VQA^2: Visual Question Answering for Video Quality Assessment

Ziheng Jia
jzhws1@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Zicheng Zhang
zzc1998@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Jiaying Qian
2022qjy@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Haoning Wu
haoning001@e.ntu.edu.sg
Nanyang Technological University
Singapore, Singapore

Wei Sun
Chunyi Li
Shanghai Jiao Tong University
Shanghai, China

Xiaohong Liu
xiaohongliu@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Weisi Lin
wslin@e.ntu.edu.sg
Nanyang Technological University
Singapore, Singapore

Guangtao Zhai
zhaiguangtao@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Xiongkuo Min*
minxiongkuo@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

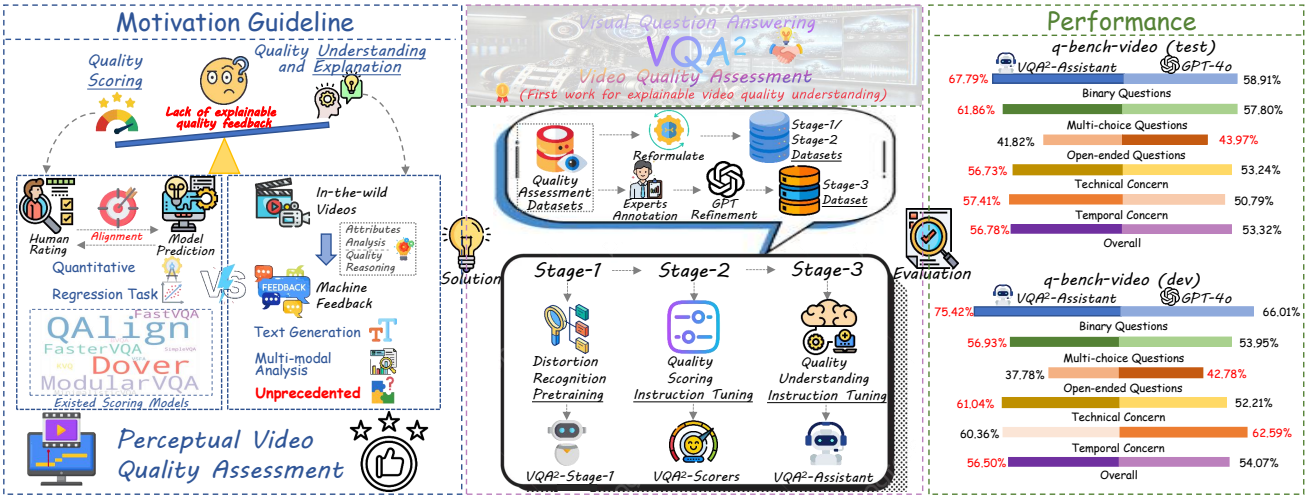

Figure 1: On the left is the motivation of our work: in the field of perception video quality assessment, the task of quantitative video quality scoring has been well solved while the task of quality understanding with explainable quality feedback still remains unprecedented, leaving a significant space for further research. The middle is an overview of the VQA^2 dataset and models. On the right side, a comparison of the performance of VQA^2-Assistant with GPT-4o on the video *q-bench-video* benchmark is presented. It can be observed that, across most dimensions, our model surpasses the performance of GPT-4o.

Abstract

The advent and proliferation of large multi-modal models (LMMs) have introduced new paradigms to computer vision, transforming various tasks into a unified visual question answering framework. Video Quality Assessment (VQA), a classic field in low-level visual perception, focused initially on quantitative video quality scoring. However, driven by advances in LMMs, it is now progressing toward more holistic visual quality understanding tasks. Recent studies in the image domain have demonstrated that Visual Question Answering (VQA) can markedly enhance low-level visual quality evaluation. Nevertheless, related work has not been explored in the video domain, leaving substantial room for improvement. To address this gap, we introduce the ***VQA^2 Instruction Dataset***—the first **visual question answering** instruction dataset that focuses on **video quality assessment**. This dataset consists of 3 subsets and covers various video types, containing 157,755 instruction question-answer pairs. Then, leveraging this foundation, we present the ***VQA^2 series models***. The *VQA^2 series models* interleave visual and motion tokens to enhance the perception of spatial-temporal

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '25, Dublin, Ireland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754696

quality details in videos. We conduct extensive experiments on video quality scoring and understanding tasks, and results demonstrate that the *VQA^2 series models* achieve excellent performance in both tasks. Notably, our final model, the ***VQA^2-Assistant***, exceeds the renowned *GPT-4o* in visual quality understanding tasks while maintaining strong competitiveness in quality scoring tasks. Our work provides a foundation and feasible approach for integrating low-level video quality assessment and understanding with LMMs. The dataset and codes are already available at *https://github.com/Q-Future/Visual-Question-Answering-for-Video-Quality-Assessment*.

