FineVQ: Fine-Grained User Generated Content Video Quality Assessment

Huiyu Duan[1], Qiang Hu[1,*], Jiarui Wang[1], Liu Yang[1], Zitong Xu[1], Lu Liu[1], Xiongkuo Min[1],
Chunlei Cai[2], Tianxiao Ye[2], Xiaoyun Zhang[1], Guangtao Zhai[1]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Bilibili Inc., Shanghai, China
{huiyuduan, qiang.hu, wangjiarui, ylyl.yl, lettieliu, minxiongkuo, xiaoyun.zhang,
zhaiguangtao}@sjtu.edu.cn, {caichunlei, yetianxiao}@bilibili.com

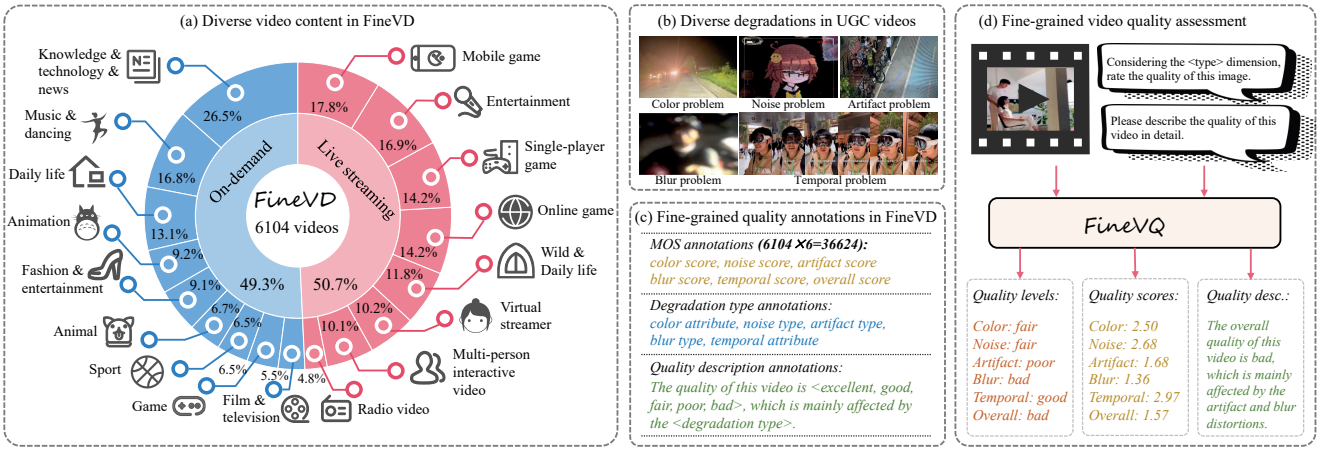


Figure 1. We present the fine-grained video quality assessment database and model, termed FineVD and FineVQ, respectively. UGC videos have diverse video content but suffer from various degradation issues as shown in (a) and (b), thus it is important to provide fine-grained quality labels for subsequent video processing and recommendation tasks in addition to only providing an overall quality score. To tackle the challenges, we construct FineVD, which includes fine-grained quality annotations for the UGC videos as shown in (c), and propose FineVQ, which has capabilities of quality rating, quality scoring, and quality attribution, as demonstrated in (d).

Abstract

*The rapid growth of user-generated content (UGC) videos has produced an urgent need for effective video quality assessment (VQA) algorithms to monitor video quality and guide optimization and recommendation procedures. However, current VQA models generally only give an overall rating for a UGC video, which lacks fine-grained labels for serving video processing and recommendation applications. To address the challenges and promote the development of UGC videos, we establish the first large-scale **Fine**-grained **V**ideo quality assessment **D**atabase, termed **FineVD**, which comprises 6104 UGC videos with fine-grained quality scores and descriptions across multiple dimensions. Based on this database, we propose a **Fine**-grained **V**ideo **Q**uality assessment (**FineVQ**) model to learn the fine-grained quality of UGC videos, with the capabilities of quality rating, quality scoring, and quality attribution. Extensive experimental results demonstrate that our proposed FineVQ can produce fine-grained video-quality results and achieve state-of-the-art performance on FineVD and other commonly used UGC-VQA datasets. Both FineVD and FineVQ are publicly available at: https://github.com/IntMeGroup/FineVQ.*

1. Introduction

Recent years have witnessed an explosion of user-generated content (UGC) videos [37, 76], thanks to the evolution of video processing technologies, and the popularity of social media platforms [53, 72]. UGC videos cover diverse video capturing and processing conditions [35], thus commonly suffering from various degradations such as noise, blur, shaking, *etc.*, which makes the visual quality of UGC videos vary greatly. Understanding and predicting the quality of UGC videos has long been an important but unsolved problem, which can be applied to capture devices and social

* Corresponding author.

media sites to monitor video quality and guide optimization and recommendation procedures [72]. However, different applications generally require quality labels considered from different dimensions, making a single overall quality rating insufficient for diverse downstream tasks [12, 13].

Many UGC video quality assessment (VQA) databases have been established [18, 53, 57, 63, 72] and numerous UGC VQA methods [2, 29, 49, 61, 62, 73, 78] have been proposed in the literature. Some UGC VQA databases have focused on the in-the-wild distortions [18, 57, 72], while several databases have considered the compression distortions in UGC videos [28, 58, 76, 79]. Recently, some works have also established quality assessment databases for short-form UGC (S-UGC) videos [35], which contain various special creation or generation modes such as special effects, numerous subtitles, *etc*. These existing databases generally only provide an overall mean quality score for a UGC video, which lacks fine-grained quality labels evaluated from multiple dimensions to facilitate subsequent diverse applications. Some studies have also provided fine-grained labels or descriptions for UGC videos [62–64, 74, 75]. However, the description of low-level attributes is coarse-grained, and in some applications, we may still need to use specific fine-grained quality scores.

To address the fine-grained VQA problems, we establish the first large-scale Fine-grained video quality assessment Database termed **FineVD**. Specifically, 6104 UGC videos are collected to cover a wide range of UGC scenarios, including on-demand and live-streaming applications, general and short-form video content, *etc.*, as shown in Figure 1(a). Based on the collected UGC videos, a professional team consisting of image-processing researchers is responsible for the quality labeling from six dimensions, including *color, noise, artifact, blur, temporal, and overall*, in a lab environment. As shown in Figure 1(c), FineVD includes *36624* mean opinion scores (MOSs) produced from over *800K* quality ratings, fine-grained degradation type labels, and quality description annotations, constituting a comprehensive quality assessment database for UGC videos.

