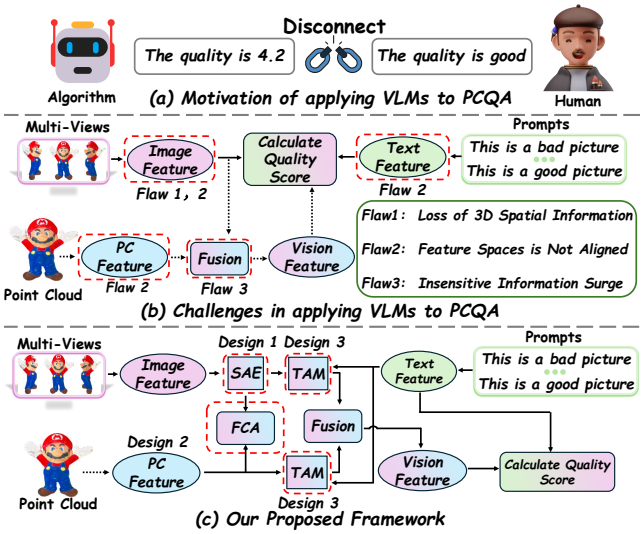

Figure 1: (a) Motivation of applying VLMs to PCQA.(b) Challenges of VLM-based PCQA Methods. (c) Our proposed framework to address these challenges.

Current PCQA research (Liu et al. 2025a) has predominantly focused on developing no-reference methods, driven by the unavailability of reference point clouds in practical applications. Despite their remarkable performance achieved by directly regressing Mean Opinion Scores (MOS), these NR-PCQA methods share a critical blind spot in characterizing human subjective perception. As shown in Fig.1 (a), humans typically use qualitative adjectives to assess an object's quality rather than relying on quantitative scores (Wu et al. 2023b). For instance, we might say "this quality is perfect," instead of stating that "this quality score is 5.6.". This insight has spurred emerging studies to explore the application of vision-language models (VLMs) in 2D media quality

*Corresponding author.

assessment (Wu et al. 2023a; Srinath et al. 2024), offering a promising path forward. However, applying VLMs to the PCQA field remains nascent.

As shown in Fig. 1 (b), VLM-based PCQA methods face a fundamental contradiction: existing VLMs, which are designed for image-text pairs, are inherently incompatible with 3D point cloud data due to the modality gap. To bridge this gap, current approaches typically project 3D point clouds into 2D plane. This transforms the task of aligning point cloud data with text into the more tractable one of aligning image data with text (Liu et al. 2025b; Wu et al. 2022). However, this projection-based approach inherently discards the essential 3D characteristics of point clouds, failing to meet the core requirements of PCQA—precise perception of structural degradations and 3D geometric distortions (Chai et al. 2024; Chu et al. 2025b).

Conversely, integrating a dedicated 3D processing branch into a VLM-based framework faces two critical bottlenecks. First, the 3D feature space is misaligned with the VLM's joint vision-text representation space, hindering the effective transfer of its pretrained knowledge. Second, fusing multi-view and 3D features introduces a significant amount of perceptually irrelevant redundancy, which dilutes the model's focus on quality-related visual clues. These challenges ultimately lead to the severe performance degradation observed when applying VLMs to PCQA.

To address these challenges, We propose the innovative **P**oint **M**eet **P**ixel for **P**oint **C**loud **Q**uality **A**ssessment framework **PMP-PCQA**, as illustrated in Fig. 1 (c). Its core innovation lies in fully leveraging the fine-grained mapping between 3D points and their projected 2D pixels. This approach ensures compatibility with powerful 2D VLMs while preserving the inherent 3D perceptual information to the greatest extent.

Specifically, to enhance geometric awareness within the 2D branch, we introduce a **Spatial Awareness Enhancer (SAE)** module, which augments image features with spatial coordinate clues. To facilitate the transfer of 2D visual-language knowledge to 3D network, we propose a **Fine-to-coarse Consistency Alignment (FCA)** module that establishes bridging features via point-pixel correspondence to eliminate the feature space gap between modalities. To address quality-insensitive redundancy, we introduce a **Text-Guided Adaptive Miner (TAM)** module that leverages the text information related to quality to dynamically mine the salient visual clues indicative of quality degradation. We summarize our main contributions as follows:

- PMP-PCQA introduces the first VLM-based PCQA framework with a dedicated 3D branch. By establishing fine-grained point-pixel mappings, it addresses the fundamental incompatibility that hinders the effective application of standard VLMs to PCQA.
- PMP-PCQA introduces a SAE module that enriches the image features with spatial coordinate clue, a FCA module that bridges the 2D-3D gap, transferring visual-language knowledge to the 3D point network, and a TAM module that dynamically suppresses quality-insensitive features to mine discriminative visual clues for PCQA.
- Across three popular PCQA benchmarks covering a wide range of content, distortion types, and scales, PMP-PCQA outperformed all competing methods, demonstrating superior correlation with human subjective perception.

Related Work

Point Cloud Quality Assessment

PCQA is broadly categorized into three types: full-reference (FR), reduced-reference (RR), and no-reference (NR). For FR-PCQA, the MPEG adopted the p2point (Mekuria et al. 2016a) metrics in point cloud compression standardization. For RR-PCQA, Viola et al. used statistical information of geometry, color and normal vector for PCQA (Viola and Cesar 2020). However, both FR and RR methods are limited by their dependency on reference objects, which are often unavailable in practice. Consequently, NR-PCQA has become the primary research focus. Xie et al. (Xie et al. 2023) enhanced the feature representation capability by projecting point clouds onto diverse 2D maps. Shan et al. proposed the GPA-Net (Shan et al. 2023), which captures structural and texture perturbations within a multi-task framework.

Despite their promising performance, these methods ignore a critical blind spot in characterizing human subjective perception: humans tend to use qualitative adjectives to describe object quality instead of quantitative scores.