CCS Concepts

• **Human-centered computing** → **Visualization design and evaluation methods**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Perceptual video quality assessment, Large multi-modal models, Instruction tuning

ACM Reference Format:

Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. 2025. VQA^2: Visual Question Answering for Video Quality Assessment. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.*https://doi.org/10.1145/3746027.3754696

1 Introduction

With the invention and rise of large multi-modal models (LMMs) [1, 15, 33, 66], the domain of computer vision pertaining to video has entered a new era. Visual Question Answering (VQA) [2], a pivotal tool for modality alignment, is widely employed in LMM applications. The paradigm of visual-language instruction tuning [33] using multi-modal instruction datasets, which encompass vast amounts of high-quality data, has markedly enhanced the performance of video LMMs in high-level visual tasks intimately related to video semantics such as video understanding [31, 71] and video temporal analysis [25]. Similarly, in low-level vision, incorporating visual question answering to aid model training and inference has tremendous potential for development. One potential field is Video Quality Assessment (VQA), which is closely related to video quality attributes (such as flicker, blur, and stuttering). We believe integrating visual question answering into developing video quality assessment models can deliver superior quantitative assessment and quality understanding capabilities compared to conventional models, thus holding greater potential for broad applications. The model can be utilized in video encoding, transmission, and decoding processes [24], providing effective feedback. Furthermore, it holds promise in the image/video generation domain as effective guidance for refining local generation details [28]. Concurrently, most video quality assessment models focus solely on video quality scoring [37]; however, they entirely lack the capability to understand and analyze quality details, resulting in significant deficiencies in model versatility. Moreover, existing models with low-level visual quality understanding function almost exclusively apply to the image field [70], lacking effective perception proficiencies for video-specific temporal and motion quality attributes. To bridge this gap, we construct the ***VQA^2 Instruction Dataset***—a large-scale instruction dataset specifically for video quality assessment based on visual question answering. The dataset lays a solid foundation for developing robust video quality assessment models with remarkable versatility. The construction pipeline can be divided into 3 stages along with their corresponding subsets:

- *Stage-1: Subset centered on distortion recognition for model pre-training.* We develop a distortion recognition instruction subset for model pre-training, leveraging the distortion information from multiple existing datasets.
- *Stage-2: Instruction tuning subset centered on video quality scoring.* We utilize the mean opinion scores (MOSs) from various existing datasets and transform them into quality-level labels serving as instruction data.
- *Stage-3: Instruction tuning subset for video quality understanding.* We curate a high-quality, diverse dataset expanded by *GPT* following human expert annotations.

Our core contributions are as follows:

(1) We construct the *first* **visual question answering** based instruction-tuning dataset for **video quality assessment** — the ***VQA^2 Instruction Dataset*** which encompasses 3 subsets and includes over 150,000 instruction pairs, covering various video types such as user-generated content (UGC), streaming, and artificial intelligence generated content (AIGC) videos. This ensures data adequacy and diversity.
(2) We design a complete training strategy and introduce the ***VQA^2 series models***, including the ***VQA^2-Scorers*** and the ***VQA^2-Assistant***. This series of models demonstrates strong functional diversity.
(3) The *VQA^2-Scorers* achieve state-of-the-art (SOTA) performance in video quality scoring tasks. Meanwhile, the *VQA^2-Assistant* excels in video quality understanding tasks, outperforming the proprietary *GPT-4o* on relevant benchmark tests. It also maintains robust performance in quality scoring tasks, showcasing the model's functional versatility and adaptability. The overview and performance of our work are summarized in Fig. 1.

