SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2026, Vol. 69, Iss. 2, 122306:1–122306:15
https://doi.org/10.1007/s11432-025-4631-2

Towards versatile multimedia quality assessment for visual communications

Zicheng ZHANG, Ziheng JIA, Chunyi LI, Yingjie ZHOU, Xiaohong LIU, Xiongkuo MIN & Guangtao ZHAI*

Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Received 28 March 2025/Revised 21 July 2025/Accepted 16 October 2025/Published online 12 January 2026

Abstract With the rapid advancement and increasing demands of multimedia applications in visual communications, the visual quality of multimedia content has emerged as a pivotal factor, profoundly impacting service quality and user experience. Traditionally, visual quality assessment has concentrated on individual modalities such as images, videos, and 3D models, with evaluation models designed separately due to the distinct characteristics of each media type. However, from the perspective of the human vision system, visual quality across these modalities is interconnected and shares a unified perceptual basis. To address this, we introduce a versatile multimedia visual quality assessment framework tailored for visual communications, which unifies quality assessment of images, videos, and 3D models within a single large multi-modal model (LMM). This integrated approach enables simultaneous quality evaluation across all three modalities, effectively harnessing cross-domain knowledge while reducing the inefficiencies and resource overhead of deploying separate models in multimodal communication systems. Experimental results show that our proposed framework, X-QA, delivers robust quality assessment performance across images, videos, and 3D models, establishing a strong technical foundation and opening new possibilities for future visual communication applications requiring sophisticated multimodal quality evaluations.

Keywords versatile visual quality assessment, multimedia, large multi-modal models

Citation Zhang Z C, Jia Z H, Li C Y, et al. Towards versatile multimedia quality assessment for visual communications. Sci China Inf Sci, 2026, 69(2): 122306, https://doi.org/10.1007/s11432-025-4631-2

1 Introduction

The escalating demand for high-quality multimedia in visual communications and an enhanced quality of experience (QoE) has driven the creation of specialized quality assessment tools to predict the quality of diverse media types [1–4]. These tools are instrumental in improving a wide range of applications critical to visual communication systems, from low-level image enhancements (e.g., dehazing [5], brightening [6], and denoising [7]) to compression and transmission frameworks [8, 9], as well as 3D scanning and reconstruction processes [10]. Depending on the assessment target, current mainstream visual quality assessment typically focuses on three distinct domains: image quality assessment (IQA), video quality assessment (VQA), and 3D quality assessment (3DQA). IQA examines traditional technical distortions (e.g., noise, blur, and compression artifacts) alongside aesthetic qualities (e.g., color harmony, balance, and emotional impact) that influence visual perception in communication. VQA evaluates spatial distortions (akin to IQA) and temporal dynamics in video streams (e.g., frame rate, resolution, and motion artifacts), ensuring seamless and clear playback essential for effective video-based communication. Meanwhile, 3DQA assesses the visual quality of 3D models (e.g., point clouds, voxels, and meshes), focusing on attributes like texture, resolution, and structural fidelity, which are vital for immersive visual experiences in modern communication platforms.

However, the structural differences between these modalities (images as pixel-based structures, videos with spatio-temporal dimensions, and 3D models with three-dimensional spatial properties) combined with the input limitations of traditional models, have historically restricted quality assessment to single-modality approaches. These conventional models are tailored to specific modalities and are challenging to adapt for cross-modal visual quality evaluation. Yet, from the perspective of the human visual system (HVS), the visual quality of these three modalities shares common features. For instance, compression in images, videos, and 3D models often results in blurring (such as unclear textures or indistinct object boundaries) while color shifts or noise similarly degrade the viewing

* Corresponding author (email: zhaiguangtao@sjtu.edu.cn)

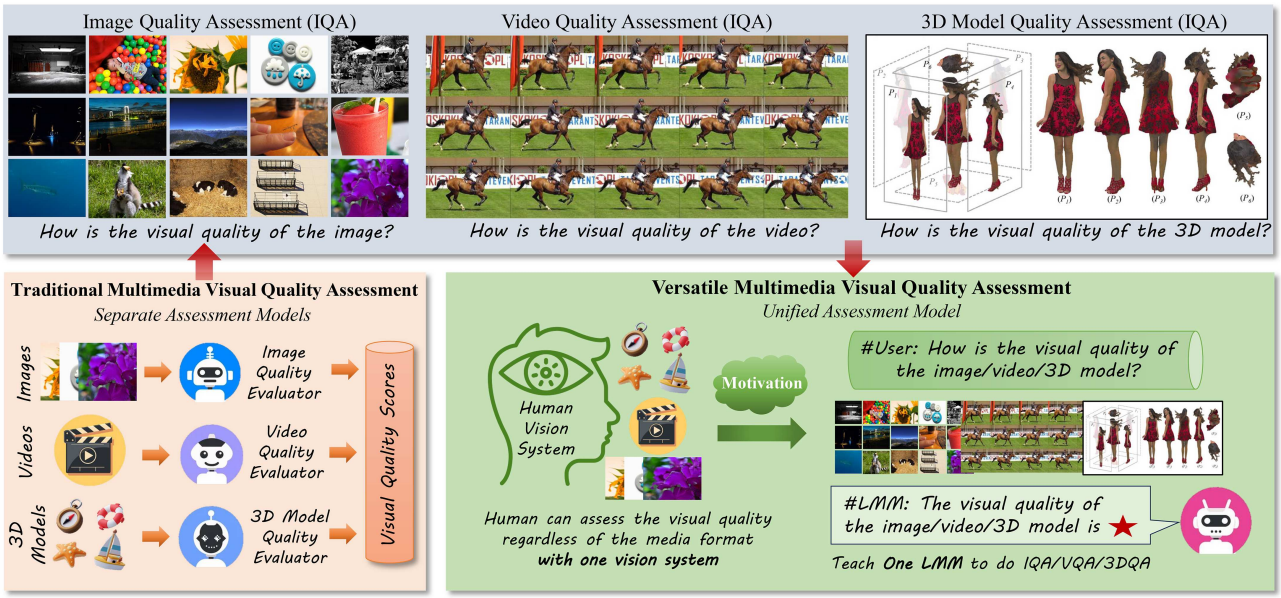

Figure 1 (Color online) Traditional multimedia visual quality assessment vs. versatile multimedia visual quality assessment. Given the diverse formats of multimedia content and the capabilities of traditional models, conventional quality assessment typically involves designing separate quality evaluators for different modalities. However, humans can accurately perceive the visual quality of various modalities within a single vision system. This inspires us to propose a versatile quality assessment framework that leverages a one LMM to simultaneously evaluate the visual quality of images, videos, and 3D content using a single model.

experience across all three. This commonality suggests that consolidating the visual quality assessment of these modalities into a single model is both logical and feasible. Traditional single-modality models, however, lack the capacity to perceive and process the complexity of multiple diverse modalities, making such an approach impractical with conventional methods. Recent advancements in large multimodal models (LMMs) and their application to visual quality assessment have changed this landscape. With their superior perceptual understanding and vast model capacity, LMMs can effectively learn to assess all three tasks simultaneously, enabling the development of a versatile multimedia visual quality assessment framework (as shown in Figure 1). Nevertheless, this approach introduces two significant challenges.

How can the input for these three modalities be unified for LMMs? LMMs natively support image inputs [11]. Videos can be treated as sequences of sampled frames, while 3D models can be rendered into a series of images from fixed viewpoints [12]. While this suffices for IQA, both frame sampling for videos and rendering for 3D models result in the loss of critical quality information (e.g., temporal continuity in videos or geometric features in 3D models). To address this, we propose an additional quality projector module that employs modality-specific feature extractors to capture video and 3D-specific characteristics. These features are then integrated into the LMM for comprehensive analysis, minimizing the loss of quality-related information.

How can mixed training of quality data across the three modalities be achieved? Quality knowledge injection into LMMs typically involves converting traditional mean opinion score (MOS) annotated datasets into question-answer pairs for supervised fine-tuning (SFT) [13]. However, the fixed question-answer pair format may interfere with the model's ability to distinguish between modalities. To overcome this, we introduce a modality-specific system prompt strategy, embedding tailored prompts in the text input to guide the LMM in differentiating between modalities. This enhances the model's ability to deliver targeted quality assessments for each modality.

Building upon the challenges and motivations outlined earlier, this paper introduces X-QA, the first experimental framework for versatile multimedia visual quality assessment. Specifically, we begin by converting the MOSs from existing quality assessment datasets into question-answer pairs with adjective-based ratings (e.g., poor, good, excellent). For the image modality, we perform no additional preprocessing, directly utilizing the raw data for quality evaluation. For the video modality, we sample frames at a rate of 1 FPS to conduct spatial quality analysis, then employ the SlowFast [14] network as a projector to extract temporal quality features from consecutive frames. For the 3D model modality, we project the model from fixed cubic viewpoints and leverage PointNet++ [15] as a projector to extract three-dimensional quality features from local patches of the model. To address the challenges of mixed-modality training, we design distinct system prompts tailored to each modality. These prompts help bridge the gap between modalities, enabling the model to optimize effectively for diverse data types. Experimental re-

sults demonstrate that the proposed X-QA framework achieves strong performance across visual quality assessment tasks for all three modalities (images, videos, and 3D models). This advancement paves the way for multimedia applications, such as enhanced content delivery, virtual reality experiences, and automated quality control in 3D reconstruction, by providing a unified and robust quality assessment solution. In all, the contributions of this paper can be summarized as follows.