Vision-Language Models for Quality Assessment

By leveraging the rich visual-language prior inherently encapsulated within VLMs, these models have demonstrated superior performance in 2D quality assessment. CLIP-IQA (Wang, Chan, and Loy 2022) shows an antonym prompt pairing strategy to convert quality regression to a binary classification task and regard the softmax results as the quality scores. CLIP-VQA (Xing et al. 2024) leverages CLIP-based language descriptions and video spatiotemporal features to predict video quality. However, applying VLMs to the PCQA field remains nascent.

Existing VLM-based PCQA methods (Zhang et al. 2024c) project 3D point clouds into 2D views to adapt them for VLM processing. However, such a projection-based approach discards the essential 3D characteristics, failing to capture their complete geometric information.

Method

While equipping VLMs with 3D perceptual promises comprehensive PCQA, three key challenges arise: spatial information loss, feature space mismatch, and redundant quality-insensitive information. To address these challenges, we propose PMP-PCQA, which leverages point-pixel correspondence. The overall architecture of PMP-PCQA, illustrated in Fig. 2, takes point clouds, multi-view images, and quality descriptions as its inputs. Following feature extraction, the representations are processed through three core modules to predict a probability distribution over a set of qualitative adjectives. The final score is computed as a linear combination of the predefined values for each quality level, weighted by their corresponding predicted probabilities. This section elaborates on the principles of PMP-PCQA.

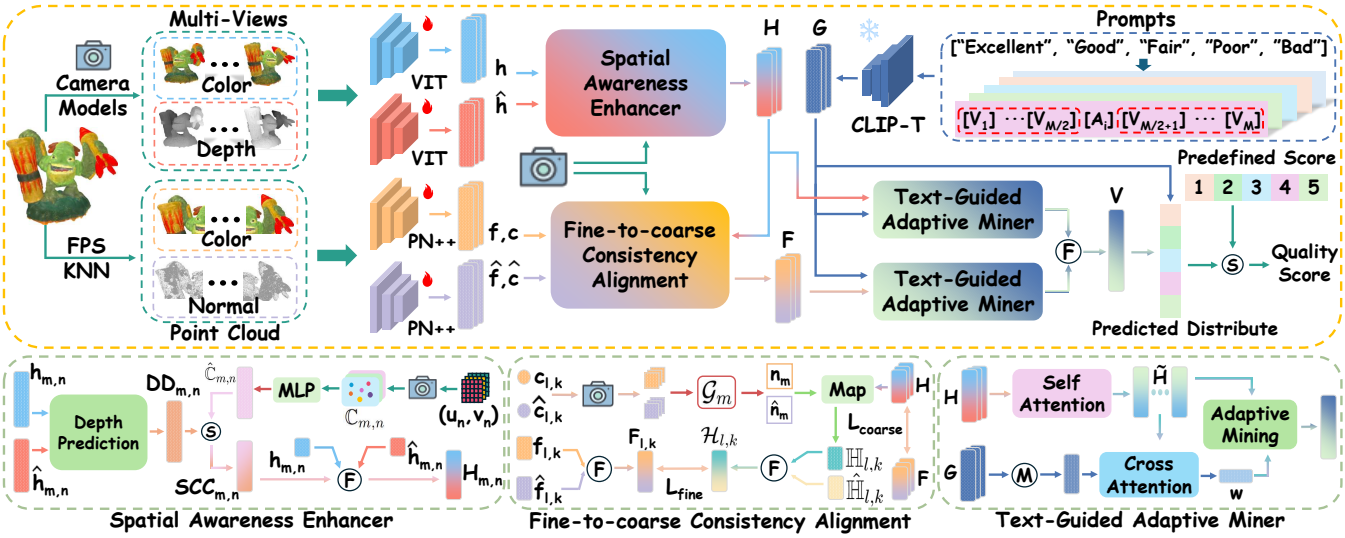

Figure 2: The overall Framework of PMP-PCQA. The main components contain three parts: Spatial Awareness Enhancer (SAE) module, Fine-to-coarse Consistency Alignment (FCA) module, and Text-Guided Adaptive Miner (TAM) module.

Multi-Modal Feature Extraction

Point Cloud Feature Extraction A colored point cloud $\mathcal{P}^{rgb} = \{(p_i^{xyz}, p_i^{rgb})\}_{i=1}^{N}$ comprises N points, each composed of spatial coordinates p_i^{xyz} and color attribute p_i^{rgb}. To better capture local structural variations, we augment each point with a normal attribute p_i^{normal}, constructing a normal point cloud $\mathcal{P}^{nor} = \{(p_i^{xyz}, p_i^{normal})\}_{i=1}^{N}$ (Wang et al. 2023). Owing to the dense nature of point clouds, we decompose them into L patches via Farthest Point Sampling (FPS) and K-Nearest Neighbors (KNN) following (Zhang et al. 2023). The resulting patch sets are denoted as $\mathbf{P}^{rgb} \in \mathbb{R}^{L\times K\times 6}$ and $\mathbf{P}^{nor} \in \mathbb{R}^{L\times K\times 6}$, where K represents the number of points per patch.

Given the distinct characteristics of color and normal point clouds, we employ two independent PointNet++ (Qi et al. 2017) to encode $\mathbf{P}^{rgb}$ and $\mathbf{P}^{nor}$ respectively.

The two PointNet++ branches generate features $f, \hat{f} \in \mathbb{R}^{L\times\mathcal{K}\times C}$ and corresponding centroid coordinates $c, \hat{c} \in \mathbb{R}^{L\times\mathcal{K}\times 3}$ for the color and normal point clouds, respectively. Here, $\mathcal{K}$ denotes the number of point clusters retained at the final sampling stage, and the centroids provide the 3D spatial locations for these clusters.

Multi-view Images Feature Extraction We render the colored point cloud into a set of 2D multi-view images using camera models following (Liu et al. 2021a). The projection process of each camera is defined by a function C:

$$[u, v, d] = C([x, y, z]). \tag{1}$$

The projection function C transforms a 3D world coordinate $[x, y, z]$ into a 2D pixel location $[u, v]$ and an associated depth value d. We employ M camera models $\{C_1, \cdots, C_M\}$ to generate multi-view color images $\mathbf{I}^{color}$ and depth images $\mathbf{I}^{depth}$. Features $h, \hat{h} \in \mathbb{R}^{M\times N\times C}$ are then extracted using two independent pretrained CLIP visual encoders , where N denotes the number of non-overlapping patches per image.