2 Related Works

2.1 Video Quality Assessment

Classical video quality assessment tasks heavily rely on the MOSs obtained from subjective experiments. Two significant tasks among them are the UGC video and streaming video quality scoring tasks. UGC video quality scoring task involves datasets like [17, 19, 40, 44, 50, 62], which contain various authentic or synthetic distortions. Streaming video quality scoring datasets include the Waterloo-SQoE series [9–12] and the LIVE-NFLX series [5, 6, 16], which mainly use simulated transmission distortions (such as rebuffering, long-term stalling, and bitrate switching).

Classic video quality assessment models can be broadly divided into knowledge-driven and data-driven approaches. Knowledge-driven models [8, 12, 38, 39, 43, 47, 48] rely on elaborately designed features to evaluate video quality. In contrast, data-driven models [4, 22, 26, 29, 34, 35, 45, 46, 51, 52, 54, 62, 69] primarily employ deep

Table 1: Source datasets, sampling information and statistic summary of the *VQA^2 Instruction Dataset*.

Stages	Data Types	Source Dataset	Original / Sampled #	Instruction Pairs
Stage-1	Images	KonIQ-10k [20]	10,073 / 3,693	7,174
		KADID-10k [32]	10,125 / 3,481	
	UGC-videos	LIVE-Qualcomm [17]	208 / 34	5,211
		LSVQ (train) [62]	28,056 / 100	
Stage-2	UGC-videos	LSVQ (train)	28,056 / 28,056	28,056
	Streaming-videos	Waterloo-I [12]	180 / 180	2,100
		Waterloo-III [11]	450 / 450	
		LIVE-NFLX-II [5]	420 / 420	
Stage-3	UGC-videos	LSVQ (train)	28,056 / 13,007	110,232
		LSVQ (1080p) [62]	3,573 / 998	
		LIVE-VQC [44]	585 / 497	
	AIGC-videos	VideoFeedback [18]	32,901 / 998	4,982
Overall	Images: 7,174 / UGC videos (no overlap): 29,585 Streaming videos: 1,050 / AIGC videos: 998			157,755

neural networks (DNNs) to extract features sensitive to video quality. With the widespread application of LMMs in low-level vision, recent works based on LMMs have achieved higher performance. For example, q-align [57] attains high precision and generalizability in video quality scoring.

However, these models possess only the capability to score the video quality but almost entirely miss the function of video quality understanding and analysis, with no capability to provide reasonable responses to diverse question types. Thus, they can not realize the boosting demand for quality understanding and analysis of spatial and temporal quality attributes in videos. Our proposed *VQA^2-Assistant* can perform precise video quality scoring while exhibiting strong capabilities in video quality understanding and question answering, marking new progress in this field.

2.2 Low-level Visual Question Answering

In the field of low-level visual question answering for images, recent research has made significant advances. Q-Bench [68] is an image-centered visual question-answering-based benchmark for evaluating LMMs on low-level quality understanding tasks. Q-Instruct [56] has substantially enhanced LMMs' capabilities in understanding the low-level visual quality of images by constructing a large and diverse instruction dataset through human annotation. DepictQA [63] leverages LMMs to provide detailed, language-driven evaluations that outperform traditional score-based methods. AesExpert [21] builds expert-level aesthetic foundational models by assembling a rich corpus of image aesthetic critique datasets. Co-Instruct [58] has significantly improved large models' ability to analyze the quality of multiple visual stimuli by creating instruction datasets based on multi-image comparisons and joint analysis.

However, for videos, no existing work has incorporated low-level visual question answering to create models with enhanced video quality and understanding ability. Our work is the first to achieve this, thus paving a promising way for deep video quality analysis through LMMs.

3 The VQA^2 Instruction Dataset

As the foundation of our work, we propose the *VQA^2 Instruction Dataset*. The dataset construction is composed of 3 stages: *Stage-1* is used for distortion recognition pretraining, while *Stage-2* and *Stage-3* focus on improving the model's capabilities in video quality scoring and video quality understanding, respectively. The data construction pipelines are shown in Fig. 2 and Fig. 3.