• The paper introduces X-QA, the first experimental framework for versatile multimedia visual quality assessment, integrating quality evaluation of images, videos, and 3D models into a single LMM. This unified approach contrasts with traditional methods that rely on separate, modality-specific evaluators.

• To enable simultaneous assessment across modalities, the framework converts MOS into adjective-based question-answer pairs (e.g., poor, good, excellent). It employs modality-specific preprocessing: direct use of raw images, 1 FPS frame sampling with SlowFast for video temporal features, and fixed cubic viewpoint projections with PointNet++ for 3D model features. This ensures comprehensive quality feature extraction despite modality differences.

• To overcome challenges in mixed-modality training, the paper proposes tailored system prompts for each modality (images, videos, 3D models). These prompts guide the LMM to distinguish and optimize quality assessments across diverse data types, enhancing cross-modal performance.

2 Related work

2.1 Image quality assessment

Image quality assessment (IQA) predicts perceptual visual quality, vital for visual communication and computational imaging [16–25]. Wang et al. [26] critiqued error sensitivity metrics like PSNR, shifting IQA toward human perception-aligned methods. For blind IQA, Moorthy et al. [27, 28] introduced visual importance pooling and NSS-based indices. Later advances include frequency-domain sharpness [29], gradient statistics with Adaboost [30], and deep learning for distortion-generic models [31]. Recent efforts target diverse content, like VR images [32], 360-degree IQA [33], smartphone photos [34], and face images [35], using techniques from full-reference [36] to unsupervised methods [36]. No-reference IQA now leverages multiscale features [37], attention networks [38], and CNNs [39]. These developments reflect ongoing efforts to enhance IQA accuracy and robustness.

2.2 Video quality assessment

Video quality assessment (VQA) is essential for compression, communication, and enhancement applications [40–42]. Wang et al. [43] focused on structural distortion for perceived quality, while Seshadrinathan et al. [44] integrated motion into VQA metrics. Advances in human visual perception models include speed perception statistics [45], multi-dimensional analysis [46], temporal distortion metrics [47], and motion artifact estimation [48]. Efficiency is boosted by fragment sampling [49–51] and exposure correction assessments [52, 53].

2.3 3D quality assessment

3D quality assessment (3DQA) has grown with VR, AR, and metaverse applications [1]. Alexiou et al. [54] and Su et al. [55] targeted point cloud quality, with Su building a subjective study database. Diniz et al. [56] introduced a multi-distance metric for uneven point distributions. Machine learning advances include PQA-Net and MS-GraphSIM [57], using multi-view projection and multiscale features. No-reference metrics for colored 3D models [10, 58] and point-to-distribution methods [57] enhance 3DQA. Yang et al. [59] tackled domain adaptation, while Zhang et al. [60, 61] predicted quality via rendered images/videos. Multi-modal MM-PCQA [62] fuses 2D texture and 3D geometry. Efficiency gains come from Zhang's work [63, 64], alongside 3D digital human assessments [65, 66], shifting 3DQA from traditional to deep learning methods.

2.4 Large multi-modal models

Large language models (LLMs) like GPT-4 [67], T5 [68], and LLaMA [69] excel in linguistic tasks across diverse knowledge domains. They expand into multimodal applications by integrating visual inputs via CLIP [70] and adaptation modules, as seen in LMMs [11, 71–74]. OpenFlamingo [75] employs gated cross-attention in pretrained language encoders, while InstructBLIP [73] enhances BLIP-2 [76] with vision-language tuning. Open-source LMMs, such as the LLaVA series [11, 77, 78], leverage GPT-4 [67] to generate data for fine-tuning vision-language models.

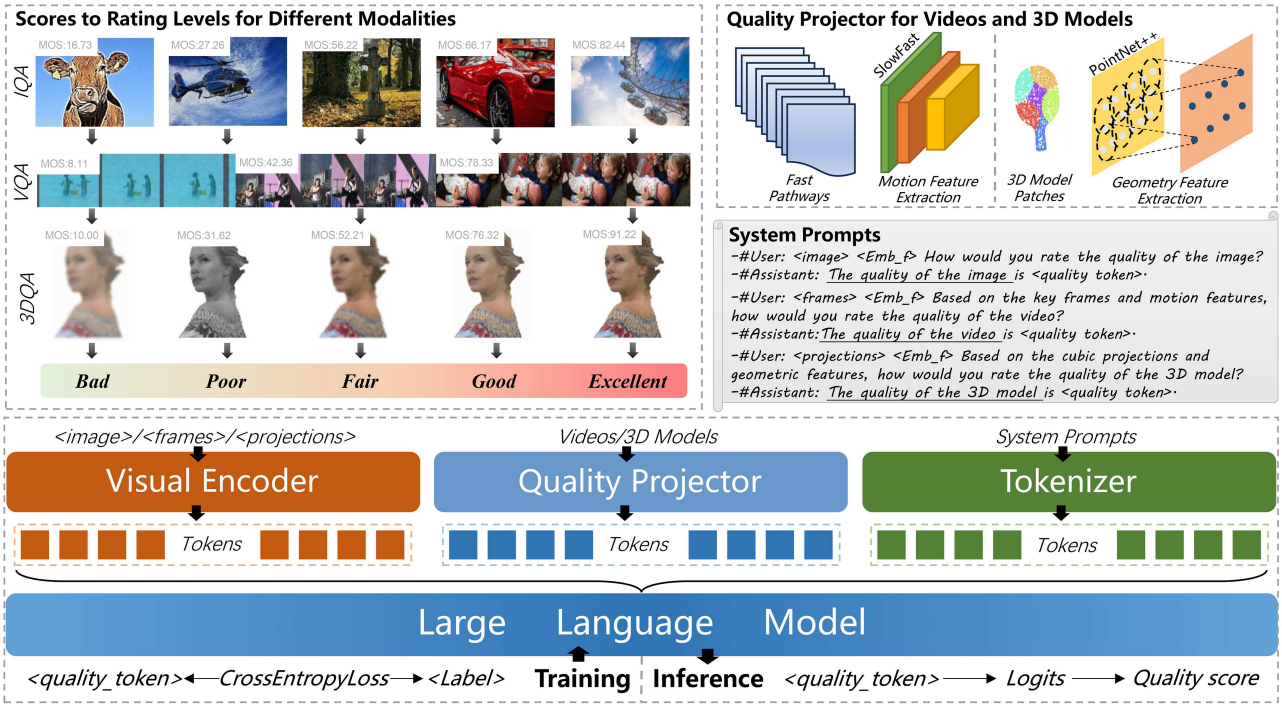

Figure 2 (Color online) Framework of the proposed X-QA. The MOSs of the images, videos and 3D models are converted to adjective quality ratings. Quality projectors are employed to capture specific quality attributes that might otherwise be overlooked, such as motion features absent in individual frames for VQA or geometry details lost in projections for 3DQA. This quality knowledge is subsequently integrated into the LMM through supervised fine-tuning, enabling X-QA to infer quality scores across different modalities.

3 Methods

In this section, we detail the methodology of the proposed X-QA framework (shown in Figure 2), designed to integrate visual quality assessment across three modalities (images, videos, and 3D models) within a single LMM. The approach addresses two primary challenges: unifying multimodal inputs for the LMM and enabling effective mixed-modality training. Below, we describe the data preprocessing, feature extraction, model architecture, and training strategy, supported by theoretical reasoning and mathematical formulations.

3.1 Data preprocessing and quality projectors

To enable the LMM to process diverse modalities (images, videos, and 3D models), we preprocess the input data to align with the model's native capabilities while preserving modality-specific quality information. The preprocessing varies by modality, as outlined below.

3.1.1 *Image modality*

For IQA, raw images are used directly as input to the LMM without additional preprocessing. Images are represented as $I \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the height and width, and 3 corresponds to RGB channels. This preserves traditional distortions (e.g., noise, blur) and aesthetic features (e.g., color harmony), which are critical for quality evaluation. The LMM processes I natively, leveraging its pre-trained vision encoder to extract spatial quality features.

3.1.2 *Video modality*

Key frames sampling. For VQA, videos are processed to capture both spatial and temporal quality aspects. A video V is represented as a sequence of frames, defined as

$$V = \{F_1, F_2, \ldots, F_T\}, \tag{1}$$

where $F_t \in \mathbb{R}^{H\times W\times 3}$ denotes the t-th frame, with H and W representing the height and width of the frame, and 3 corresponding to the RGB channels. T is the total number of frames in the video. To address spatial quality, we

sample frames at a rate of 1 frame per second (FPS), resulting in a subset

$$V_s = \{F_{t_1}, F_{t_2}, \ldots, F_{t_k}\}, \tag{2}$$

where $k = \lfloor T/\text{fps} \rfloor$ and $\text{fps} = 1$. These sampled frames are directly fed into the LMM for spatial quality analysis, analogous to the approach used in IQA.