Based on the FineVD benchmark, we further introduce the first Fine-grained Video Quality evaluator (**FineVQ**) to perform multi-dimensional fine-grained video quality assessment in a ***one-for-all*** manner. To achieve the comprehensive capabilities of quality rating, quality scoring, and quality depicting from multiple dimensions, FineVQ is built upon large multimodal models (LMMs) [6, 26, 71] and leverages instruction tuning [32] and low-rank adaptation (LoRA) [19] techniques. Specifically, our FineVQ adopts the vision encoder used in InternVL [6], *i.e.*, InternViT, as the spatial visual backbone, along with a motion feature extraction encoder [14] as the temporal visual backbone. A large language model, *i.e.*, InternLM [51], is used as the language backbone to facilitate the quality-related score regression and text generation. Moreover, to extract quality-related features and refine quality-specific attributes, we impose LoRA weights on both the spatial visual backbone and the large language backbone. Additionally, we also employ instruction tuning by inputting various question-answering pairs and training the feature projectors to enable task-specific quality awareness. Extensive experimental results demonstrate the fine-grained quality rating, scoring, depicting abilities, and the state-of-the-art performance of FineVQ. The main highlights of this work include:

- We construct the **FineVD**, the first large-scale UGC-VQA database that contains 6104 videos with corresponding over 800K subjective quality ratings and descriptions, annotated from multiple dimensions.
- We propose the **FineVQ**, which leverages the powerful visual representation, understanding, and description capabilities to perform quality rating, quality scoring, and quality depicting tasks.
- The instruction tuning and LoRA techniques are employed to enable the task-specific quality awareness in FineVQ.
- The extensive experimental results on FineVD and other commonly-used UGC VQA databases manifest the fine-grained VQA capabilities and state-of-the-art performance of the proposed FineVQ model.

2. Related Work

UGC-VQA Databases. Many VQA databases have been established in the literature to study the human perceptual quality characteristics of videos. The early databases typically focus only on synthetic distortions, using limited source videos and manually introduced degradation types [10, 41, 42, 47, 55]. In recent years, with the popularity of UGC, many UGC VQA databases have also been established considering authentic distortions in real-world or real applications. Some UGC VQA databases [16, 18, 42, 48, 57, 72] have been proposed focusing on authentic distortions, *i.e.*, in-capture or in-the-wild degradations. Several databases [17, 28, 58, 76, 79] have been presented considering both synthetic and authentic distortions. Moreover, considering the UGC videos in the aforementioned databases are sourced from general media platforms (*e.g.*, Youtube), a recent database, KVQ [35], also contributes a UGC VQA database specifically focused on short-form videos. In contrast, our FineVD includes both on-demand and live-streaming videos, as well as general UGC and short-form UGC formats. And fine-grained labels in FineVD can lead to more applications compared to an overall score.

UGC-VQA Models. With the establishment of numerous UGC VQA databases, a variety of UGC-VQA models have also been proposed in the literature. Traditional

Figure 2. An overview of the content and construction process of FineVD. (a) Example videos from our database, which contains both common UGC videos and short-form UGC videos. (b) The illustration of subjective data annotation methods, including both quality scoring and quality attribute labeling processes. (c) The quality-related question-answering pairs generated by GPT-4 and revised by human annotators.

methods [21, 22, 40, 44, 80] generally extract and regress handcrafted features to predict the video quality, which lack adaptability to complex UGC scenarios. With the development of deep learning, many deep neural network-based (DNN-based) VQA models have been presented and have achieved superior performance. A large number of VQA methods [2, 15, 24, 29, 49, 60–62, 73, 78] employ pre-trained DNN models to extract semantic features, and train quality regressors to predict quality scores. Some studies [3, 4, 33, 38, 67] have also explored using self-supervised learning strategies to train VQA models. However, these methods lack distortion understanding ability and can only predict an overall quality score for a UGC video.

Large Multimodal Models. The rapid evolution of large language models [1, 52, 59] and vision language pre-training techniques [25, 43] has explored the advancement of large multimodal models [6, 26, 31, 32, 34, 71, 81], which has superior multimodal joint understanding and interpretation abilities. Based on LMMs, some works have also explored the ability of the LMMs on low-level visual understanding [5, 64–66], which can also effectively predict and describe the quality levels. However, precise visual quality assessment, *i.e.,* MOS prediction, is still useful in real applications, and these models lack abilities to give fine-grained quality levels and scores from multi-dimensions.

3. The FineVD Database

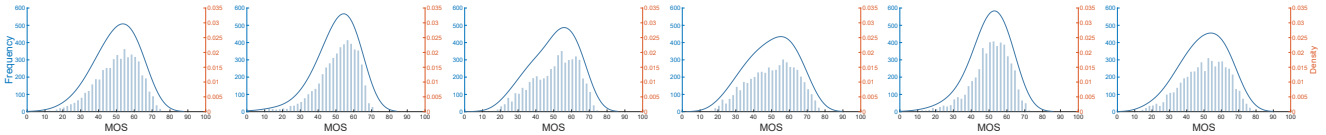
In this section, we introduce the proposed large-scale fine-grained video quality assessment database (**FineVD**). The database includes 6104 UGC videos, with corresponding 805728 human opinion ratings. Our FineVD exhibits the advantages of comprehensive UGC content and multi-dimensional labeling, which can significantly facilitate the advancement of UGC research.

3.1. Video Collection

The video collection process of FineVD follows two key principles: (1) covering a wide range of UGC scenarios and diverse distortions as comprehensively as possible, and (2) reflecting practical online statistics and applications of popular video platforms. To this end, we first collect a large-scale video dataset from the popular UGC video platform Bilibili [20]. Then we manually screen the collected videos to ensure the diversity of video scenarios and quality attributes. Finally, a total of 6104 UGC videos are included in our database. As shown in Figure 1, FineVD contains both on-demand videos and live-streaming videos, in which the on-demand UGC scenes consist of knowledge&technology&news, music&dancing, daily life, animation, fashion&enterainment, animal, sport, *etc.*, while live-streaming videos include scenes of mobile games, entertainment, single-player games, online games, wild, virtual streamer, *etc.* Figure 2 demonstrates that our FineVD contains both common UGC scenarios as aforementioned, and short-form UGC contents [35] such as virtual streamer, subtitle, special effect, multi-window interaction, *etc.*, which further manifest the content diversity of the established database. Moreover, Figure 1(b) also shows that FineVD covers a variety of degradations.