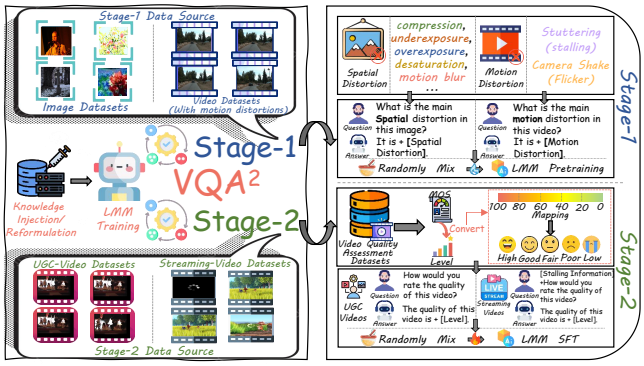

Figure 2: Data construction pipelines of *Stage-1* and *Stage-2*.

3.1 Video Selection or Sampling

To ensure the diversity of video content, the videos collected for the *VQA^2 Instruction Dataset* are sourced from various image/video datasets [5, 11, 12, 17, 18, 20, 32, 44, 62], providing a wide range of visual content and quality variations. We determine the sampling proportion of different quality levels according to each dataset's original quality distribution. This is because almost all source datasets exhibit a normal distribution of quality, and uniform sampling fails to capture sufficient videos across all quality levels. More importantly, we believe that moderate-quality videos encompass both positive and negative attributes, offering greater annotation value than videos of purely high or low quality, and this sampling strategy can guarantee sufficient medium-quality videos to be selected. Tab. 1 summarizes our dataset's sampling and statistic information.

3.2 Distortion-recognition Based Pretraining Set

Many classic multi-modal works [30, 33, 41, 61] follow the well-established pretraining-finetuning paradigm, which has been proven to be an effective way for foundation model development. Since we believe that distortion is central to low-level quality assessment and the recognition of distortion types is fundamental to achieving high performance on such tasks, we design *VQA^2 Instruction Dataset*

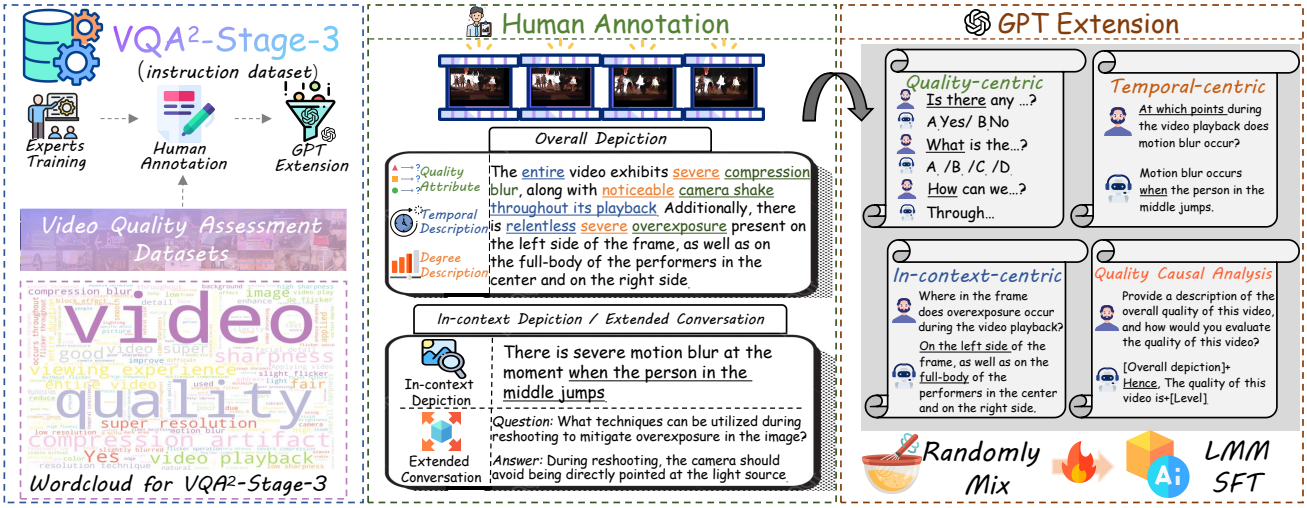

Figure 3: Data construction pipeline of *Stage-3*, which is annotated by human and then refined and expanded through *GPT*.

Stage-1, a distortion recognition-centered instruction subset, as the model's pre-training instruction set.