VQA projector. To capture temporal quality aspects, such as motion artifacts and frame continuity, we employ the SlowFast network [14] with frozen weights as the VQA quality projector. In its original formulation, SlowFast processes the full video V using dual pathways: a slow pathway for sparse frames and a fast pathway for dense frames. In our approach, we utilize only the fast pathway to extract temporal features, as it operates on a higher frame rate and is better suited to capture fine-grained temporal dynamics. To align with the 1 FPS sampling setting used for spatial quality analysis, we divide the video V into non-overlapping 1-s clips and extract features from each clip. Formally, the video $V = \{F_1, F_2, \ldots, F_T\}$ is segmented into a set of clips

$$C = \{C_1, C_2, \ldots, C_M\}, \tag{3}$$

where $C_m = \{F_{t_{(m-1)S+1}}, F_{t_{(m-1)S+2}}, \ldots, F_{t_{mS}}\}$ represents the m-th 1-s clip, S is the video's frame rate, and $M = \lfloor T/S \rfloor$ is the total number of 1-s clips. For each clip C_m, the fast pathway of SlowFast processes the consistent frame sequences and produces a temporal feature vector

$$F_{V_m} = \sigma_v(C_m), \tag{4}$$

where $\sigma_v(\cdot)$ denotes the fast pathway feature extraction function, and $F_{V_m} \in \mathbb{R}^{D_t}$ is the temporal feature vector for the m-th clip, with D_t as the feature dimension. The full set of temporal features for the video is then represented as

$$F_V = \{F_{V_1}, F_{V_2}, \ldots, F_{V_M}\}, \tag{5}$$

where F_V encapsulates the temporal quality information extracted at 1-s intervals. Finally, we employ a multilayer perceptron (MLP) to map F_V into video motion quality embeddings, which are subsequently fed into the LMM for perception and analysis. The process described above constitutes the complete VQA projector, which takes a video as input and produces video motion quality embeddings as output, which can be denoted as

$$\text{Emb}_V = \alpha(V), \tag{6}$$

where $\alpha(\cdot)$ indicates the VQA projector, V stands for the input video, and Emb_V are the output video motion quality embeddings, respectively.

3.1.3 *3D model modality*

Projections rendering. For 3DQA, 3D models (e.g., meshes, voxel grids, or point clouds) are preprocessed to align with the image-based input requirements of LMMs while preserving geometric quality information. Given a 3D model O, if it is not already in point cloud format (i.e., $O \neq P$, where $P \in \mathbb{R}^{N\times 3}$ represents a point cloud with N points and 3D coordinates), it is converted into a point cloud P through sampling or reconstruction techniques (e.g., uniform sampling from mesh surfaces or voxel-to-point conversion). The resulting point cloud P is then rendered from six fixed cubic viewpoints (front, back, left, right, top, and bottom) yielding a set of 2D RGB images

$$P_r = \{I_1, I_2, \ldots, I_6\}, \tag{7}$$

where each $I_i \in \mathbb{R}^{H\times W\times 3}$ has height H, width W, and 3 color channels. These rendered images are fed into the LMM to assess surface-level quality attributes, such as texture fidelity and color distortions.

3DQA projector. To capture 3D-specific quality attributes (e.g., structural integrity and geometric detail), we employ PointNet++ [15] as a quality feature projector with frozen weights. For quality-specific 3D models, which often contain a large number of points ($N \gg 2048$), directly processing the full point cloud P may be computationally inefficient. Instead, we sample patches of 2048 points based on the visibility from the projection viewpoints as the PointNet++ input. For each viewpoint v_i (where $i = 1, 2, \ldots, 6$), we identify the subset of visible points $P_{v_i} \subseteq P$ by determining which points are unobstructed from the camera's perspective. From P_{v_i}, we sample fixed-size patches of 2048 points using farthest point sampling to ensure uniform coverage of the visible geometry

$$P' = \{P'_1, P'_2, \ldots, P'_L\}, \tag{8}$$

where P' represents the sampled point cloud patches. Then we use PointNet++ to extract the local geometric features from each patch P'_l

$$F_{P'_l} = \sigma_p(P'_l), \tag{9}$$

where $\sigma_p(\cdot)$ is the PointNet++ feature extraction function. The full set of geometry features for the point cloud is then represented as

$$F_{P'} = \{F_{P'_1}, F_{P'_2}, \ldots, F_{P'_L}\}, \tag{10}$$

where $F_{P'}$ represents the set of the geometry features extracted from all the patches. Similarly, we employ an MLP layer to map $F_{P'}$ into point cloud geometry quality embeddings and the overall 3DQA projector can be denoted as

$$\mathrm{Emb}_P = \beta(P), \tag{11}$$

where $\beta(\cdot)$ indicates the VQA projector, P stands for the input point cloud, and Emb_P are the output point cloud geometry quality embeddings, respectively.

3.2 Model architecture

The X-QA framework builds upon a pre-trained LMM (mPLUG-Owl2-7B as default [79]) with a vision encoder E_v and a language decoder D_l. The architecture is extended with modality-specific quality projectors and a unified quality assessment head, as described below. For an input X (where X is I for images, V_s for the sampled video frames, or P_r for the rendering projections of 3D models), the vision encoder generates visual embeddings

$$\mathrm{Emb}_X = E_v(X), \tag{12}$$

where $\mathrm{Emb}_X \in \mathbb{R}^{D_v}$ and D_v is the embedding dimension. For videos and 3D models, modality-specific features (F_V and F_P, respectively) are projected into the same embedding space via the MLP layers $\alpha(\cdot)$ and $\beta(\cdot)$ as described previously

$$\mathrm{Emb}_V = \alpha(V), \quad \mathrm{Emb}_P = \beta(P), \tag{13}$$

where $\mathrm{Emb}_V, \mathrm{Emb}_P \in \mathbb{R}^{D_v}$. These embeddings are concatenated to form a fused representation

$$\mathrm{Emb}_f = \mathrm{Emb}_V \oplus \mathrm{Emb}_P, \tag{14}$$

where $\mathrm{Emb}_f \in \mathbb{R}^{2D_v}$ is the combined feature vector. If the input does not include video data (i.e., $V = \emptyset$), then $\mathrm{Emb}_V = \mathbf{0} \in \mathbb{R}^{D_v}$; similarly, if the input does not include a 3D model (i.e., $P = \emptyset$), then $\mathrm{Emb}_P = \mathbf{0} \in \mathbb{R}^{D_v}$. The visual embeddings Emb_X and fused embeddings Emb_f are fed into the language decoder D_l, guided by modality-specific prompt text T (detailed in Section 3.3)

$$Q_t = D_l(\mathrm{Emb}_X, \mathrm{Emb}_f, T), \tag{15}$$

where Q_t represents the predicted quality score token, which can be regarded as the probability distribution (denoted as $\mathcal{X}$) on all possible tokens of the language model. To evaluate quality levels, we perform a closed-set softmax over a predefined set of five quality scores $\{\gamma_i\}_{i=1}^5$, yielding probabilities η_{γ_i} for each level, where $\sum_{i=1}^5 \eta_{\gamma_i} = 1$. The probability for each γ_i is computed as

$$\eta_{\gamma_i} = \frac{\mathrm{e}^{\mathcal{X}_{\gamma_i}}}{\sum_{j=1}^5 \mathrm{e}^{\mathcal{X}_{\gamma_j}}}, \tag{16}$$

where $\mathcal{X}_{\gamma_i}$ is the logit corresponding to the quality score γ_i. The final predicted score of the LMM, denoted S_{LMM}, is calculated as a weighted sum of the quality scores $G(\gamma_i)$, where $G(\gamma_i) = i$ represents the numerical value of each level (e.g., $i = 1, 2, 3, 4, 5$)

$$S_{\mathrm{LMM}} = \sum_{i=1}^5 \eta_{\gamma_i} G(\gamma_i) = \sum_{i=1}^5 i \cdot \frac{\mathrm{e}^{\mathcal{X}_{\gamma_i}}}{\sum_{j=1}^5 \mathrm{e}^{\mathcal{X}_{\gamma_j}}}, \tag{17}$$

where S_{LMM} represents the final predicted quality score.

3.3 Mixed-modality training strategy

3.3.1 *Quality assessment datasets transformation*

To train the LMM on mixed-modality quality data, we first convert MOSs from existing quality assessment datasets into adjective-based question-answer pairs via equidistant interval partition [13]. Specifically, we uniformly divide the range between the highest score (M) and the lowest score (m) into five distinct intervals, and assign the scores in each interval as respective levels

$$L(s) = l_i \text{ if } m + \frac{i-1}{5} \times (M-m) < s \leqslant m + \frac{i}{5} \times (M-m), \tag{18}$$

where $\{l_i|_{i=1}^5\} = \{\text{bad}, \text{poor}, \text{fair}, \text{good}, \text{excellent}\}$ are the standard text rating levels as defined by ITU [80].