Textual Feature Extraction In the language branch, we adopt the standard text rating levels as quality level tokens defined by ITU-R BT.500 (Drogu 2019), such as ("excellent", "good", "fair", "poor", and "bad"). Following findings in (Zhou et al. 2022b,a) that fixed prompts may limit model optimization, we insert these adjective tokens into the middle of a unified learnable prompt template to construct a set of quality-aware prompts $\mathbf{T} = \{\mathbf{T}_i | i = 1, 2, \cdots, \mathcal{L}\}$:

$$\mathbf{T}_i = [V_1] \cdots [V_{\mathcal{M}/2}][A_i][V_{\mathcal{M}/2+1}] \cdots [V_{\mathcal{M}}]. \tag{2}$$

where $\mathcal{L}$ is the number of prompts variants, $[A_i]$ is the i-th quality-level adjective token, $[V]$ is a learnable context token embedding, and $\mathcal{M}$ is the hyper-parameter controlling context token count. The corresponding textual features $G \in \mathbb{R}^{\mathcal{L}\times C}$ are extracted via the frozen CLIP text encoder.

Spatial Awareness Enhancer

We propose a SAE module to augment image features with spatial information. Let $h_{m,n}$ denote the feature vector of the n-th image patch (from either a color or depth view) in the m-th projected view, which is generated by camera C_m.

Meanwhile, each image patch is associated with the pixel coordinates of its center, serving as a 2D anchor for its feature representation. These pixel coordinates can be calculated based on the grid layout of the N patches:

$$\begin{aligned} u_n &= \frac{W}{\sqrt{N}}\left(\left(n \mod \sqrt{N}\right) + \frac{1}{2}\right), \\ v_n &= \frac{H}{\sqrt{N}}\left(\left\lfloor \frac{n}{\sqrt{N}} \right\rfloor + \frac{1}{2}\right). \end{aligned} \tag{3}$$

Where $\lfloor\rfloor$ denotes the floor function.

A single pixel (u_n, v_n) corresponds to a ray in 3D space extending from the camera center through that pixel. To capture comprehensive spatial information, we uniformly

sampling $\mathcal{N}$ point along this ray within the valid depth range $[d_{\min}, d_{\max}]$, obtaining a series of 3D points at varying depths, and project them into world coordinates:

$$\begin{aligned}\mathbb{C}_{m,n} &= \{C_m^{-1}(u_n, v_n, d_i) | i = 0, 1, \cdots, \mathcal{N}-1\} \in \mathbb{R}^{\mathcal{N}\times 3}, \\ where \quad d_i &= d_{min} + \frac{i \times (d_{max} - d_{min})}{\mathcal{N}-1}.\end{aligned} \tag{4}$$

This gives the 3D point set $\mathbb{C}_{m,n}$ for the n-th patch in the m-th view. We then map $\mathbb{C}_{m,n}$ into a high-dimensional feature space using a Multi-Layer Perceptron (MLP), yielding the transformed feature $\widehat{\mathbb{C}}_{m,n} \in \mathbb{R}^{\mathcal{N}\times C}$.

To address the inherent uncertainty in image-based depth estimation, we predict a depth distribution $DD_{m,n}$ for each patch from its corresponding image and depth features. This distribution is used to compute a weighted combination of the high-dimensional point features, yielding the 3D spatial coordinate clue $SCC_{m,n}$ for the n-th patch in the m-th view:

$$\begin{aligned}DD_{m,n} &= \mathrm{MLP}(h_{m,n} + \hat{h}_{m,n}) \in \mathbb{R}^{\mathcal{N}}, \\ SCC_{m,n} &= \sum_{i=1}^{\mathcal{N}} \left(DD_{m,n}[i] \cdot \hat{\mathbb{C}}_{m,n}[i,:] \right) \in \mathbb{R}^{C}.\end{aligned} \tag{5}$$

Finally, we fuse the color feature, depth feature, and the 3D spatial coordinate clues to obtain the updated patch feature. The resulting spatially enhanced image feature $H \in \mathbb{R}^{M\times N\times C}$ is given by:

$$H[m,n,:] = \mathrm{Fusion}(h_{m,n}, \hat{h}_{m,n}, SCC_{m,n}). \tag{6}$$

Fine-to-coarse Consistency Alignment

To bridge this gap between feature representation spaces and enable the effective transfer of prior knowledge from pre-trained 2D VLMs to 3D networks, we introduce the FCA module. Specifically, for the k-th local region within the l-th point cloud patch, we define its features $f_{l,k}, \hat{f}_{l,k}$ and 3D centroid coordinates $c_{l,k}, \hat{c}_{l,k}$ as described previously.

Subsequently, we map the 3D coordinates $c_{l,k}$ and $\hat{c}_{l,k}$ onto the m-th view plane using the camera model C_m, which yields the corresponding pixel coordinates:

$$(u,v)_m = C_m(c_{l,k}), (\hat{u},\hat{v})_m = C_m(\hat{c}_{l,k}). \tag{7}$$

Based on the mapped pixel coordinates, the patch partitioning rules of the 2D view space are further utilized to determine that the pixel falls into the n-th patch region of the m-th view. This process establishes local feature correspondences across modalities as follows:

$$\begin{aligned}n_m = \mathcal{G}_m((u,v)_m), \hat{n}_m = \mathcal{G}_m((\hat{u},\hat{v})_m), \\ f_{l,k} \Longleftrightarrow \mathbb{H}_{l,k} = \frac{\sum_{m=1}^{M} H_{m,n_m}}{M}, \\ \hat{f}_{l,k} \Longleftrightarrow \hat{\mathbb{H}}_{l,k} = \frac{\sum_{m=1}^{M} H_{m,\hat{n}_m}}{M}.\end{aligned} \tag{8}$$

Here, the function $\mathcal{G}_m$ maps pixel coordinates $(u,v)_m$ to their corresponding patch index n in the m-th view.