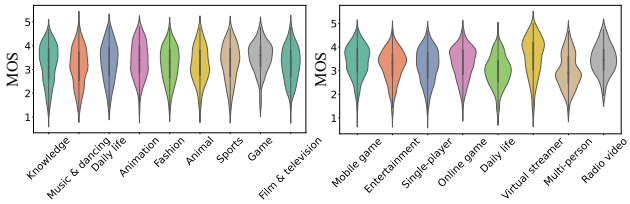
3.2. In-lab Subjective Study

Different from the crowd-sourced method adopted by some previous UGC VQA databases, such as LSVQ [72], our subjective study is carried out in a lab environment. A total of 22 professional annotators participate in the experiments. The videos are displayed on a DELL UltraSharp monitor with a resolution of 3840×2160, and the viewers are seated at a distance of about 2 feet from the monitor. Before the subjective experiments, we provide clear rating criteria with numerous examples to train the subjects. As



(a) Color (b) Noise (c) Artifact (d) Blur (e) Temporal (f) Overall

Figure 3. The MOS distribution of FineVD in terms of different perspectives, *i.e.*, color, noise, artifact, blur, temporal, and overall.



(a) On-demand (b) Live-streaming

Figure 4. The MOS distribution of *"overall score"* in terms of different video contents. (a) MOS distribution for on-demand videos. (b) MOS distribution for live-streaming videos.

shown in Figure 2(b), the subjects are introduced to evaluate the quality of a video from six perspectives, including *noise, artifact, blur, color, temporal, and overall*. For the quality rating task, subjects are asked to give their opinions on the videos in five categories from six perspectives. In particular, for *noise, artifact, and blur* distortions, the five-category levels are introduced as severe, strong, mild, slight, and undistorted, respectively, while for *color, temporal, and overall* perspectives, the five-category levels are explained as bad, poor, fair, good, and excellent, respectively. Besides the quality scores, identifying the distortion or quality attributes is also important for applications such as video processing, video recommendation, *etc.* Thus, we also collect the quality attribute labels as shown in Figure 2(b). For each distortion category, we provide three commonly encountered distortion types and two additional options "other" and "none". In particular, for the "other" option, the subjects can manually enter the quality attributes using a jump window.

Based on the collected quality scores and attribute labels, we further prepare language question-answering (QA) pairs for tuning LMMs [66, 74, 75]. Specifically, we first generate QA pairs automatically using an excellent large language model, *i.e.,* GPT-4 [1], and then we revise the generated texts manually to avoid ambiguous QA pairs. Finally, we introduce three tasks including quality-level prediction, quality score regression, and quality description, respectively, as shown in Figure 2(c). The generated QA pairs can significantly facilitate the subsequent training for our FineVQ model.

3.3. Subjective Data Analysis

We follow the suggestions given in [46] to conduct the outlier detection and subject rejection. As a result, no subject is rejected, and each image is rated by 22 valid subjects. Among all scores, about 3.16% of the total subjective evaluations are identified as outliers and are subsequently removed. For the remaining valid subjective ratings, we convert these raw ratings into Z-scores [11] as follows:

$$z_{ij} = \frac{m_{ij} - \mu_i}{\sigma_i}, \tag{1}$$

where m_{ij} is the raw rating given by the i-th subject to the j-th image, μ_i is the mean rating given by subject i, σ_i is the standard deviation. Then these scores are linearly scaled to the range $[0, 100]$ and obtain the rescaled z-scores z'_{ij}, which are averaged over subjects to obtain the final mean opinion scores (MOSs):

$$\text{MOS}_j = \frac{1}{N}\sum_{i=1}^{N} z'_{ij}, \tag{2}$$

where N is the total number of subjects.

Based on the MOSs, we further analyze the collected data. Specifically, we first visualize the MOS distribution for six evaluation dimensions, respectively, as shown in Figure 3. It can be observed that all MOS distributions of six dimensions cover a wide range. Moreover, although the distributions from different evaluation perspectives are generally similar, there are distinct differences in the details. To investigate the MOS distribution differences in terms of different content and applications, we further visualize the MOS distribution for 9 on-demand categories and 8 live-streaming categories in Figure 4, respectively. We can observe that the MOS distributions of different contents are generally similar, except for the game content in on-demand videos and virtual streamer and multi-person interaction categories for live-streaming videos. This manifests that these video contents are unusual compared to other UGC videos. Comparing Figure 4(a)&(b), we can also observe that the MOS distributions are significantly different for on-demand and live-streaming videos. Additionally, the extensive coverage of all video categories also demonstrates the abundance of our database.

4. The FineVQ Approach

In this section, we introduce our ***one-for-all*** video quality assessment method, **FineVQ**, towards giving quality levels, predicting quality scores, and depicting quality attributes from multiple perspectives using one model.

4.1. Overall Architecture

The overall framework of FineVQ is depicted in Figure 5, which takes both UGC videos and user prompts as input,