3.3.2 *System prompts*

Due to the inherent differences across modalities, using a fixed prompt template for all quality assessment tasks may confuse the model and lead to conflicting interpretations. To address this, we design modality-specific system prompts tailored to each input type (images, videos, and 3D models) to facilitate clearer differentiation and improve the model's learning process. Specifically, the system prompts are formulated as follows.

Prompt template for image.

–#User: `<image> <Emb_f>` How would you rate the quality of the image?

–#Assistant: the quality of the image is `<quality_token>`.

Prompt template for video.

–#User: `<frames> <Emb_f>` Based on the key frames and motion features, how would you rate the quality of the video?

–#Assistant: the quality of the video is `<quality_token>`.

Prompt template for 3D model.

–#User: `<projections> <Emb_f>` Based on the cubic projections and geometric features, how would you rate the quality of the 3D model?

–#Assistant: the quality of the 3D model is `<quality_token>`,

where `<image>`, `<frames>`, and `<projections>` represent the embedding placeholders for the corresponding image, video frames, and 3D model projections, respectively. `<Emb_f>` denotes the embedding placeholder mapped by the quality projector, and `<quality_token>` is the final predicted quality score token output by the LMM.

3.3.3 *Loss function*

The training minimizes the cross-entropy loss between predicted probabilities $Q = \{\eta_{\gamma_i}\}_{i=1}^5$ and ground-truth labels Q^*:

$$\mathcal{L} = -\sum_{i=1}^{5} Q_i^* \log(\eta_{\gamma_i}), \tag{19}$$

where Q^* is the one-hot label. The model is fine-tuned using supervised fine-tuning (SFT) on a mixed dataset with modality-specific prompts.

4 Experiment

4.1 Experimental setup

To comprehensively cover IQA, VQA, and 3DQA, we select 11 popular datasets across these domains for experimentation. Additionally, to ensure a fair and robust performance comparison, we evaluated our proposed X-QA framework against 23 mainstream methods spanning IQA, VQA, and 3DQA.

4.1.1 *Training datasets*

For training, we utilized the training subsets of five datasets: KonIQ, SPAQ, LSVQ, SJTU-PCQA, and WPC, as the foundation for X-QA. These datasets are chosen for their diversity and prominence in their respective domains, with no overlap between their training and test splits. Notably, while X-QA is designed to handle multiple modalities within a single model, the competing methods were trained solely within their specific domains due to their modality-specific limitations.

Table 1 Brief introduction of the employed quality assessment datasets.

Dataset	Year	Scale	Content	Task
Quality assessment datasets for images				
LIVEC [81]	2016	1162	Real-world images with authentic distortions from user devices	Quality assessment of authentically-distorted images
KonIQ [82]	2018	10073	Images from public multimedia dataset YFCC100m [83]	Quality assessment of in-the-wild images
KADID-10K [84]	2019	10125	24 different types of distorted images	Weakly-supervised image quality assessment
SPAQ [34]	2020	11125	Smartphone-captured images with diverse scenes	Quality assessment of smartphone images
AGIQA-3K [85]	2023	2982	AI-generated images from various generative models	Quality assessment of AI-generated images
Quality assessment datasets for videos				
KoNViD-1K [86]	2017	1200	Public-domain video sequences from YFCC100m [83]	Unified video quality assessment
LIVE-VQC [87]	2018	585	Videos of unique content, captured by users	Quality assessment of real-world videos
LSVQ [88]	2021	39075	Real-world distorted videos and video patches	Quality assessment of user-generated videos
Quality assessment datasets for 3D contents				
SJTU-PCQA [89]	2020	420	6 different types of distorted point clouds	Colorful point cloud quality assessment
WPC [90]	2021	740	3 different types of distorted point clouds	Colorful point cloud quality assessment
WPC2.0 [91]	2021	400	Video-based compressed point clouds	Compressed point cloud quality assessment

4.1.2 *Testing datasets*

To evaluate X-QA's performance, we used the test splits of the aforementioned datasets (KonIQ, SPAQ, LSVQ, SJTU-PCQA, WPC) and further included six additional datasets: LIVEC, KADID-10K, AGIQA-3K, KoNViD-1K, LIVE-VQC, and WPC2.0. These datasets, detailed in Table 1, are excluded from training to rigorously assess generalzaition performance.

4.1.3 *Quality assessment competitors*

We compared X-QA against 23 established methods, categorized by domain:

• IQA methods (8): BRISQUE [92], NIQE [93], NIMA [94], DBCNN [95], HyperIQA [96], MUSIQ [97], CLIP-IQA+ [98], LIQE [99];

• VQA methods (8): TLVQM [100], VSFA [101], VIDEVAL [102], PVQ [88], BVQA [103], DisCoVQA [104], SimpleVQA [105], FAST-VQA [106];

• 3DQA methods (7): IT-PCQA [107], ResSCNN [108], PQA-Net [90], 3D-NSS [10], GMS-3DQA [64], MM-PCQA [62], LMM-PCQA [12].

Performance results are reported with Spearman's rank correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) as metrics. For the main experiment, competing methods are evaluated in an intra-dataset setting (trained and tested within the same dataset), while X-QA is trained across all listed training datasets, demonstrating its versatility and superior performance across modalities.

4.2 Training details

The training of X-QA is conducted with 8 NVIDIA A800-80 GB GPU (requiring about 48 h in total). The vision transformer (ViT) initialization is derived from the pre-training stage, utilizing an updated CLIP-L/14 model [70]. Similarly, the large language model (LLM) initialization is based on LLaMA-2 [109]. The visual abstractor initialization combines the pre-training stage with mPLUG-Owl2. The image resolution is set to 448×448 pixels for both configurations. A batch size of 256 is employed, alongside a maximum learning rate (lr max) of 2e−5, which follows a cosine decay learning rate schedule. The learning rate warmup ratio is configured at 0.03, with no weight decay applied (set to 0). Gradient accumulation is performed over 16 steps, and the numerical precision is maintained at bfloat16. Training is conducted for a single epoch, with 250 warm-up steps. The optimizer used is AdamW, with optimizer sharding enabled. Additional configurations include activation checkpointing and a model parallelism level of 2, while pipeline parallelism is set to 1.

The SlowFast [14] network is fixed and initialized with the pre-trained weights on the Kinetics-400 [110] dataset while the PointNet++ network is fixed and initialized with the pre-trained weights on the ModelNet40 [111] dataset. This choice of weights fixing is based on two considerations: (1) SlowFast and PointNet++ are pre-trained on large-scale datasets (Kinetics-400, ModelNet40), allowing them to extract robust motion and geometric features without further tuning; (2) freezing these networks reduces training cost and memory usage, while avoiding overfitting on the limited quality-labeled video and 3D datasets.

4.3 Main performance

To evaluate the effectiveness of the proposed X-QA, we conduct extensive experiments across IQA, VQA, and 3DQA tasks, with results reported in Table 2. Unlike competing methods, which are trained and tested within their specific domains (intra-dataset setting), X-QA is trained on a mixed dataset of KonIQ, SPAQ, LSVQ, SJTU-PCQA, and WPC, enabling it to handle all three modalities simultaneously.

Table 2 (Color online) Performance comparison across IQA, VQA, and 3DQA datasets. The proposed X-QA is trained on all the listed datasets (KonIQ, SPAQ, LSVQ, SJTU-PCQA, and WPC). Other quality assessment competitors are tested with the intra-dataset setting (trained and tested within the same dataset). First in red, second in blue.