Building on this,we construct a bridge feature $\mathcal{H}$ from the original 2D features H, ensuring each token is spatially aligned with the 3D features F. Then we perform the consistency constraint between the bridge feature and 3D features:

$$\begin{aligned}F[l,k,:] &= \mathrm{Fusion}(f_{l,k}, \hat{f}_{l,k}) \in \mathbb{R}^{C}, \\ \mathcal{H}[l,k,:] &= \mathrm{Fusion}(\mathbb{H}_{l,k}, \hat{\mathbb{H}}_{l,k}) \in \mathbb{R}^{C}, \\ \mathcal{L}_{\text{fine}} &= \frac{1}{L\times\mathcal{K}} \sum_{l=1}^{L}\sum_{k=1}^{\mathcal{K}} \mathcal{C}(\mathcal{H}[l,k,:], F[l,k,:]).\end{aligned} \tag{9}$$

In addition to fine-grained alignment, we also enforce coarse-grained alignment to maintain consistency in the global representations after multi-view and multi-patch fusion. This is achieved by applying a consistency constraint on the aggregated features:

$$\begin{aligned}\mathcal{L}_{coarse} &= \mathcal{C}(Agg(H), Agg(F)), \\ \mathcal{L}_{con} &= \mathcal{L}_{\text{fine}} + \mathcal{L}_{coarse}.\end{aligned} \tag{10}$$

Where $\mathcal{C}(\cdot)$ is the NT-Xent contrastive loss (Chen et al. 2020) and $\mathrm{Agg}(\cdot)$ aggregates dimensions via mean pooling over all but the last dimension.

Text-Guided Adaptive Miner

Integrating multi-view features with point cloud features inevitably introduces abundant visual clues that is irrelevant to quality. This redundant information distracts the model from key quality indicators, limiting PMP-PCQA's performance. To address this issue, we introduce the TAM module, which leverages semantic guidance from the text branch to mine the most salient quality-degradation visual clues.

Taking the 2D branch as an example, we flatten the visual feature H and apply self-attention to obtain $\tilde{H} \in \mathbb{R}^{MN\times C}$. We then compute its semantic correlation with the text description via cross-attention, generating an attention map:

$$\mathbf{w} = \mathrm{softmax}\left(\frac{\tilde{H}(\mathrm{Mean}(G))^T}{\sqrt{d_k}} \right) \in \mathbb{R}^{MN}. \tag{11}$$

Here, $\mathrm{Mean}(\cdot)$ performs mean-pooling, d_k is the key dimension, and $\mathbf{w}$ quantifies the relevance of each visual token in $\tilde{H}$ to the global quality description in G.

Building on the attention weights, we introduce an adaptive mining mechanism to retain the most critical visual clues. This mechanism dynamically sets a threshold θ_{ada} based on the attention entropy, which modulates the focus akin to the "attentional focus" in human vision: in complex (high-entropy) scenes, it automatically filters out irrelevance to concentrate on salient regions. Formally, the attention entropy E is computed, and the adaptive threshold is derived as:

$$\begin{aligned}E &= -\sum_{i=1}^{MN} w_i \cdot \log_2(w_i + \epsilon), \\ \theta_{\text{ada}} &= \theta_{\min} + (\theta_{\max} - \theta_{\min}) \cdot \frac{E}{\log_2(MN)}.\end{aligned} \tag{12}$$

Here, ϵ is a smoothing term. $\theta_{\max}, \theta_{\min}$ are threshold bounds.

We then retain only the tokens whose attention scores exceed θ_{ada}. For instance, if θ_{ada} corresponds to the 40th percentile, the top 60% of tokens are kept. The preserved

attention weights are renormalized and used for a weighted average of their corresponding features, yielding the refined 2D representation $\tilde{\mathbb{H}}$:

$$\tilde{\mathbb{H}} = \sum_{i=1}^{MN} \left(\frac{w_i \cdot \mathbb{I}(w_i \geq \mathcal{Q}(\mathbf{w}, \theta_{ada}))}{\sum_{j=1}^{MN} w_j \cdot \mathbb{I}(w_j \geq \mathcal{Q}(\mathbf{w}, \theta_{ada}))} \right) \cdot \tilde{H}[i, :]. \tag{13}$$

Here, $\mathcal{Q}$ denotes the function for calculating quantiles, and $\mathbb{I}(.)$ is an indicator function. The same adaptive mining operation is applied to the 3D features $\mathbf{F}$, producing the refined representation $\tilde{\mathbf{F}} \in \mathbb{R}^C$. The final visual representation $\mathbf{V} \in \mathbb{R}^C$ is obtained by averaging $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{F}}$.

Quality Assessment and Loss Function

We compute cosine similarity between visual and text features, and apply softmax to get the probability that the point cloud's quality matches each text descriptor.

$$s_i = \frac{\mathbf{V} \odot \mathbf{G}_i}{\|\mathbf{V}\| \cdot \|\mathbf{G}_i\|}, \quad p_i = \frac{\exp(s_i)}{\sum_{j=1}^{\mathcal{L}} \exp(s_j)}. \tag{14}$$

Subsequently, we define the mapping function $G(.)$ from qualitative adjectives to quantitative scores (e.g., $G(\text{excellent}) = 5$). The final predicted scores can be calculated as follows:

$$\text{Score} = \sum_{i=1}^{L} p_i * G(A_i). \tag{15}$$

Our loss function includes two parts: the Wasserstein Distance loss (Hou, Yu, and Samaras 2017) and the Consistency loss. The WD loss optimizes the model by quantifying the discrepancy between the Cumulative Distribution Function (CDF) of the predicted probability distribution and the ground truth. The consistency loss as described by Eq.15; The total loss of our entire model is finally expressed as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{WD} + \lambda_2 \mathcal{L}_{con}. \tag{16}$$

where the hyper-parameter λ_1, λ_2 is to balance these losses.