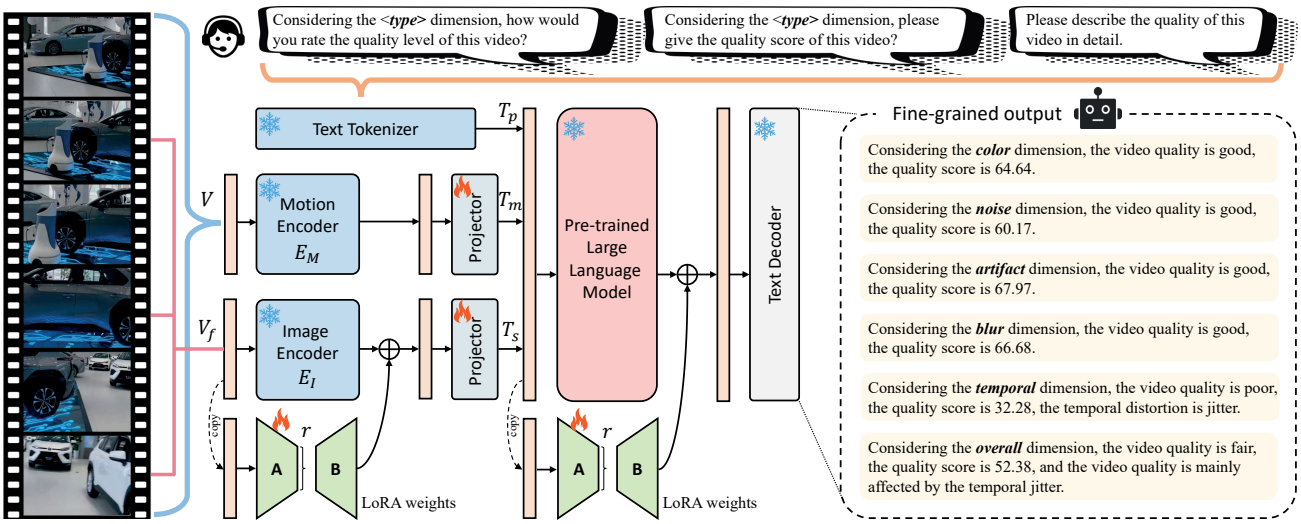


Figure 5. An overview of our proposed FineVQ model. Our model consists of three feature encoders, including an image feature extractor for extracting spatial features from sparse video frames, a motion feature extractor for extracting motion features from the entire video, and a text encoder for extracting aligned text features from prompts. The extracted features are then aligned through projectors and fed into a pre-trained LLM to generate the output results. LoRA weights are introduced to the pre-trained image encoder and the large language model to adapt the models to the quality assessment task.

and outputs quality-related answers according to both questions and videos. FineVQ begins with extracting visual features and text features from UGC videos and user prompts respectively. For the visual feature extraction part, two vision backbones are introduced. An image encoder is used to extract spatial features from selected video frames, and a motion encoder is applied to extract motion features from the entire video. Then two projectors P_I and P_M are used to project the extracted features into language space and output tokens T_s and T_m. For text feature extraction, a text tokenizer is utilized to encode the user prompt into tokens T_p. The extracted three feature tokens T_s, T_m, and T_p are concatenated and input into a pre-trained large language model to generate the final outputs.

4.2. Model Design

Visual Encoding. The visual encoding part involves two visual feature extractors, *i.e.,* an image encoder and a motion encoder. Specifically, for an input UGC video V, we uniformly collect 8 frames V_f from the video and then extract the features from these frame images using the image encoder. The image encoder E_I is built on a pre-trained vision transformer (ViT), *i.e.*, InternViT [6], which is pre-trained on the LAION-en dataset [45] using text-image contrastive learning. To align the extracted features with the input space of the large language model, a projector P_I with two multi-layer perceptron (MLP) layers is applied. The process can be formulated as:

$$T_s = P_I(E_I(V_f)), \tag{3}$$

where T_s is the mapped frame feature tokens. Considering the sparse frame can not include meticulous temporal features, a motion feature extractor E_M is also introduced to extract slow-fast motion features from a video [14]. Similarly, a projector P_M with two MLP layers is also introduced to ensure the feature dimension consistency and semantic feature alignment between the motion features and LLM input. The process can be formulated as:

$$T_m = P_M(E_M(V)), \tag{4}$$

where T_m is the mapped moton feature tokens.

Feature Fusion via the LLM. For a given prompt P, it is first encoded into the text tokens T_p via a text tokenizer. Then the text tokens T_p are concatenated with the well-aligned visual tokens T_s and T_m as the input to the LLM. In particular, InternLM-8B [51] is used as the pre-trained LLM to combine visual tokens and text tokens to perform multi-modal learning. Finally, the output features of the LLM are decoded with a text decoder and then projected to the prompt-corresponded quality space.

4.3. Instruction Tuning and LoRA Adaptation

Instruction Tuning. It is of great significance to achieve **one-for-all** VQA, which is conducive to multi-dimensional assessment. However, previous VQA models generally need to train multiple weights for different dimensions respectively. Benefiting from the generalization ability of LLMs, many recent works have verified the effectiveness of using the instruction tuning strategy for one-for-all tasks [31, 32]. As shown in Figure 5, our input prompts contain various question types from multiple dimensions. For different instructions, FineVQ trains visual feature projectors to tune the LLM to align textual and visual semantics for

Table 1. Performance of state-of-the-art models and the proposed FineVQ on our established FineVD database in terms of the *quality scoring task*. ♠, ♢, and ♡ denote the traditional IQA methods, traditional VQA methods, and DNN-based VQA metrics, respectively. It should be noted that the DNN-based models are trained with separate weights for different dimensions, while FineVQ complies with a ***one-for-all*** fashion. The best results are highlighted in **red**, and the second-best results are highlighted in **blue**.

Dimension	Color			Noise			Artifact			Blur			Temporal			Overall		
Methods / Metrics	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑
♠NIQE [39]	0.3273	0.2246	0.2368	0.2682	0.1873	0.1417	0.3006	0.2106	0.1649	0.3236	0.2259	0.1870	0.2777	0.1927	0.1630	0.3192	0.2222	0.2019
♠QAC [69]	0.1160	0.0852	0.0789	0.0072	0.0559	0.0067	0.0549	0.0154	0.0393	0.0862	0.0146	0.0600	0.0410	0.0103	0.0280	0.0739	0.0180	0.0512
♠HOSA [68]	0.3593	0.2449	0.3233	0.2937	0.2440	0.2385	0.3457	0.2419	0.3067	0.3878	0.2719	0.3451	0.2983	0.2078	0.2429	0.3627	0.2525	0.3236
♢TLVQM [22]	0.6346	0.4550	0.6417	0.6357	0.4543	0.5909	0.6608	0.4756	0.6526	0.6639	0.4768	0.6588	0.6816	0.4968	0.6996	0.6537	0.4698	0.6546
♢VIDEVAL [53]	0.6922	0.5034	0.6943	0.6912	0.5033	0.6514	0.7440	0.5500	0.7370	0.7610	0.5649	0.7637	0.7174	0.5305	0.7123	0.7310	0.5373	0.7307
♢RAPIQUE [54]	0.6203	0.4470	0.6330	0.6048	0.4284	0.5607	0.6258	0.4484	0.6276	0.6572	0.4778	0.6717	0.5554	0.3945	0.5429	0.6379	0.4597	0.6501
♡VSFA [24]	0.7617	0.5673	0.7837	0.7635	0.5707	0.7278	0.8006	0.6081	0.8170	0.7772	0.5858	0.8001	0.7276	0.5355	0.7018	0.7730	0.5825	0.7929
♡GSTVQA [2]	0.7747	0.5831	0.7761	0.7883	0.5974	0.7448	0.8121	0.6200	0.8187	0.8101	0.6185	0.8202	0.7533	0.5593	0.7093	0.7834	0.5906	0.7825
♡SimpleVQA [49]	0.8086	0.6097	0.8058	0.8070	0.6133	0.7634	**0.8465**	**0.6550**	**0.8487**	**0.8466**	**0.6558**	**0.8519**	**0.7746**	**0.5744**	0.7417	0.8311	0.6401	0.8358
♡FAST-VQA [60]	0.8017	0.6078	0.8183	**0.8093**	**0.6209**	**0.7758**	0.8176	0.6295	0.8328	0.8352	0.6500	0.8513	0.7560	0.5609	0.7393	0.8348	0.6476	**0.8474**
♡DOVER [62]	**0.8244**	**0.6311**	**0.8311**	0.8018	0.6055	0.7424	0.8265	0.6338	0.8289	0.8404	0.6504	0.8355	0.7664	0.5700	**0.7569**	**0.8422**	**0.6517**	0.8393
FineVQ (Ours)	**0.8495**	**0.6665**	**0.8527**	**0.8444**	**0.6613**	**0.7986**	**0.8852**	**0.7111**	**0.8921**	**0.8711**	**0.6957**	**0.8833**	**0.8085**	**0.6171**	**0.7597**	**0.8834**	**0.7118**	**0.8891**