Index	Datasets	IQA				VQA				3DQA			
		KonIQ		SPAQ		$LSVQ_{test}$		$LSVQ_{1080P}$		SJTU-PCQA		WPC	
	Criteria	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
	IQA methods												
A	BRISQUE (TIP 2012) [92]	0.645	0.641	0.633	0.620			–			–		
B	NIQE (TIP 2013) [93]	0.430	0.451	0.413	0.426			–			–		
C	NIMA (TIP 2018) [94]	0.859	0.896	0.907	0.910			–			–		
D	DBCNN (TCSVT 2020) [95]	0.875	0.884	0.908	0.913			–			–		
E	HyperIQA (CVPR 2020) [96]	0.906	0.917	0.908	0.909			–			–		
F	MUSIQ (ICCV 2021) [97]	0.929	0.924	0.917	0.921			–			–		
G	CLIP-IQA+ (AAAI 2023) [98]	0.895	0.909	0.881	0.883			–			–		
H	LIQE (CVPR 2023) [99]	0.928	0.912	0.922	0.919			–			–		
	VQA methods												
I	TLVQM (TIP 2019) [100]			–		0.772	0.774	0.589	0.616		–		
J	VSFA (ACMMM 2019) [101]			–		0.801	0.796	0.675	0.704		–		
K	VIDEVAL (TIP 2021) [102]			–		0.794	0.783	0.545	0.554		–		
L	PVQ (CVPR 2021) [88]			–		0.827	0.828	0.711	0.739		–		
M	BVQA (TCSVT 2022) [103]			–		0.852	0.854	0.772	0.788		–		
N	DisCoVQA (TCSVT 2023) [104]			–		0.859	0.850	0.734	0.772		–		
O	SimpleVQA (ACMMM 2022) [105]			–		0.867	0.861	0.764	0.803		–		
P	FAST-VQA (ECCV 2022) [106]			–		0.876	0.877	0.779	0.814		–		
	3DQA methods												
Q	IT-PCQA (CVPR 2022) [107]			–			–			0.865	0.828	0.487	0.432
R	ResSCNN (TOMM 2022) [108]			–			–			0.860	0.810	0.722	0.714
S	PQA-net (TCSVT 2021) [90]			–			–			0.837	0.858	0.702	0.712
T	3D-NSS (TCSVT 2022) [10]			–			–			0.714	0.738	0.647	0.651
U	GMS-3DQA (TOMM 2024) [64]			–			–			0.910	0.917	0.830	0.833
V	MM-PCQA (IJCAI 2023) [62]			–			–			0.910	0.922	0.841	0.855
W	LMM-PCQA (ACMMM 2024) [12]			–			–			0.937	0.940	0.882	0.873
	Proposed versatile method												
X	X-QA (ours)	0.939	0.942	0.931	0.932	0.886	0.884	0.799	0.828	0.940	0.941	0.891	0.880

X-QA consistently outperforms domain-specific methods across all evaluated datasets, achieving the highest SRCC and PLCC scores in nearly all cases. For IQA, on KonIQ and SPAQ, X-QA achieves SRCC/PLCC of 0.939/0.942 and 0.931/0.932, respectively, surpassing top IQA methods like MUSIQ (0.929/0.924 on KonIQ) and LIQE (0.922/0.919 on SPAQ). In VQA, X-QA excels on $LSVQ_{test}$ and $LSVQ_{1080P}$ with SRCC/PLCC of 0.886/0.884 and 0.799/0.828, respectively, outperforming FAST-VQA (0.876/0.877 and 0.779/0.814). For 3DQA, X-QA achieves SRCC/PLCC of 0.940/0.941 on SJTU-PCQA and 0.891/0.880 on WPC, exceeding LMM-PCQA (0.937/0.940 and 0.882/0.873). These results are visually summarized in Figure 3, where radar charts compare the top-six methods across each modality. X-QA demonstrates superior performance across all axes, reflecting its ability to leverage cross-domain knowledge effectively. Notably, its unified training approach not only enhances accuracy but also eliminates the need for modality-specific models, reducing computational overhead. This versatility and high performance establish X-QA as a robust solution for multimedia quality assessment, aligning closely with the human visual system's integrated perception of quality across diverse modalities.

4.4 Generalization performance

To assess the generalization capability of X-QA, we evaluate its performance on six datasets excluded from training: KADID-10K, LIVEC, AGIQA-3K (IQA); KoNViD-1K, LIVE-VQC (VQA); and WPC2.0 (3DQA). These datasets, detailed in Table 1, test the framework's ability to handle unseen data across modalities. Results are reported in Table 3, with SRCC and PLCC as metrics. For comparison, only the top-performing are included.

X-QA demonstrates exceptional generalization, outperforming domain-specific methods across all tested datasets. In IQA, X-QA achieves SRCC/PLCC of 0.701/0.698 on KADID-10K (vs. CLIP-IQA+ at 0.661/0.663), 0.881/0.880 on LIVEC (vs. LIQE at 0.871/0.869), and 0.731/0.784 on AGIQA-3K (vs. LIQE at 0.718/0.765). For VQA, it scores 0.870/0.878 on KoNViD-1K (vs. SimpleVQA at 0.860/0.861) and 0.834/0.841 on LIVE-VQC (vs. FAST-VQA at 0.823/0.844). In 3DQA, X-QA reaches 0.845/0.856 on WPC2.0, surpassing LMM-PCQA (0.834/0.821). These results highlight X-QA's ability to adapt to diverse, unseen distortions and content types, from AI-generated images to compressed point clouds. The superior generalization stems from X-QA's unified training on a mixed dataset (KonIQ, SPAQ, LSVQ, SJTU-PCQA, WPC), which exposes it to a broad range of quality features across modalities. In contrast, competing methods, trained on narrower domain-specific datasets, struggle to extrapolate effectively to new data. This cross-modal robustness positions X-QA as a versatile solution for real-world multimedia applications, where quality assessment often involves unpredictable content and distortions.

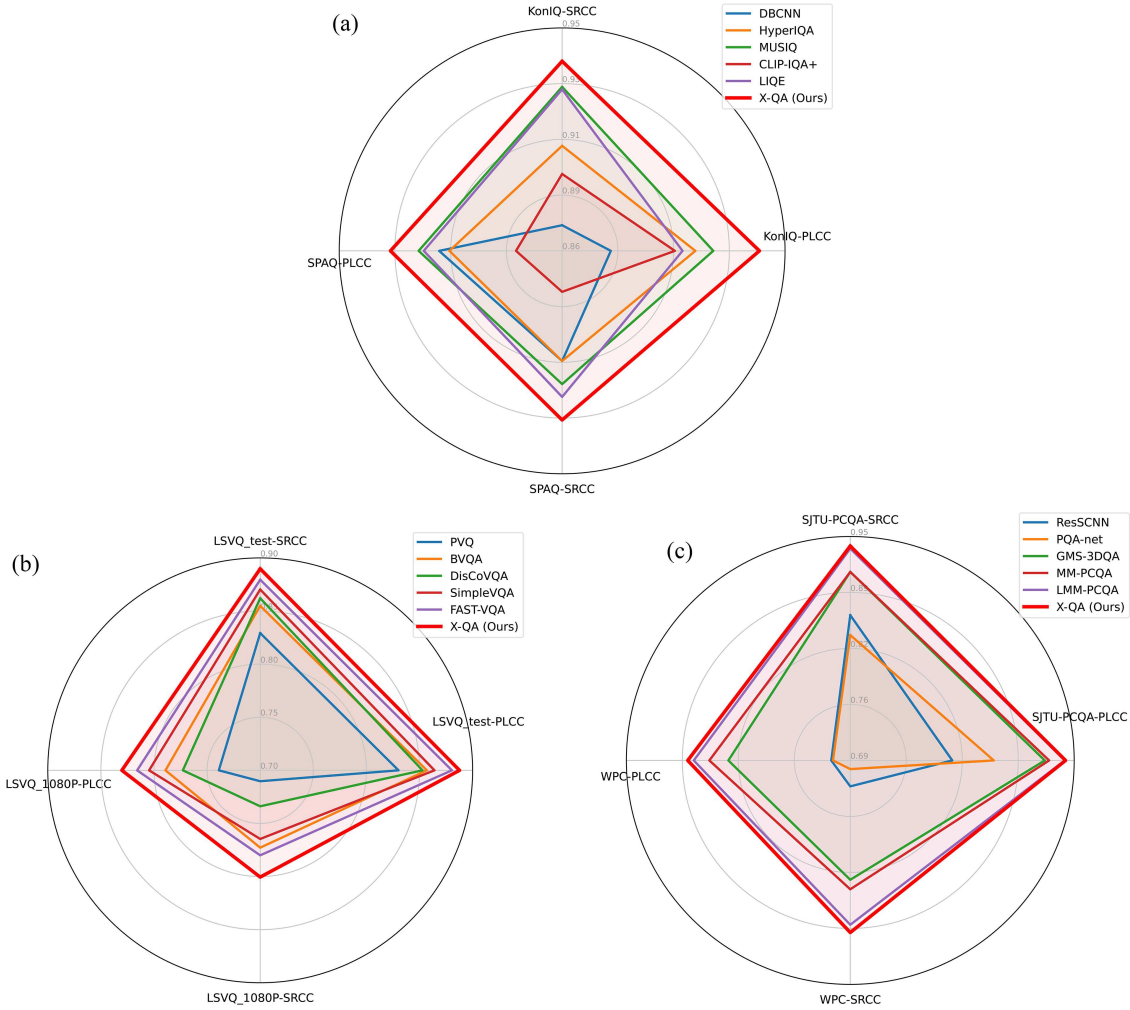

Figure 3 (Color online) Illustrations of performance comparison for top-six methods across (a) IQA, (b) VQA, and (c) 3DQA tasks.

Table 3 (Color online) Generalization performance comparison across IQA, VQA, and 3DQA datasets. The proposed X-QA is trained on KonIQ, SPAQ, LSVQ, SJTU-PCQA, and WPC datasets. The IQA competitors are trained on KonIQ and SPAQ datasets, the VQA methods are trained on LSVQ datasets, and the 3DQA methods are trained on SJTU-PCQA and WPC datasets.