Experimental

Experimental Settings

Datasets. We employ three popular PCQA datasets: SJTU-PCQA (Yang et al. 2021) (9 references, 7 distortions, 378 samples), LS-PCQA (Liu et al. 2023a) (85 references, 31 distortions, 930 samples), and BASICS (Ak et al. 2024) (75 references, 4 distortions, 1494 samples). Considering the limited dataset scale, we employ k-fold cross-validation to evaluate model performance: 9-fold for SJTU-PCQA and 5-fold for LS-PCQA and BASICS.

Comparison Methods and Evaluation Metrics. We select 13 state-of-the-art quality assessment methods for comparison, which consists of 5 FR-PCQA and 8 NR-PCQA methods. The specific methods are listed in the "Method" column of Tab.1. We employ Spearman's Rank-Order Correlation Coefficient (SRCC), Pearson's Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE) to quantify model performance. Following established practice in PCQA (Zhu et al. 2024), we employ a five-parameter nonlinear regression to compensate for scale discrepancies between algorithmic predictions and human-rated MOS.

Implementation Details. For data preprocessing, the number of point cloud patches L and image projections M are both set to 6. Each point cloud patch has $K = 2048$ points and images have a resolution of 224×224. The camera models select the FoV perspective cameras. The hyper-parameter $\mathcal{L}$ and $\mathcal{M}$ is set to 5 and 16, respectively. For feature extraction process, the number of local regions $\mathcal{K}$ is set to 196 and we select the CLIP-ViT-B/16 variant in our experiments. For our proposed modules, we set the $\mathcal{N}, d_{max}, d_{min}, \theta_{max}$ and θ_{min} to 64, 4, 0.8, 0.2. For loss function, we set λ_1, λ_2 to 1, 0.2. We set the Fusion$(.)$ function to calculate the average of features. For the training process, We employ the Adam optimizer with an initial learning rate of 4×10^{-6} and a weight decay of 10^{-4}. Our model is trained for 100 epochs with a batch size of 6 on one NVIDIA RTX 4090 GPU.

Overall Performance

The performance comparison results are presented in Tab.1. We can draw some following conclusions: 1) Our method outperforms all the compared methods. Specifically, our model achieves improvements in SRCC by 3.5% (0.957 vs. 0.925), 4.4% (0.763 vs. 0.731), and 4.2% (0.892 vs. 0.856) compared to the second-best method on the SJTU-PCQA, LS-PCQA, and BASICS databases, respectively. 2) The performance of other PCQA method drops sharply on the LS-PCQA and BASICS datasets, which have more complex data scales, content, and distortion types. In contrast, our model still achieves stable performance. 3) Compared to other PCQA methods with "T" modality, our model consistently achieves superior performance. Further analysis reveals that these methods typically project point cloud data directly into 2D images, failing to effectively leverage the inherent 3D nature of point clouds, which consequently leads to their performance decline.

Ablation Study

Ablation on Different Components We conducted systematic module ablation experiments on the SJTU-PCQA and LS-PCQA datasets, with detailed results presented in Tab.2. Removing the FCA module causes significant performance degradation, primarily because pretrained VLMs are trained on 2D images, creating a representational gap that hinders the seamless transfer of prior knowledge from 2D pre-trained VLM into 3D networks. Building upon FCA, the SAE module supplements missing spatial structural information in the 2D branch, enhancing perception of spatial hierarchical distortion. The TAM module adaptively filters critical visual clues with decision-making value for quality assessment, improving feature discriminability. Ultimately, through synergistic interactions with the FCA module, SAE and TAM collectively establish the model's performance advantages and enhance overall effectiveness.

Hyperparameter Analysis In this section, we conduct hyperparameter analysis on SJTU dataset. For the SAE module, we vary the number of sampling points (16 to 256). As Tab.3

Type	Methods	Modal	SJTU-PCQA			LS-PCQA			BASICS		
			SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓
FR	MSE-p2po (Mekuria et al. 2016b)	PC	0.729	0.812	1.361	0.301	0.427	0.744	0.774	0.849	0.559
	MSE-p2pl (Tian et al. 2017)	PC	0.628	0.594	2.282	0.286	0.454	0.734	0.836	0.895	0.468
	PCQM (Meynet et al. 2020)	PC	0.864	0.885	1.086	0.426	0.208	0.789	0.808	0.891	0.478
	GrahSim (Yang et al. 2022)	PC	0.878	0.845	1.032	0.331	0.358	0.767	0.813	0.895	0.465
	TCDM (Zhang et al. 2024a)	PC	0.930	0.910	0.891	0.413	0.438	0.739	0.757	0.874	0.505
NR	3D-NSS (Zhang et al. 2022)	PC	0.714	0.738	1.769	0.532	0.584	0.682	0.617	0.657	0.883
	RES-SCNN (Liu et al. 2023b)	PC	0.880	0.889	0.878	0.620	0.648	0.615	0.352	0.391	0.975
	MM-PCQA (Zhang et al. 2023)	PC+I	0.910	0.923	0.772	0.605	0.644	0.621	0.738	0.793	0.628
	GMS-3DQA (Zhang et al. 2024b)	I	0.911	0.918	0.787	0.645	0.678	0.606	0.807	0.895	0.472
	PIT-PCQA (Xie et al. 2024)	PC+I+T	0.903	0.927	0.725	0.712	0.723	0.574	0.831	0.851	0.514
	CoPA (Shan et al. 2024)	I	0.897	0.913	0.739	0.613	0.636	0.594	0.785	0.812	0.534
	CLIP-PCQA (Liu et al. 2025b)	I+T	0.922	0.937	0.693	0.727	0.741	0.552	0.856	0.911	0.462
	LMM-PCQA (Zhang et al. 2024c)	I+T	0.925	0.929	0.724	0.731	0.744	0.547	0.845	0.867	0.443
	PMP-PCQA	**PC+I+T**	**0.957**	**0.967**	**0.624**	**0.763**	**0.776**	**0.512**	**0.892**	**0.936**	**0.337**

Table 1: Performance comparison among the proposed and the state-of-the-art PCQA methods on the SJTU, LS-PCQA, and BASICS datasets. "PC","I","T" stand for the method is based on the point cloud, image and text modality, respectively.