Table 2. Performance of state-of-the-art LMMs and the proposed FineVQ on our established FineVD database in terms of the *quality attribute prediction task*. The *"yes-or-no"* type represents the judgment on whether the corresponding dimension is degraded. The *"which"* type indicates which distortion exists or has the most impact on the quality of the video.

Question Type	*Yes-or-no*					*Which*	
Model / Attribute	Color	Noise	Artifact	Blur	Temporal	Exist	Most
VideoLLaMA2 (*7B*) [7]	57.87%	63.98%	51.87%	61.91%	**86.42%**	25.59%	28.44%
Video-LLaVA (*7B*) [30]	26.48%	27.26%	43.90%	64.67%	13.09%	16.73%	36.71%
VideoChat2 (*7B*) [27]	18.25%	18.80%	24.26%	31.15%	15.19%	10.93%	13.99%
Video-ChatGPT (*7B*) [36]	28.35%	27.66%	44.49%	62.89%	18.60%	20.37%	24.11%
InternVL2 (*8B*) [6]	58.46%	63.58%	50.69%	54.33%	70.28%	**28.25%**	**43.21%**
MiniCPM-V (*8B*) [70]	70.47%	**65.85%**	**52.66%**	46.26%	71.95%	25.98%	30.81%
Qwen2-VL (*7B*) [56]	**71.16%**	44.78%	42.91%	41.14%	54.33%	23.33%	27.85%
LLaVA-NeXT (*7B*) [23]	26.87%	27.26%	43.90%	**64.76%**	13.09%	17.81%	37.89%
FineVQ (Ours)	**73.52%**	**72.74%**	**51.87%**	**64.76%**	**86.91%**	**91.93%**	**65.06%**

joint reasoning. As a result, FineVQ can evaluate the video quality from multiple dimensions using one model weight.

LoRA Adaptation. Fine-tuning LLMs is generally resource consuming but can lead to better performance. To further improve the performance of FineVQ, we adopt the LoRA technique [19] to fine-tune the model. Specifically, we employ LoRA to the image encoder E_I and the LLM. LoRA models the changes $\Delta\mathbf{W} \in \mathbb{R}^{n\times m}$ for each layer's weights $\mathbf{W} \in \mathbb{R}^{n\times m}$ as $\Delta\mathbf{W} = \mathbf{AB}$, where $\mathbf{A} \in \mathbb{R}^{n\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times m}$. The rank is constrained by $r \ll \{n, m\}$ to achieve parameter efficiency. Given the original output $\mathbf{h} = \mathbf{Wx}$, the forward pass of LoRA is:

$$\mathbf{h} = \mathbf{Wx} + \Delta\mathbf{Wx} = (\mathbf{W} + \mathbf{AB})\mathbf{x}. \tag{5}$$

With the LoRA, FineVQ can effectively adapt to the fine-grained VQA task.

5. Experimental Evaluation

5.1. Implementation Details

For the image encoder, we use a pre-trained InternViT [6] as the backbone and uniformly extract 8 frames from the video as the input. The frame images are resized to 448×448 and then input. For the motion extractor, we use the slow-fast [14] as the backbone and input the entire video frames. The video frames are also resized to 448×448 as the input. The LLM used in FineVQ is InternLM-8B [51], and the input token channel dimension is 4096. The image encoder, motion encoder, and LLM are all frozen during training. The rank parameter of LoRA [19] in FineVQ is set to 16. The model is trained on 4 NVIDIA RTX A6000 GPUs and flash-attn [8, 9] is used to save the GPU memory. We use AdamW Optimizer to train the network for 10 epochs with a batch size of 8. The initial learning rate is $2e^{-4}$ and gradually reduces to $1e^{-6}$ with cosine annealing.

5.2. Evaluation on the FineVD

Experimental Settings. We first follow the previous VQA research settings [35, 72] and split the UGC videos into training, validation, and test sets with a ratio of 4:1:1. A total of 11 state-of-the-art VQA models are selected as benchmark methods in our FineVD, which can be categories into three groups: (1) traditional image quality assessment (IQA) methods, including NIQE [39], QAC [69], HOSA [68]; (2) traditional video quality assessment (VQA) methods, including TLVQM [22], VIDEVAL [53], RAPIQUE [54]; (3) DNN-based VQA methods, including VSFA [24], GSTVQA [2], SimpleVQA [49], FAST-VQA [60], DOVER [62]. The DNN-based VQA models are trained using their official codes with the default settings. We train these models separately for different dimensions in FineVD. Three evaluation metrics, including Spearman rank-order correlation coefficient (SRCC), Kendall rank-order correlation coefficient (KRCC), and Pearson linear correlation coefficient (PLCC), are adopted to measure the performance of VQA models.

For the quality attribute prediction task, we use the same dataset split setting as mentioned above. A total of 7 large multimodal models that accept video as input are selected as benchmark methods, including VideoLLaMA2 [7], Video-LLaVA [30], VideoChat2 [27], Video-ChatGPT [36], InternVL2 [6], MiniCPM-V [70], Qwen2-VL [56], LLaVA-NeXT [23]. We set two types of questions including *"yes-or-no"* and *"which"* to test the ability of LMMs and our FineVQ on quality-related low-level vision tasks [64].

Table 3. Performance comparison between state-of-the-art VQA methods and the proposed FineVQ on six UGC VQA databases. The "N/A" means missing results in the original paper. The best results and the second-best results are highlighted in **red** and **blue**, respectively.