Datasets	IQA						VQA				3DQA	
	KADID-10K		LIVEC		AGIQA-3K		KoNViD-1K		LIVE-VQC		WPC2.0	
Criteria	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
IQA methods												
MUSIQ (ICCV 2021) [97]	0.549	0.529	0.821	0.829	0.664	0.677			–		–	
CLIP-IQA+ (AAAI 2023) [98]	0.661	0.663	0.769	0.787	0.675	0.721			–		–	
LIQE (CVPR 2023) [99]	0.625	0.626	0.871	0.869	0.718	0.765			–		–	
VQA methods												
BVQA (TCSVT 2022) [103]			–				0.819	0.800	0.776	0.784	–	
SimpleVQA (ACMMM 2022) [105]			–				0.860	0.861	0.799	0.813	–	
FAST-VQA (ECCV 2022) [106]			–				0.859	0.855	0.823	0.844	–	
3DQA methods												
GMS-3DQA (TOMM 2024) [64]			–						–		0.781	0.771
MM-PCQA (IJCAI 2023) [62]			–						–		0.798	0.795
LMM-PCQA (ACMMM 2024) [12]			–						–		0.834	0.821
Proposed versatile method												
X-QA (ours)	0.701	0.698	0.881	0.880	0.731	0.784	0.870	0.878	0.834	0.841	0.845	0.856

4.5 Ablation study

We conduct an ablation study to evaluate the contributions of each component in the X-QA, including modalities (IQA, VQA, 3DQA), quality projectors (SlowFast for VQA, PointNet++ for 3DQA), and system prompts. Results are presented in Table 4.

(1) Removing any component results in a performance drop across all metrics, confirming the necessity of each module. For instance, excluding IQA reduces VQA from 0.843 to 0.828 (SRCC) and 3DQA from 0.916 to 0.901,

Table 4 Ablation study on the X-QA framework. Columns indicate the presence (✓) or absence (×) of modalities (IQA, VQA, 3DQA), quality projectors (Proj.), and system prompt (Prompt). Performance is reported as the average SRCC and PLCC across datasets: KonIQ and SPAQ (IQA), $\text{LSVQ}_{\text{test}}$ and $\text{LSVQ}_{1080\text{P}}$ (VQA), SJTU-PCQA and WPC (3DQA).

Ablation variant	Components					IQA Avg.		VQA Avg.		3DQA Avg.	
	IQA	VQA	3DQA	Proj.	Prompt	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
Full model (X-QA)	✓	✓	✓	✓	✓	0.935	0.937	0.843	0.856	0.916	0.911
w/o IQA	×	✓	✓	✓	✓	0.869	0.866	0.828	0.841	0.901	0.896
w/o VQA	✓	×	✓	✓	✓	0.924	0.926	0.768	0.767	0.905	0.900
w/o 3DQA	✓	✓	×	✓	✓	0.928	0.930	0.833	0.846	0.812	0.801
w/o projector	✓	✓	✓	×	✓	0.932	0.935	0.798	0.811	0.868	0.863
w/o system prompt	✓	✓	✓	✓	×	0.911	0.913	0.816	0.829	0.887	0.882

Table 5 Parameters and inference latencies (average time cost for each instance) of the X-QA framework for IQA, VQA, and 3DQA tasks.

Task format	Parameters		Latencies	
	Projector	LMM	Projector	LMM
IQA	0	7.3 B	0	101 ms
VQA	34.6 M	7.3 B	807 ms	350 ms
3DQA	1.7 M	7.3 B	1165 ms	276 ms

while eliminating projectors lowers VQA from 0.843 to 0.798 and 3DQA from 0.916 to 0.868. This consistent decline underscores that all components—IQA, VQA, 3DQA, projectors, and prompts—contribute to the framework's overall effectiveness. (2) Notably, removing a modality's data most significantly impacts its corresponding domain, highlighting each domain's critical role in its own performance. Without IQA, the IQA average drops sharply from 0.935/0.937 to 0.869/0.866 (SRCC/PLCC), a decrease of 0.066/0.071. Similarly, excluding VQA reduces VQA performance from 0.843/0.856 to 0.768/0.767 (drop of 0.075/0.089), and removing 3DQA lowers 3DQA from 0.916/0.911 to 0.812/0.801 (drop of 0.104/0.110). These substantial declines demonstrate that each modality's training data primarily enhances its own quality assessment, though cross-modal benefits remain evident in smaller drops elsewhere. (3) Eliminating projectors has a pronounced effect on VQA and 3DQA, where additional temporal and geometric information is lost, but minimally affects IQA. VQA performance falls from 0.843/0.856 to 0.798/0.811 (drop of 0.045/0.045), and 3DQA decreases from 0.916/0.911 to 0.868/0.863 (drop of 0.048/0.048), reflecting the loss of SlowFast's motion features and PointNet++'s structural features. In contrast, IQA only drops from 0.935/0.937 to 0.932/0.935 (drop of 0.003/0.002), as it relies solely on raw image inputs without projector enhancement, making it less sensitive to this ablation. (4) Removing the system prompt introduces confusion in the model, leading to performance degradation across all modalities due to conflicting interpretations of mixed-modality data. These consistent reductions emphasize the prompt's role in guiding the LMM to differentiate modalities, preventing quality assessment inconsistencies.

4.6 Computational cost analysis

Table 5 presents the parameter scale and average inference latency of the proposed X-QA framework across IQA, VQA, and 3DQA tasks, distinguishing between the costs incurred by the quality projectors and the LMM. Notably, the LMM remains consistent across all tasks, with 7.3 billion parameters and moderate latency, underscoring its scalability. In contrast, the task-specific projectors (absent in IQA but crucial for VQA and 3DQA) introduce only modest parameter overhead (34.6 M for SlowFast and 1.7 M for PointNet++), yet substantially improve performance as evidenced in the ablation studies. However, these projectors increase latency significantly: 807 ms for VQA and 1165 ms for 3DQA. This reflects the computational burden of extracting temporal and geometric features, particularly in 3DQA where point cloud processing is more complex.

4.7 Statistical test

To rigorously validate the performance of the proposed X-QA framework, we perform a statistical significance test in this section. Adopting the experimental methodology from [112], we assess the differences between predicted quality scores and subjective ratings for all pairwise model combinations. Results are presented in Figure 4, comprising heatmaps for IQA, VQA, and 3DQA modalities. In IQA, X-QA (index X) significantly outperforms six of the eight competing methods. For VQA, X-QA excels over seven of the eight methods. In 3DQA, X-QA surpasses six of the seven methods. These findings, grounded in the heatmap data, affirm that X-QA delivers consistently superior or equivalent performance across all three modalities compared to most state-of-the-art methods.

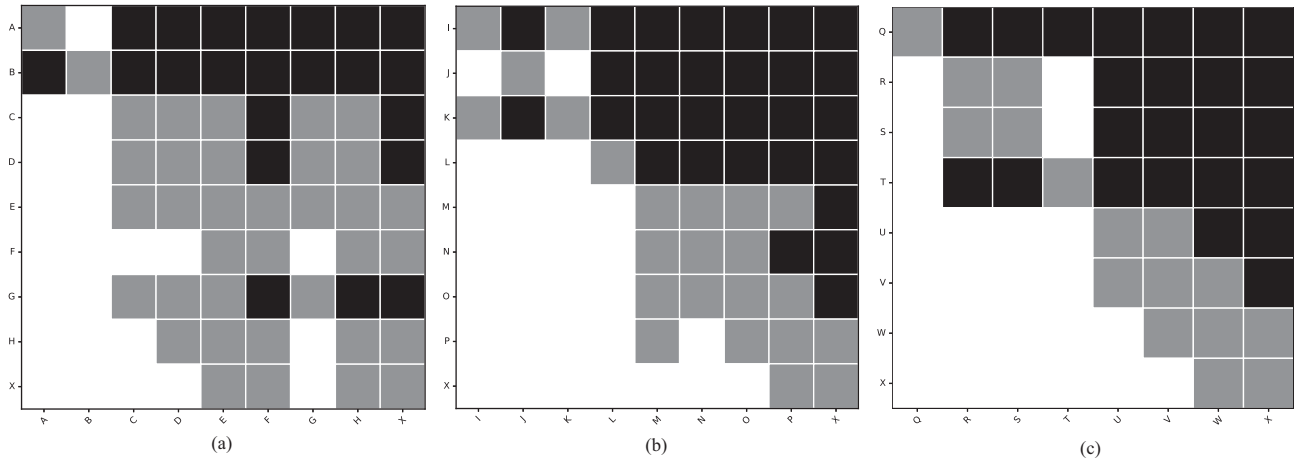

Figure 4 Statistical significance heatmaps for X-QA and competing methods across (a) IQA, (b) VQA, and (c) 3DQA methods (the index is indicated in Table 2). A black/white block means the row method is statistically worse/better than the column one. A gray block means the row method and the column method are statistically indistinguishable.