Our distortion recognition design includes both spatial and motion aspects. We sample 7,174 images from the KonIQ-10K [20] and KADID-10K [32] datasets for spatial distortion recognition. Using the distortion type annotations from [65], we select 11 different types of spatial distortions: *'compression artifact'*, *'spatial blur'*, *'motion blur'*, *'noise'*, *'overexposure'*, *'underexposure'*, *'low contrast'*, *'high contrast'*, *'oversaturation'*, *'desaturation'* and *'block effect'*. For motion distortion recognition, we focus on the video distortions—*'flickers (camera shake)' and 'stuttering'*. We use 34 videos with flicker distortion from the LIVE-Qualcomm [17] and 100 videos with flicker and stuttering distortions in the LSVQ (train) [62] and extend them by extracting temporal and spatial clips, yielding 5,211 video clips containing motion distortions. The format of instructions is shown in Fig. 2.

3.3 Instruction Set for Quality Scoring

Video quality scoring has always been a key focus of video quality assessment. The accurate assessment of video quality forms the foundation for understanding and analyzing quality attributes within the video. To this end, we design the *VQA² Instruction Dataset Stage-2*, a finetuning subset primarily aimed at scoring the video quality. We use the LSVQ (train) as the video source for the offline UGC video quality scoring task while using Waterloo-I [12], Waterloo-III [12], and LIVE-NFLX-II [5] as the video sources for the streaming video quality scoring task. To ensure that the subjective experimental scores from all datasets are on the same scale, we normalize the MOSs in each dataset to the [0, 100] range. After scaling, we transform the video quality into five quality levels: *'High'*, *'Good'*, *'Fair'*, *'Poor'* and *'Low'*, with each level representing a 20-point interval. This approach minimizes the impact of inconsistent quality distributions across datasets.

Since long-term stalling is a significant distortion in the streaming video datasets, we add **stalling information** in the instruction set. We design two formats to present the stalling information. The first format uses a *'0/1'* sequence to directly indicate stalling for each frame ('*1*' represents stalling, and '*0*' represents smooth playback). In the second format, the **stalling information** is summarized as follows: *the total number of stalling events*, *the duration of each stalling event*, *the proportion of stalling events duration to the total video length*, *the initial buffering time*, and *the time elapsed between the end of the last stalling event and the end of the playback*. The UGC and streaming video prompt formats are shown in Fig. 2.

3.4 Instruction Set for Quality Understanding

The primary part of the dataset is *VQA² Instruction Dataset Stage-3*, an instruction subset for low-level video visual quality understanding and question answering. We select 14,005 videos from LSVQ (train) and LSVQ (1080p) [62], 497 videos from thise LIVE-VQC [44], and 998 videos from the AIGC video dataset Videofeedback [18]. This subset focuses on low-level visual quality question-answering for videos while also involving a small number of data related to video aesthetic assessment (VAA) and AIGC video quality analysis. For constructing more question-answer (Q&A) pairs, we adopt the method of extending human expert annotations using *GPT*.

Human Annotation Process. The human annotation process for each video is divided into two parts. First, we require a comprehensive **overall quality depiction** focused on video quality attributes. Each depiction includes several quality attributes described with three key elements: **Quality Attribute+Degree+Temporal Description**. A detailed example is shown in Fig. 3. This design highlights the key point that the viewer perceives the video quality by focusing on salient quality attributes. Furthermore, the depiction incorporates rich temporal information, significantly enhancing the model's capacity to answer questions about the temporal quality within videos. We instruct annotators to select quality attributes for overall depiction according to the transformed video quality levels. To maximize the value of medium-quality videos, annotators are encouraged to identify a broader range of quality attributes in these videos, encompassing both positive and negative aspects.

In the second part, we require annotators to provide a brief quality depiction containing in-context (local) temporal or spatial quality. If such a depiction can not be found, it may be substituted with extended conversations, such as designing a Q&A pair about possible causes of certain distortions or proposing feasible solutions to enhance video quality. We provide more detailed information for annotation experiments and examples in the *Supplementary Materials (supp.)*.

GPT Extension. Most of our Q&A pairs are generated using *GPT* to rewrite and refine the overall quality depictions, in-context depictions, and extended conversations. For each overall depiction, we instruct *GPT* to extract the information and reformulate it into three quality (attributes) centric Q&A pairs: a binary-choice single-answer pair, a multiple-choice single-answer pair, and an open-ended pair. Additionally, we require *GPT* to generate one temporal-centric Q&A pair and another extended conversation (except for the human-annotated one) based on the original overall depiction. For the human-annotated in-context depictions and extended conversations, the *GPT* also rewrites them to the format of formal Q&A pairs with their meaning unchanged. The prompts for GPT extension are detailed in *supp.*.