Method	LIVE-YT-Gaming [77]		KoNViD-1k [18]		YouTube-UGC [57]		LIVE-VQC [48]		$LSVQ_{test}$ [72]		$LSVQ_{1080p}$ [72]	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
VIQE [80]	N/A	N/A	0.628	0.638	0.513	0.476	0.659	0.694	N/A	N/A	N/A	N/A
TLVQM [22]	0.748	0.756	0.773	0.768	0.669	0.659	0.798	0.802	0.772	0.774	0.589	0.616
RAPIQUE [54]	0.771	0.815	0.803	0.817	0.759	0.768	0.754	0.786	N/A	N/A	N/A	N/A
VIDEVAL [53]	0.807	0.812	0.773	0.768	0.669	0.659	0.752	0.751	0.795	0.783	0.545	0.554
VSFA [24]	0.776	0.801	0.773	0.775	0.724	0.743	0.773	0.795	0.801	0.796	0.675	0.704
GSTVQA [2]	N/A	N/A	0.814	0.825	N/A	N/A	0.788	0.796	N/A	N/A	N/A	N/A
PVQ [72]	N/A	N/A	0.791	0.786	N/A	N/A	0.827	0.837	0.827	0.828	0.711	0.739
SimpleVQA [49]	0.861	0.866	0.856	0.860	0.847	0.856	N/A	N/A	0.867	0.861	0.764	0.803
FastVQA [60]	**0.869**	0.880	0.891	0.892	0.855	0.852	0.849	0.862	0.876	0.877	0.779	0.814
Dover [62]	0.852	0.868	0.908	0.910	0.841	0.851	0.844	0.875	0.877	0.878	0.778	0.812
MinimalisticVQA [50]	0.857	**0.888**	0.889	0.890	0.890	0.891	0.842	0.854	0.881	0.879	0.781	0.820
KSVQE [35]	N/A	N/A	**0.922**	**0.921**	**0.900**	**0.912**	**0.861**	**0.883**	**0.886**	**0.888**	**0.790**	**0.823**
FineVQ (Ours)	**0.912**	**0.926**	**0.915**	**0.910**	**0.910**	**0.914**	**0.895**	**0.895**	**0.900**	**0.900**	**0.828**	**0.857**

Table 4. The cross-dataset evaluation results. The models are trained on other datasets and tested on our FineVD.

Train on:	KoNViD-1k [18]		YouTtbe-UGC [57]		LIVE-VQC [48]	
Test on: FineVD	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
SimpleVQA [49]	0.5947	0.6330	0.6050	0.6265	0.4631	0.4753
FastVQA [60]	0.5155	0.5155	0.5990	0.6454	0.5730	0.5883
FineVQ (Ours)	**0.6475**	**0.7020**	**0.6261**	**0.6520**	**0.6125**	**0.6550**

Table 5. The cross-dataset evaluation results. The models are trained on our FineVD and tested on other datasets.

Test on:	KoNViD-1k [18]		YouTtbe-UGC [57]		LIVE-VQC [48]	
Train on: FineVD	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
SimpleVQA [49]	0.4818	0.3402	0.7421	0.7231	0.5922	0.6137
FastVQA [60]	0.6793	0.6490	0.5607	0.5651	0.6767	0.6817
FineVQ (Ours)	**0.7171**	**0.7312**	**0.7797**	**0.7880**	**0.7417**	**0.7620**

Specifically, for "yes-or-no" questions, we query the LMMs with the prompts of *"Is these any <dimension> distortion in this video?"*, while for "which" questions, we infer the LMMs with the prompts of *"Which distortion exists in this video?"* and *"Which distortion has the most impact on the quality of this video?"* We adopt prediction accuracy as the metric for evaluating the performance of the quality attribute prediction task.

Results on The Quality Scoring Task. We first evaluate the performance of our FineVQ and state-of-the-art VQA models on the established FineVD dataset, as shown in Table 1. We first observe that the traditional IQA models typically perform poorly on FineVD, especially for the temporal dimension. Traditional VQA models can generate meaningful quality scores, but are still far from satisfactory. DNN-based models generally achieve better results than traditional VQA and IQA methods, and our FineVQ consistently achieves the best performance in terms of all dimensions and all metrics. By comparing the performance of FineVQ across different dimensions, it can be observed that the proposed model performs relatively worse on the temporal dimension compared to other dimensions. This indicates that although we have incorporated the motion features into FineVQ, the temporal features are still insufficient and can be further improved.

Results on The Quality Attribute Prediction Task. Table 2 shows the results of the state-of-the-art large language model and the proposed FineVQ on the constructed FineVD dataset in terms of the quality attribute prediction task. For *"yes-or-no"* question, the accuracy of random selection is 50%. It can be observed that the performance of some LMMs is lower than the random selection, which manifests that these LMMs have poor quality-related low-level inference ability. Among the five specific dimensions, almost all LMMs have similar performance for the "artifact" dimension, which manifests that this dimension is difficult for all LMMs. Moreover, different LMMs have different advantages, *e.g.*, Qwen2-VL[56] has good performance on the color distortion judgment, while having relatively poor ability to judge the blur distortion; LLaVA-NeXT [23] achieves the best performance in judging the color distortion, but its ability to judge the color distortion is relatively poor. For *"which"* question, the accuracy of random selection is 25%. Although some LMMs, such as VideoLLaMA2 [7] and MiniCPM-V [70] have relatively good performance for *"yes-or-no"* questions, we observe that almost all LMMs perform poorly in selecting the distortion type and judging the most influential distortion. Finally, our FineVQ achieves the best performance for almost all tasks compared with state-of-the-art LMMs, which demonstrates the superiority of our model.

5.3. Evaluation on Existing UGC-VQA Databases

We further test the performance of the proposed FineVQ on other six VQA benchmark datasets, including LIVE-YT-Gaming [77], KoNViD-1k [18], YouTube-UGC [57], LIVE-VQC [48], LSVQ_test [72], LSVQ_1080p [72]. We follow the settings used in previous studies to conduct the experiments [35, 49, 60, 62]. Specifically, for LSVQ [72], we follow the public training/test split setting to validate our method, which is trained on the training set of LSVQ and tested on the $LSVQ_{test}$ and $LSVQ_{1080p}$ test sets. For the rest databases, we follow the previous principles, which split the dataset into training and test sets with a ratio of 4:1 and conduct multi-round cross-validation experiments.

Table 3 demonstrates the results of our proposed FineVQ and other state-of-the-art VQA methods on the six VQA

Table 6. Ablation study of FineVQ on three datasets, including our FineVD, LIVE VQC [48], and YouTube-UGC [57].