5 Conclusion

In this paper, we tackle the challenge of unifying visual quality assessment across images, videos, and 3D models, drawing inspiration from the human visual system's cohesive perception of multimedia content in visual communications. Conventional approaches, limited by modality-specific models, overlook cross-domain quality insights and impose substantial resource demands, hindering their applicability in integrated communication systems. To address these shortcomings, we introduce X-QA, the first framework to seamlessly integrate image quality assessment (IQA), video quality assessment (VQA), and 3D quality assessment (3DQA) into a single large multimodal model (LMM) tailored for visual communications. Leveraging modality-specific preprocessing (raw images, SlowFast for videos, PointNet++ for 3D models) and customized system prompts, X-QA adeptly manages diverse inputs and enhances mixed-modality training, ensuring robust quality evaluation across varied visual communication scenarios. Looking forward, X-QA paves the way for future advancements, such as expanding to additional modalities and optimizing prompt strategies for more precise assessments in complex communication contexts. Its versatility and efficiency establish it as a transformative tool for next-generation visual communication applications, where consistent and comprehensive quality evaluation across diverse content types is paramount for delivering superior user experiences.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 623B2073, 62225112, 62301310).

References

1 Zhou Y, Zhang Z, Sun W, et al. Perceptual quality assessment for point clouds: a survey. ZTE Commun, 2023, 21: 3–16
2 Zhang Z, Zhou Y, Li C, et al. Quality assessment in the era of large models: a survey. ACM Trans Multimedia Comput Commun Appl, 2025, 21: 1–31
3 Skorin-Kapov L, Varela M, Hoßfeld T, et al. A survey of emerging concepts and challenges for QoE management of multimedia services. ACM Trans Multimedia Comput Commun Appl, 2018, 14: 1–29
4 Min X K, Duan H Y, Sun W, et al. Perceptual video quality assessment: a survey. Sci China Inf Sci, 2024, 67: 211301
5 Wang R, Li W, Liu X, et al. Hazeclip: towards language guided real-world image dehazing. In: Proceedings of the IEEE International Conference on Acoustics, Hyderabad, 2025. 1–5
6 Zhang Z, Sun W, Min X, et al. A no-reference evaluation metric for low-light image enhancement. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2021. 1–6
7 Zhang C, Cheng W, Hirakawa K. Corrupted reference image quality assessment of denoised images. IEEE Trans Image Process, 2018, 28: 1732–1747
8 Zhang Z, Sun W, Wu W, et al. Perceptual quality assessment for fine-grained compressed images. J Visual Commun Image Represent, 2023, 90: 103696
9 Li Y, Wang S, Zhang X, et al. Quality assessment of end-to-end learned image compression: the benchmark and objective measure. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021. 4297–4305
10 Zhang Z, Sun W, Min X, et al. No-reference quality assessment for 3D colored point cloud and mesh models. IEEE Trans Circ Syst Video Technol, 2022, 32: 7618–7631

11 Liu H, Li C, Wu Q, et al. Visual instruction tuning. In: Proceedings of the Advances in Neural Information Processing Systems, 2023. 34892–34916

12 Zhang Z, Wu H, Zhou Y, et al. LMM-PCQA: assisting point cloud quality assessment with LMM. ArXiv:2404.18203

13 Wu H, Zhang Z, Zhang W, et al. Q-align: teaching LMMs for visual scoring via discrete text-defined levels. ArXiv:2312.17090

14 Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition. In: Proceedings of the ICCV, 2019. 6201–6210

15 Qi C R, Yi L, Su H, et al. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the Advances in Neural Information Processing Systems, 2017

16 Zhu H, Li L, Wu J, et al. MetaIQA: deep meta-learning for no-reference image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 14143–14152

17 Li L, Lin W, Wang X, et al. No-reference image blur assessment based on discrete orthogonal moments. IEEE Trans Cybern, 2015, 46: 39–50

18 Li L, Lin W, Wang X, et al. No-reference image blur assessment based on discrete orthogonal moments. IEEE Trans Cybern, 2015, 46: 39–50

19 Li L, Wu D, Wu J, et al. Image sharpness assessment by sparse representation. IEEE Trans Multimedia, 2016, 18: 1085–1097

20 Ma J, Wu J, Li L, et al. Blind image quality assessment with active inference. IEEE Trans Image Process, 2021, 30: 3650–3663

21 Chahine N, Conde M V, Carfora D, et al. Deep portrait quality assessment. a ntire 2024 challenge survey. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 6732–6744

22 Wan Z, Yang Q, Li Z, et al. Dual-stream perception-driven blind quality assessment for stereoscopic omnidirectional images. In: Proceedings of the 32nd ACM International Conference on Multimedia, 2024. 10431–10439

23 Wan Z, Yan X, Li Z, et al. No-reference stereoscopic omnidirectional image quality assessment via a binocular viewport hypergraph convolutional network. IEEE Trans Circ Syst Video Technol, 2025, 35: 7196–7209

24 Wan Z, Gu K, Zhao D. Reduced reference stereoscopic image quality assessment using sparse representation and natural scene statistics. IEEE Trans Multimedia, 2019, 22: 2024–2037

25 Zhang Z, Wang J, Guo Y, et al. Aibench: towards trustworthy evaluation under the 45° law. 2025. https://aiben.ch/

26 Wang Z, Bovik A C, Lu L. Why is image quality assessment so difficult? In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002

27 Moorthy A K, Bovik A C. Visual importance pooling for image quality assessment. IEEE J Sel Top Signal Process, 2009, 3: 193–201

28 Moorthy A K, Bovik A C. A two-step framework for constructing blind image quality indices. IEEE Signal Process Lett, 2010, 17: 513–516

29 De K, Masilamani V. Image sharpness measure for blurred images in frequency domain. Procedia Eng, 2013, 64: 149–158

30 Liu L, Hua Y, Zhao Q, et al. Blind image quality assessment by relative gradient statistics and Adaboosting neural network. Signal Process-Image Commun, 2016, 40: 1–15

31 Bianco S, Celona L, Napoletano P, et al. On the use of deep learning for blind image quality assessment. SIViP, 2018, 12: 355–362

32 Kim H G, Lim H T, Ro Y M. Deep virtual reality image quality assessment with human perception guider for omnidirectional image. IEEE Trans Circ Syst Video Technol, 2020, 30: 917–928

33 Sun W, Min X, Zhai G, et al. MC360IQA: a multi-channel CNN for blind 360-degree image quality assessment. IEEE J Sel Top Signal Process, 2019, 14: 64–77

34 Fang Y, Zhu H, Zeng Y, et al. Perceptual quality assessment of smartphone photography. In: Proceedings of the CVPR, 2020

35 Schlett T, Rathgeb C, Henniger O, et al. Face image quality assessment: a literature survey. ACM Comput Surv, 2022, 54: 1–49

36 Cheon M, Yoon S J, Kang B, et al. Perceptual image quality assessment with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 433–442

37 Zhang Z, Sun W, Min X, et al. A no-reference evaluation metric for low-light image enhancement. In: Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021. 1–6

38 Lu W, Sun W, Min X, et al. Deep neural network for blind visual quality assessment of 4K content. IEEE Trans Broadcast, 2022, 69: 406–421

39 Cao J, Wu W, Wang R, et al. No-reference image quality assessment by using convolutional neural networks via object detection. Int J Mach Learn Cyber, 2022, 13: 3543–3554

40 Li X, Yuan K, Pei Y, et al. Ntire 2024 challenge on short-form ugc video quality assessment: methods and results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 6415–6431

41 Conde M V, Zadtootaghaj S, Barman N, et al. Ais 2024 challenge on video quality assessment of user-generated content: methods and results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 5826–5837

42 Zhang Z, Lu W, Sun W, et al. Surveillance video quality assessment based on quality related retraining. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2022. 4278–4282

43 Wang Z, Lu L, Bovik A C. Video quality assessment based on structural distortion measurement. Signal Process-Image Commun, 2004, 19: 121–132

44 Seshadrinathan K, Bovik A C. A structural similarity metric for video based on motion models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. I–869

45 Wang Z, Li Q. Video quality assessment using a statistical model of human visual speed perception. J Opt Soc Am A, 2007, 24: B61

46 Zhai G, Cai J, Lin W, et al. Cross-dimensional perceptual quality assessment for low bit-rate videos. IEEE Trans Multimedia, 2008, 10:

1316–1324

47 Ninassi A, Le Meur O, Le Callet P, et al. Considering temporal variations of spatial visual distortions in video quality assessment. IEEE J Sel Top Signal Process, 2009, 3: 253–265

48 Vu P V, Vu C T, Chandler D M. A spatiotemporal most-apparent-distortion model for video quality assessment. In: Proceedings of the 18th IEEE International Conference on Image Processing, 2011. 2505–2508

49 Wu H, Chen C, Liao L, et al. Neighbourhood representative sampling for efficient end-to-end video quality assessment. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 15185–15202