Settings	SJTU		LS-PCQA	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑
w/o SAE	0.943	0.951	0.743	0.756
w/o FCA	0.923	0.936	0.722	0.729
w/o TAM	0.946	0.953	0.740	0.755
Only SAE	0.904	0.909	0.707	0.714
Only FCA	0.936	0.947	0.734	0.741
Only TAM	0.913	0.921	0.713	0.717
PMP-PCQA	**0.957**	**0.967**	**0.763**	**0.776**

Table 2: Ablation study of different components on SJTU and LS-PCQA datasets.

Points-Num	SR↑	PL↑
16	0.945	0.956
32	0.950	0.956
64	**0.957**	**0.967**
128	0.948	0.957
256	0.941	0.952

Table 3: Hyperparameter Analysis of SAE module.

Threshold	SR↑	PL↑
0.2	0.936	0.941
0.4	0.941	0.947
0.6	0.947	0.952
0.8	0.943	0.947
Adaptive	**0.957**	**0.967**

Table 4: Hyperparameter Analysis of TAM module.

Index	Image	Point	Text	SRCC↑	PLCC↑
①	✓	×	×	0.913	0.918
②	×	✓	×	0.907	0.912
③	✓	✓	×	0.925	0.939
④	✓	×	✓	0.934	0.942
⑤	×	✓	✓	0.839	0.847
⑥	✓	✓	✓	**0.957**	**0.967**

Table 5: Ablation study of Different modalities.

Train on	SJTU		LS-PCQA	
Test on	LS-PCQA	BASICS	SJTU	BASICS
MM-PCQA	0.201	0.178	0.723	0.423
GMS-3DQA	0.396	0.568	0.776	0.459
CLIP-PCQA	0.324	0.625	0.812	0.496
LMM-PCQA	0.405	0.552	0.662	0.511
PMP-PCQA	**0.431**	**0.678**	**0.843**	**0.560**

Table 6: Cross-dataset validation. Both the training and testing are on the complete dataset. The final result is expressed as SRCC.

shows, performance does not improve monotonically with more sampling points. Under our experimental setup, optimal model performance is achieved with 64 sampling points. For the TAM module, we compare our entropy-adaptive threshold with fixed thresholds (Tab.4). Our adaptive method consistently outperforms fixed thresholds because different point clouds vary significantly in content, distortion types, and severity. Using a uniform threshold for salient visual cue mining cannot universally adapt to all samples.

Ablation on Different Modalities To verify the impact of multimodal information, we performed modality ablation experiments on the SJTU dataset, with the results presented in Tab.5. Notably, We assess the performance of removing the text branch thorough an MLP for quality regression. Results demonstrate that optimal model performance is achieved only when incorporating multiple modalities. Moreover, combining language and point clouds causes significant performance degradation due to the gap between point cloud and textual features. In addition, while combining language and image delivers competent performance, it remains inferior to our method. This limitation fundamentally stems from images' inherent lack of spatial information, resulting in insufficient sensitivity to structural distortions.

Cross Dataset Validation

To verify the generalization and robustness of PMP-PCQA when confronted with diverse data distributions, we conducted cross-database validation experiments, with the results summarized in Tab.6. Given that LS-PCQA and SJTU-PCQA share certain reference point clouds, we excluded these overlapping point cloud groups from LS-PCQA to prevent information leakage. It can be clearly observed from Tab.6 that all methods exhibit relatively low performance in cross-dataset evaluations, particularly when trained on the SJTU dataset and tested on other datasets. This is primarily due

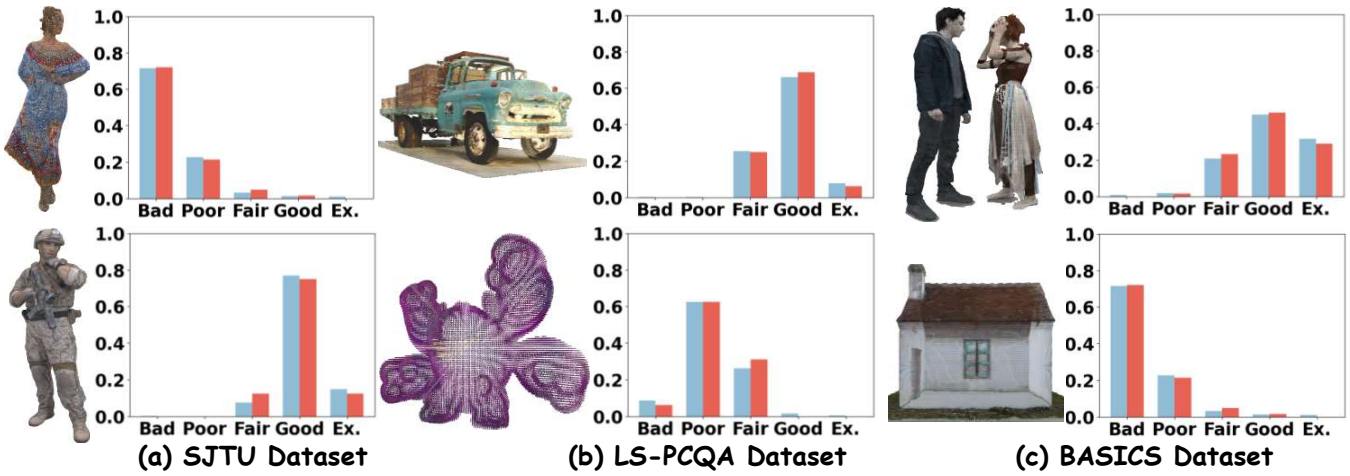

Figure 3: Comparison the True Probability Distribution (Red) versus the Predicted Probability Distribution (Blue). Ex. denotes Excellent.

Figure 4: Left: Original image with reference patch marked in red. Right: Heatmap showing cosine similarity between the 3D spatial coordinate clues of all image patches and the reference patch.