Strategy					FineVD			LIVE VQC [48]			YouTube-UGC [57]		
spatial	motion	LoRA$_{r=8}$ (vision)	LoRA$_{r=8}$ (llm)	LoRA$_{r=16}$	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
✔					0.8210	0.8434	0.6356	0.7832	0.8206	0.6065	0.8084	0.8175	0.6149
✔	✔				0.8264	0.8459	0.6427	0.8061	0.8431	0.6294	0.8349	0.8525	0.6574
✔	✔	✔			0.8345	0.8478	0.6497	0.8207	0.8475	0.6316	0.8442	0.8476	0.6600
✔	✔		✔		0.8365	0.8549	0.6521	0.8224	0.8261	0.6426	0.8640	0.8685	0.6842
✔	✔	✔	✔		0.8838	0.8874	0.7128	0.8624	0.8661	0.6826	0.8921	0.9039	0.7212
✔	✔	✔	✔	✔	**0.8887**	**0.8964**	**0.7176**	**0.8951**	**0.8950**	**0.7297**	**0.9107**	**0.9141**	**0.7451**

Table 7. Running time comparison with other methods.

Model	NIQE [39]	QAC [69]	HOSA [68]	TLVQM [22]	VIDEVAL [53]	RAPIQUE [54]	VSFA [24]	GSTVQA [2]	SimpleVQA [49]	FastVQA [60]	Dover [62]	FineVQ
Time(s)↓	918.4	369.0	250.0	848.2	1881	152.7	95.05	164.1	71.62	**13.12**	**9.006**	**18.57**
FPS↑	2.613	6.505	9.600	2.830	1.277	15.72	25.26	14.63	33.52	**182.9**	**266.5**	**129.3**

benchmark datasets. It can be observed that FineVQ achieves state-of-the-art performance on all six datasets in terms of both SRCC and PLCC metrics. Specifically, the proposed FineVQ achieves 3.8% performance improvement compared to KSVQE [35] on LSVQ$_{1080p}$, demonstrating the effectiveness of FineVQ in high-resolution video quality assessment.

5.4. Cross-dataset Evaluation

We also conduct two cross-database evaluations, including (1) training on other datasets and testing on FineVD, and (2) training on FineVD and testing on other datasets. The results are shown in Table 4 and Table 5, respectively.

It can be concluded from the tables that, compared to the other two state-of-the-art VQA models, our proposed FineVQ model exhibits superior performance across both experimental settings, demonstrating the superior generalization capability of our model. Furthermore, comparing the generalization performance in Table 4 and Table 5, we find that our FineVD database is more challenging, as models trained on FineVD typically perform better on other datasets but exhibit poorer results in the reverse setting. This validates the diversity and broad distribution of our dataset, which suggests that training on FineVD should result in models with enhanced generalization ability.

5.5. Ablation Study

We conduct ablation studies to validate the utility of the core components of FineVQ. The results are shown in Table 6. The experiments are conducted on three datasets, including FineVD, LIVE-VQC [48] and YouTube-UGC [57], with consistent data partitioning and parameter settings.

Effectiveness of Motion Features. We first conduct ablation studies on the introduced motion feature extractor to validate the effectiveness of the extracted meticulous temporal features. Comparing the performance in the first two rows in Table 6, introducing motion features yields a significant improvement over the baseline for all datasets, which highlights the importance of integrating motion features into the proposed model.

Effectiveness of LoRA Adaptation. We then demonstrate the effectiveness of LoRA adaptation, as shown in rows 3 to 5 in Table 6. It can be observed that employing LoRA to either the image encoder (LoRA$_{r=8}$(vision)) or to the LLM (LoRA$_{r=8}$(llm)) enhances the performance of the proposed model. Moreover, combining both adaptations leads to a substantial improvement in performance.

Rank r in LoRA. Finally, we compare the influence of different rank r on the performance of FineVQ. As shown in the last two rows in Table 6, using $r = 16$ generally performs better than using $r = 8$ on all databases, which not only further validates the effect of LoRA, but also manifests the importance of task-specific adaptation.

Runtime. We also compare the runtime and FPS of state-of-the-art models on 10 2K 240-frame videos with an A6000 card or matlab and show results in Table 7. FineVQ achieves comparable results.

6. Conclusion

In this paper, we conduct a large-scale fine-grained UGC video quality assessment study. Specifically, we first present FineVD, a large-scale VQA dataset containing 6104 UGC videos with fine-grained quality scores and descriptions from multiple dimensions. Based on the database, we propose FineVQ to learn the fine-grained quality of UGC videos in a *one-for-all* fashion via instruction tuning, which has excellent capabilities of quality rating, quality scoring, and quality attribution. Extensive experiments on FineVD and other commonly used UGC-VQA datasets manifest that our FineVQ achieves state-of-the-art performance and can produce fine-grained video quality results.

Acknowledgment

This work was supported in part by the NSFC 62401365, 62225112, 62271312, 62132006, 62271308, in part by the STCSM 24ZR1432000, 24511106902, 22511105700, 22DZ2229005, and in part by the 111 plan BP0719010.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 4

[2] Baoliang Chen, Lingyu Zhu, Guo Li, Fangbo Lu, Hongfei Fan, and Shiqi Wang. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(4):1903–1916, 2021. 2, 3, 6, 7, 8

[3] Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Contrastive self-supervised pre-training for video quality assessment. *IEEE Transactions on Image Processing (TIP)*, 31:458–471, 2021. 3

[4] Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Unsupervised curriculum domain adaptation for no-reference video quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5178–5187, 2021. 3

[5] Yixiong Chen, Li Liu, and Chris Ding. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *arXiv preprint arXiv:2305.10843*, 2023. 3

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024. 2, 3, 5, 6

[7] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 6, 7

[8] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 6

[9] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 35: 16344–16359, 2022. 6

[10] Francesca De Simone, Marco Tagliasacchi, Matteo Naccari, Stefano Tubaro, and Touradj Ebrahimi. A h. 264/avc video database for the evaluation of quality metrics. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2430–2433, 2010. 2

[11] Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, Xiaokang Yang, and Patrick Le Callet. Confusing image quality assessment: Toward better augmented reality experience. *IEEE Transactions on Image Processing (TIP)*, 31: 7206–7221, 2022. 4