50 Wu H, Chen C, Hou J, et al. Fast-VQA: efficient end-to-end video quality assessment with fragment sampling. In: Proceedings of the European Conference of Computer Vision (ECCV), 2022

51 Wu H, Zhang E, Liao L, et al. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proceedings of the ICCV, 2023

52 Zhou X, Liu X, Dong Y, et al. Light-VQA+: a video quality assessment model for exposure correction with vision-language guidance. ArXiv:2405.03333

53 Dong Y, Liu X, Gao Y, et al. Light-VQA: a multi-dimensional quality assessment model for low-light video enhancement. In: Proceedings of the 31st ACM International Conference on Multimedia, 2023. 1088–1097

54 Alexiou E, Ebrahimi T. Point cloud quality assessment metric based on angular similarity. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2018. 1–6

55 Su H, Duanmu Z, Liu W, et al. Perceptual quality assessment of 3D point clouds. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2019. 3182–3186

56 Diniz R, Freitas P G, Farias M C. Multi-distance point cloud quality assessment. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2020. 3443–3447

57 Zhang Y, Yang Q, Xu Y. Ms-graphsim: inferring point cloud quality via multiscale graph similarity. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021. 1230–1238

58 Zhang Z, Sun W, Min X, et al. A no-reference visual quality metric for 3D color meshes. In: Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2021. 1–6

59 Yang Q, Liu Y, Chen S, et al. No-reference point cloud quality assessment via domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 21179–21188

60 Zhang Z, Sun W, Zhu Y, et al. Evaluating point cloud from moving camera videos: a no-reference metric. IEEE Trans Multimedia, 2025, 27: 927–939

61 Fan Y, Zhang Z, Sun W, et al. A no-reference quality assessment metric for point cloud based on captured video sequences. In: Proceedings of the 24th International Workshop on Multimedia Signal Processing (MMSP), 2022. 1–5

62 Zhang Z, Sun W, Min X, et al. MM-PCQA: multi-modal learning for no-reference point cloud quality assessment. ArXiv:2209.00244

63 Zhang Z, Sun W, Zhou Y, et al. EEP-3DQA: efficient and effective projection-based 3D model quality assessment. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2023. 2483–2488

64 Zhang Z, Sun W, Wu H, et al. GMS-3DQA: projection-based grid mini-patch sampling for 3D model quality assessment. ACM Trans Multimedia Comput Commun Appl, 2024, 20: 1–19

65 Zhang Z, Zhou Y, Li C, et al. A reduced-reference quality assessment metric for textured mesh digital humans. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. 2965–2969

66 Chen S, Zhang Z, Zhou Y, et al. A no-reference quality assessment metric for dynamic 3D digital human. Displays, 2023, 80: 102540

67 OpenAI. GPT-4 technical report. ArXiv:2303.08774

68 Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. ArXiv:2210.11416

69 Touvron H, Lavril T, Izacard G, et al. Llama: open and efficient foundation language models. ArXiv:2302.13971

70 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, 2021

71 Li B, Zhang Y, Chen L, et al. Otter: a multi-modal model with in-context instruction tuning. ArXiv:2305.03726

72 Gao P, Han J, Zhang R, et al. Llama-adapter v2: parameter-efficient visual instruction model. ArXiv:2304.15010

73 Dai W, Li J, Li D, et al. Instructblip: towards general-purpose vision-language models with instruction tuning. In: Proceedings of the Advances in Neural Information Processing Systems, 2023

74 Zhang P, Dong X, Wang B, et al. Internlm-xcomposer: a vision-language large model for advanced text-image comprehension and composition. ArXiv:2309.15112

75 Awadalla A, Gao I, Gardner J, et al. Openflamingo: an open-source framework for training large autoregressive vision-language models. ArXiv:2308.01390

76 Li J, Li D, Savarese S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the International Conference on Machine Learning, 2023. 19730–19742

77 Liu H, Li C, Li Y, et al. Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023

78 Liu H, Li C, Li Y, et al. LlaVA-NeXT: improved reasoning, OCR, and world knowledge. 2024. https://llava-vl.github.io/blog/2024-01-30-llava-next/

79 Ye Q, Xu H, Ye J, et al. Mplug-owl2: revolutionizing multi-modal large language model with modality collaboration. ArXiv:2311.04257

80 ITU. Recommendation 500-10: methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT.500, 2000

81 Ghadiyaram D, Bovik A C. Massive online crowdsourced study of subjective and objective picture quality. IEEE Trans Image Process, 2016, 25: 372–387

82 Hosu V, Lin H, Sziranyi T, et al. KonIQ-10k: an ecologically valid database for deep learning of blind image quality assessment. IEEE Trans Image Process, 2020, 29: 4041–4056

83 Thomee B, Shamma D A, Friedland G, et al. Yfcc100m: the new data in multimedia research. Commun ACM, 2016, 59: 64–73

84 Lin H, Hosu V, Saupe D. Kadid-10k: a large-scale artificially distorted iqa database. In: Proceedings of the QoMEX, 2019. 1–3

85 Li C, Zhang Z, Wu H, et al. AGIQA-3K: an open database for AI-generated image quality assessment. IEEE Trans Circ Syst Video Technol, 2024, 34: 6833–6846

86 Hosu V, Hahn F, Jenadeleh M, et al. The konstanz natural video database (konvid-1k). In: Proceedings of the QoMEX, 2017. 1–6

87 Sinno Z, Bovik A C. Large-scale study of perceptual video quality. IEEE Trans Image Process, 2019, 28: 612–627

88 Ying Z, Mandal M, Ghadiyaram D, et al. Patch-vq: 'patching up' the video quality problem. In: Proceedings of the CVPR, 2021

89 Yang Q, Chen H, Ma Z, et al. Predicting the perceptual quality of point cloud: a 3D-to-2D projection-based exploration. IEEE Trans Multimedia, 2020, 23: 3877–3891

90 Liu Q, Su H, Duanmu Z, et al. Perceptual quality assessment of colored 3D point clouds. IEEE Trans Visual Comput Graphics, 2023, 29: 3642–3655

91 Liu Q, Yuan H, Hamzaoui R, et al. Reduced reference perceptual quality model with application to rate control for video-based point cloud compression. IEEE Trans Image Process, 2021, 30: 6623–6636

92 Mittal A, Moorthy A K, Bovik A C. No-reference image quality assessment in the spatial domain. IEEE Trans Image Process, 2012, 21: 4695–4708

93 Mittal A, Soundararajan R, Bovik A C. Making a "completely blind" image quality analyzer. IEEE Signal Process Lett, 2013, 20: 209–212

94 Talebi H, Milanfar P. NIMA: neural image assessment. IEEE Trans Image Process, 2018, 27: 3998–4011

95 Zhang W, Ma K, Yan J, et al. Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans Circ Syst Video Technol, 2020, 30: 36–47

96 Su S, Yan Q, Zhu Y, et al. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the CVPR, 2020

97 Ke J, Wang Q, Wang Y, et al. MUSIQ: multi-scale image quality transformer. In: Proceedings of the ICCV, 2021. 5148–5157

98 Wang J, Chan K C K, Loy C C. Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2022

99 Zhang W, Zhai G, Wei Y, et al. Blind image quality assessment via vision-language correspondence: a multitask learning perspective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023

100 Korhonen J. Two-level approach for no-reference consumer video quality assessment. IEEE Trans Image Process, 2019, 28: 5923–5938

101 Li D, Jiang T, Jiang M. Quality assessment of in-the-wild videos. In: Proceedings of the ACM MM, 2019

102 Tu Z, Wang Y, Birkbeck N, et al. UGC-VQA: benchmarking blind video quality assessment for user generated content. IEEE Trans Image Process, 2021, 30: 4449–4464

103 Li B, Zhang W, Tian M, et al. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. IEEE Trans Circ Syst Video Technol, 2022, 32: 5944–5958

104 Wu H, Chen C, Liao L, et al. DisCoVQA: temporal distortion-content transformers for video quality assessment. IEEE Trans Circ Syst Video Technol, 2023, 33: 4840–4854

105 Sun W, Min X, Lu W, et al. A deep learning based no-reference quality assessment model for ugc videos. ArXiv:2204.14047

106 Wu H, Chen C, Hou J, et al. Fast-VQA: efficient end-to-end video quality assessment with fragment sampling. In: Proceedings of the ECCV, 2022

107 Yang Q, Liu Y, Chen S, et al. No-reference point cloud quality assessment via domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022

108 Liu Y, Yang Q, Xu Y, et al. Point cloud quality assessment: dataset construction and learning-based no-reference metric. ACM Trans Multimedia Comput Commun Appl, 2023, 19: 1–26

109 Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. 2023. ArXiv:2307.09288

110 Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset. 2017. ArXiv:1705.06950

111 Wu Z, Song S, Khosla A, et al. 3D shapenets: a deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 1912–1920

112 Sheikh H R, Sabir M F, Bovik A C. A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans Image Process, 2006, 15: 3440–3451