Figure 5: (a) (b) display attention score after TAM Module processing for two samples. Left: red boxes highlight distorted regions. Right: weight maps show focused areas and regions without red overlay represent discarded features.

to the significant differences in data distributions. However, our model still outperforms other models by a large margin, demonstrating superior generalization ability.

Visualization Analysis

Scoring Distribution Visualization We present the predicted probability distributions alongside the true probability distributions for multiple samples across different quality levels of text descriptions in Fig.3. The figure clearly demonstrates that our method utilizing discrete text descriptions can effectively and accurately model the authentic human subjective scoring process.

Spatial Correlation Visualization To validate the 3D spatial coordinate clues can supplement spatial information for 2D visual features, we examined the correlation of these clues within a sample image. As shown in Fig.4, the left panel displays the original image, where a red bounding box marks a randomly selected reference location. Subsequently, we computed the cosine similarity between the SCC of all other image patches and those of the reference location. The heatmap is shown in the right panel, which clearly demonstrates the clues encode spatial information.

Quality Cue Mining Comparison Fig.5 compares the weight updates resulting from our proposed TAM module for two different samples. The results demonstrate that our approach effectively guides the model to focus on visual clues more relevant to quality assessment. Furthermore, the adaptive threshold mechanism exhibits better adaptation to different samples. As illustrated in the figure, under the adaptive scheme, the θ value adjusts accordingly: it decreases for sample (a) and increases for sample (b), demonstrating its adaptability to varying sample characteristics.

Conclusion

In our paper, to address the numerous challenges arising from applying VLMs to PCQA, we propose an innovative framework called PMP-PCQA. This framework enhances the geometric perception ability of 2D visual representations by embedding spatial coordinate clues, and constructs a cross-modal feature bridge through the mapping relationship between points and pixels, thereby bridging the gap between 2D and 3D modalities and transferring 2D vision-language knowledge to 3D branch. Furthermore, our adaptive mining mechanism empowers the model to autonomously mine salient visual clues most relevant to quality assessment. According to the extensive experiments on three datasets, PMP-PCQA gets better performance against SOTA methods.

Acknowledgments

This paper was supported in part by the National Science Foundation of China under Grant 62301041, and in part by Beijing Institute of Technology Research Fund Program for Young Scholars.

References

Ak, A.; Zerman, E.; Quach, M.; Chetouani, A.; Smolic, A.; Valenzise, G.; and Le Callet, P. 2024. BASICS: Broad Quality Assessment of Static Point Clouds in a Compression Scenario. *IEEE Transactions on Multimedia*, 26: 6730–6742.

Chai, X.; Shao, F.; Mu, B.; Chen, H.; Jiang, Q.; and Ho, Y.-S. 2024. Plain-PCQA: No-Reference Point Cloud Quality Assessment by Analysis of Plain Visual and Geometrical Components. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 6207–6223.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709.

Chu, X.; Duan, H.; Wen, Z.; Xu, L.; Hu, R.; and Xiang, W. 2025a. Union-Domain Knowledge Distillation for Underwater Acoustic Target Recognition. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–16.

Chu, X.; Zhou, H.; Zhang, Y.; Zhang, Y.; Hu, R.; Duan, H.; Huang, Y.; Zheng, Y.; and Ji, R. 2025b. Attention-driven acoustic properties learning for underwater target ranging. *Pattern Recognition*, 164: 111560.

Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; and Cao, D. 2022. Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review. *IEEE Transactions on Intelligent Transportation Systems*, 23(2): 722–739.

Drogu, J. 2019. Recommendation ITU-R BT.500-13 Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service (television). Technical report, International Telecommunication Union - Radiocommunication Sector (ITU - R).

Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; and Bennamoun, M. 2021. Deep Learning for 3D Point Clouds: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12): 4338–4364.

Hou, L.; Yu, C.-P.; and Samaras, D. 2017. Squared Earth Mover's Distance-based Loss for Training Deep Neural Networks. arXiv:1611.05916.

Hu, R.; Chu, X.; Dou, D.; Liu, X.; Liu, Y.; and Qi, B. 2025. Toward Real-World Applicability: Lightweight Underwater Acoustic Localization Model Through Knowledge Distillation. *IEEE Journal of Oceanic Engineering*, 50(2): 1429–1442.

Liu, Q.; Yuan, H.; Su, H.; Liu, H.; Wang, Y.; Yang, H.; and Hou, J. 2021a. PQA-Net: Deep No Reference Point Cloud Quality Assessment via Multi-View Projection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12): 4645–4660.

Liu, Y.; Yang, Q.; Xu, Y.; and Yang, L. 2023a. Point Cloud Quality Assessment: Dataset Construction and Learning-based No-reference Metric. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(2s).

Liu, Y.; Yang, Q.; Xu, Y.; and Yang, L. 2023b. Point Cloud Quality Assessment: Dataset Construction and Learning-based No-reference Metric. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(2s).

Liu, Y.; Yang, Q.; Zhang, Y.; Xu, Y.; Yang, L.; Xu, X.; and Liu, S. 2025a. Once-Training-All-Fine: No-Reference Point Cloud Quality Assessment via Domain-Relevance Degradation Description. *IEEE Transactions on Broadcasting*, 71(2): 616–630.

Liu, Y.; Zhang, Y.; Shan, Z.; and Xu, Y. 2025b. CLIP-PCQA: Exploring Subjective-Aligned Vision-Language Modeling for Point Cloud Quality Assessment. arXiv:2501.10071.

Liu, Z.; Li, Q.; Chen, X.; Wu, C.; Ishihara, S.; Li, J.; and Ji, Y. 2021b. Point Cloud Video Streaming: Challenges and Solutions. *IEEE Network*, 35(5): 202–209.

Mekuria, R.; Li, Z.; Tulvan, C.; and Chou, P. 2016a. Evaluation criteria for PCC (Point Cloud Compression).

Mekuria, R.; Li, Z.; Tulvan, C.; and Chou, P. 2016b. Evaluation criteria for point cloud compression. Technical report, ISO/IEC MPEG.