[12] Huiyu Duan, Wei Shen, Xiongkuo Min, Yuan Tian, Jae-Hyun Jung, Xiaokang Yang, and Guangtao Zhai. Develop then rival: A human vision-inspired framework for superimposed image decomposition. *IEEE Transactions on Multimedia (TMM)*, 25:4267–4281, 2022. 2

[13] Huiyu Duan, Xiongkuo Min, Sijing Wu, Wei Shen, and Guangtao Zhai. Uniprocessor: A text-induced unified low-level image processor. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019. 2, 5, 6

[15] Qihang Ge, Wei Sun, Yu Zhang, Yunhao Li, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, and Guangtao Zhai. Lmm-vqa: Advancing video quality assessment with large multimodal models. *arXiv preprint arXiv:2408.14008*, 2024. 3

[16] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 28(9): 2061–2077, 2017. 2

[17] Wang Haiqiang, Li Gary, Liu Shan, and Kuo C.-C. Jay. Icme 2021 ugc-vqa challenge. http://ugcvqa.com/, 2021. [Online]. 2

[18] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *Proceedings of the IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2017. 2, 7

[19] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2, 6

[20] Bilibili Inc. Bilibili. https://www.bilibili.com, 2010. 3

[21] Parimala Kancharla and Sumohana S Channappayya. Completely blind quality assessment of user generated video content. *IEEE Transactions on Image Processing (TIP)*, 31:263–274, 2021. 3

[22] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing (TIP)*, 28(12):5923–5938, 2019. 3, 6, 7, 8

[23] Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, 2024. 6, 7

[24] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 2351–2359, 2019. 3, 6, 7, 8

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022. 3

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 19730–19742. PMLR, 2023. 2, 3

[27] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, 2024. 6

[28] Yang Li, Shengbin Meng, Xinfeng Zhang, Shiqi Wang, Yue Wang, and Siwei Ma. Ugc-video: Perceptual quality assessment of user-generated videos. In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 35–38, 2020. 2

[29] Liang Liao, Kangmin Xu, Haoning Wu, Chaofeng Chen, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring the effectiveness of video perceptual representation in blind video quality assessment. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 837–846, 2022. 2, 3

[30] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 6

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 3, 5

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 2, 3, 5

[33] Yongxu Liu, Jinjian Wu, Leida Li, Weisheng Dong, Jinpeng Zhang, and Guangming Shi. Spatiotemporal representation learning for blind video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(6):3500–3513, 2021. 3

[34] Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*, 2023. 3

[35] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25963–25973, 2024. 1, 2, 3, 6, 7, 8

[36] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. 6

[37] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. Perceptual video quality assessment: A survey. *Science China Information Sciences*, 67(11):211301, 2024. 1

[38] Shankhanil Mitra and Rajiv Soundararajan. Multiview contrastive learning for completely blind video quality assessment of user generated content. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1914–1924, 2022. 3

[39] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters (SPL)*, 20(3):209–212, 2012. 6, 8

[40] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing (TIP)*, 25(1):289–300, 2015. 3

[41] Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo De Veciana. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, 6(6):652–671, 2012. 2

[42] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. Cvd2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing (TIP)*, 25(7):3073–3086, 2016. 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 3

[44] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing (TIP)*, 23(3):1352–1365, 2014. 3

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294, 2022. 5

[46] BT Series. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT*, pages 500–13, 2012. 4

[47] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing (TIP)*, 19(6):1427–1441, 2010. 2

[48] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing (TIP)*, 28(2):612–627, 2018. 2, 7, 8

[49] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 856–865, 2022. 2, 3, 6, 7, 8

[50] Wei Sun, Wen Wen, Xiongkuo Min, Long Lan, Guangtao Zhai, and Kede Ma. Analysis of video quality datasets via

design of minimalistic video quality models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. 7

[51] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023. 2, 5, 6

[52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[53] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing (TIP)*, 30:4449–4464, 2021. 1, 2, 6, 7, 8

[54] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021. 6, 7, 8

[55] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 1509–1513, 2016. 2

[56] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6, 7

[57] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5, 2019. 2, 7, 8

[58] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of ugc videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13435–13444, 2021. 2

[59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3

[60] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 538–554. Springer, 2022. 3, 6, 7, 8

[61] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 2

[62] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 20144–20154, 2023. 2, 3, 6, 7, 8

[63] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards explainable in-the-wild video quality assessment: a database and a language-prompted approach. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1045–1054, 2023. 2

[64] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 2, 3, 6

[65] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25490–25500, 2024.

[66] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 3, 4

[67] Jinjian Wu, Yongxu Liu, Leida Li, Weisheng Dong, and Guangming Shi. No-reference video quality assessment with heterogeneous knowledge ensemble. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 4174–4182, 2021. 3

[68] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing (TIP)*, 25(9):4444–4457, 2016. 6, 8

[69] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 995–1002, 2013. 6, 8

[70] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6, 7

[71] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051, 2024. 2, 3

[72] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq:'patching up'the video quality

problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14019–14029, 2021. 1, 2, 3, 6, 7

[73] Junyong You. Long short-term convolutional transformer for no-reference video quality assessment. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 2112–2120, 2021. 2, 3

[74] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. *arXiv preprint arXiv:2312.08962*, 2023. 2, 4

[75] Zhiyuan You, Jinjin Gu, Zheyuan Li, Xin Cai, Kaiwen Zhu, Chao Dong, and Tianfan Xue. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842*, 2024. 2, 4

[76] Xiangxu Yu, Neil Birkbeck, Yilin Wang, Christos G Bampis, Balu Adsumilli, and Alan C Bovik. Predicting the quality of compressed videos with pre-existing distortions. *IEEE Transactions on Image Processing (TIP)*, 30:7511–7526, 2021. 1, 2

[77] Xiangxu Yu, Zhenqiang Ying, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Subjective and objective analysis of streamed gaming videos. *IEEE Transactions on Games*, 2023. 7

[78] Kun Yuan, Zishang Kong, Chuanchuan Zheng, Ming Sun, and Xing Wen. Capturing co-existing distortions in user-generated content for no-reference video quality assessment. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1098–1107, 2023. 2, 3

[79] Zicheng Zhang, Wei Wu, Wei Sun, Danyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. Md-vqa: Multi-dimensional quality assessment for ugc live videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1746–1755, 2023. 2

[80] Qi Zheng, Zhengzhong Tu, Xiaoyang Zeng, Alan C Bovik, and Yibo Fan. A completely blind video quality evaluator. *IEEE Signal Processing Letters (SPL)*, 29:2228–2232, 2022. 3, 7

[81] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 3