Meynet, G.; Nehmé, Y.; Digne, J.; and Lavoué, G. 2020. PCQM: A Full-Reference Quality Metric for Colored 3D Point Clouds. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. PointNet++: deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5105–5114.

Selvaratnam, D.; and Bazazian, D. 2025. 3D Reconstruction in Robotics: A Comprehensive Review. *Computers Graphics*, 130: 104256.

Shan, Z.; Yang, Q.; Ye, R.; Zhang, Y.; Xu, Y.; Xu, X.; and Liu, S. 2023. GPA-Net:No-Reference Point Cloud Quality Assessment with Multi-task Graph Convolutional Network. arXiv:2210.16478.

Shan, Z.; Zhang, Y.; Yang, Q.; Yang, H.; Xu, Y.; Hwang, J.-N.; Xu, X.; and Liu, S. 2024. Contrastive Pre-Training with Multi-View Fusion for No-Reference Point Cloud Quality Assessment. arXiv:2403.10066.

Srinath, S.; Mitra, S.; Rao, S.; and Soundararajan, R. 2024. Learning Generalizable Perceptual Representations for Data-Efficient No-Reference Image Quality Assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 22–31.

Tian, D.; Ochimizu, H.; Feng, C.; Cohen, R.; and Vetro, A. 2017. Geometric distortion metrics for point cloud compression. In *2017 IEEE International Conference on Image Processing (ICIP)*, 3460–3464.

Tliba, M.; Zhou, X.; Viola, I.; Cesar, P.; Chetouani, A.; Valenzise, G.; and Dufaux, F. 2024. Enhancing Immersive Experiences through 3D Point Cloud Analysis: A Novel Framework for Applying 2D Visual Saliency Models to 3D Point Clouds. In *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*, 307–313.

Viola, I.; and Cesar, P. 2020. A Reduced Reference Metric for Visual Quality Evaluation of Point Cloud Contents. *IEEE Signal Processing Letters*, 27: 1660–1664.

Wang, J.; Chan, K. C. K.; and Loy, C. C. 2022. Exploring CLIP for Assessing the Look and Feel of Images. arXiv:2207.12396.

Wang, J.; Chen, P.; Zheng, N.; Chen, B.; Principe, J. C.; and Wang, F.-Y. 2021. Associations between MSE and SSIM as cost functions in linear decomposition with application to bit allocation for sparse coding. *Neurocomputing*, 422: 139–149.

Wang, S.; Wang, X.; Gao, H.; and Xiong, J. 2023. Non-Local Geometry and Color Gradient Aggregation Graph Model for No-Reference Point Cloud Quality Assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6803–6810.

Wu, H.; Chen, C.; Hou, J.; Liao, L.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2022. FAST-VQA: Efficient End-to-end Video Quality Assessment with Fragment Sampling. *Proceedings of European Conference of Computer Vision (ECCV)*.

Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2023a. Towards Explainable In-the-Wild Video Quality Assessment: A Database and a Language-Prompted Approach. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 1045–1054. ACM.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Li, C.; Liao, L.; Wang, A.; Zhang, E.; Sun, W.; Yan, Q.; Min, X.; Zhai, G.; and Lin, W. 2023b. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. *arXiv preprint arXiv:2312.17090*. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Corresponding Authors: Zhai, Guangtao and Lin, Weisi.

Xie, W.; Liu, Y.; Wang, K.; and Wang, M. 2024. LLM-Guided Cross-Modal Point Cloud Quality Assessment: A Graph Learning Approach. *IEEE Signal Processing Letters*, 31: 2250–2254.

Xie, W.; Wang, K.; Ju, Y.; and Wang, M. 2023. pmBQA: Projection-based Blind Point Cloud Quality Assessment via Multimodal Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3250–3258.

Xing, F.; Li, M.; Wang, Y.-G.; Zhu, G.; and Cao, X. 2024. CLIPVQA:Video Quality Assessment via CLIP. arXiv:2407.04928.

Yang, Q.; Chen, H.; Ma, Z.; Xu, Y.; Tang, R.; and Sun, J. 2021. Predicting the Perceptual Quality of Point Cloud: A 3D-to-2D Projection-Based Exploration. *IEEE Transactions on Multimedia*, 23: 3877–3891.

Yang, Q.; Ma, Z.; Xu, Y.; Li, Z.; and Sun, J. 2022. Inferring Point Cloud Quality via Graph Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3015–3029.

Zhang, Y.; Yang, Q.; Zhou, Y.; Xu, X.; Yang, L.; and Xu, Y. 2024a. TCDM: Transformational Complexity Based Distortion Metric for Perceptual Point Cloud Quality Assessment. *IEEE Transactions on Visualization and Computer Graphics*, 30(10): 6707–6724.

Zhang, Z.; Sun, W.; Min, X.; Wang, T.; Lu, W.; and Zhai, G. 2022. No-Reference Quality Assessment for 3D Colored Point Cloud and Mesh Models. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7618–7631.

Zhang, Z.; Sun, W.; Min, X.; Zhou, Q.; He, J.; Wang, Q.; and Zhai, G. 2023. MM-PCQA: Multi-Modal Learning for No-reference Point Cloud Quality Assessment. *IJCAI*.

Zhang, Z.; Sun, W.; Wu, H.; Zhou, Y.; Li, C.; Chen, Z.; Min, X.; Zhai, G.; and Lin, W. 2024b. GMS-3DQA: Projection-Based Grid Mini-patch Sampling for 3D Model Quality Assessment. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(6).

Zhang, Z.; Wu, H.; Zhou, Y.; Li, C.; Sun, W.; Chen, C.; Min, X.; Liu, X.; Lin, W.; and Zhai, G. 2024c. LMM-PCQA: Assisting Point Cloud Quality Assessment with LMM. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 7783–7792. Association for Computing Machinery.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision (IJCV)*.

Zhu, L.; Cheng, J.; Wang, X.; Su, H.; Yang, H.; Yuan, H.; and Korhonen, J. 2024. 3DTA: No-Reference 3D Point Cloud Quality Assessment With Twin Attention. *IEEE Transactions on Multimedia*, 26: 10489–10502.