Quality Causal Analysis. For the annotated overall depictions mentioned before, to fully utilize the information during training, we manually formulate them into the form of quality causal analysis Q&A pairs with the specific format shown in Fig. 3.

4 The VQA^2 Series Models

We propose the ***VQA^2 series models*** and a unique training pipeline. The structure of the model is illustrated in Fig. 4. The loss function in all three training stages adopts the standard generation loss used in text generation tasks, such as GPT-loss [42].

4.1 Model Structure

Base Model. We select *LLaVA-OneVision-Chat-7B* [27] as the foundation of our model. The foundation model achieves excellent performance on multiple high-level visual question answering benchmarks [14, 36] and demonstrates outstanding capabilities in video semantic understanding and reasoning. The foundation model includes: a vision tower constructed from the SigLIP [64], which is used to extract feature tokens from keyframe sequences; a vision projector composed of fully connected layers for feature mapping; and the Qwen-2 model [59] with its tokenizer serving as the LLM and text embedding layers.

Motion Extractor and Motion Projector. We observe that many video LMMs [7, 27, 60] can achieve excellent performance by only inputting keyframe sequences extracted at sparse sampling rates (*1 frame per second (fps), etc.*) when evaluated on high-level visual question answering tasks. We believe this is due to the high redundancy of temporal semantic information in videos. The semantic content in adjacent video frames is highly consistent, and changes or connections in video contents over longer periods can be effectively captured through sparsely sampled keyframe sequences. However, this redundancy is sometimes almost nonexistent in video quality assessment. For example, video stuttering or shaking occurring within a short time can significantly degrade a video's perception quality, and such distortion is closely related to the adjacent frames. Therefore, keyframe sequences extracted at long intervals will completely miss this distortion representation.

In summary, we propose that the model should incorporate a video motion extraction module that processes adjacent frames from densely input video frames. We select the SlowFast-R50 [13] for motion extraction, inputting the entire video after spatial preprocessing. To ensure that the number of the motion tokens aligns with the number of frames of the video, we set $\tau = 4$[1] and $\alpha = 4$[2] and use only the fast path features as the motion tokens. We then employ a motion projector identical in structure to the vision projector to map the motion tokens, ensuring their dimensions are consistent with the visual tokens and text tokens. Additionally, we apply positional encoding by performing token-wise addition on the motion tokens and learnable absolute positional embeddings. During training, the visual tokens sequence and motion tokens sequence are input in an interleaved manner.

4.2 Distortion-recognition Based Pretraining

We first pre-train the model using the data from *Stage-1*. When training with the spatial distortion instruction subset, we freeze the LLM, motion extractor, and motion projector, allowing only the vision tower and vision projector to be trained. In contrast, when using the motion distortion instruction subset, we exclusively train the motion extractor and projector. This is because we think the pre-training data has a relatively simple format. Unfreezing all model parameters for training could easily lead to severe overfitting, thereby affecting the subsequent training process.

4.3 The VQA^2-Scorers and the VQA^2-Assistant

The *VQA^2-Scorers* consist of the *VQA^2-UGC-Scorer* and the *VQA^2-Streaming-Scorer*, which are specially developed for scoring the quality of UGC and streaming videos, respectively. The *VQA^2-UGC-Scorer* is trained on the instruction subset from the *Stage-2* UGC video data portion upon the pre-trained model. Subsequently, the *VQA^2-Streaming-Scorer* is further trained on the instruction set from the *Stage-2* streaming video data portion based on the *VQA^2-UGC-Scorer*. The reason for this setting is given in the *supp.*. The scoring methodology for model inference, referenced from [57], is shown in Fig. 4 and detailed in the *supp.*. At this stage, all training involves full parameter tuning.

The *VQA^2-Assistant* is designed to master more nuanced video quality understanding tasks and efficient low-level visual question answering while still possessing the capabilities for precise quality scoring. Building on the *VQA^2-UGC-Scorer*, this model also undergoes full parameter tuning using the instruction subset *Stage-3*.

5 Experiments

We conduct comprehensive experiments on video quality scoring and understanding tasks to validate the performance of our model family. Additionally, we perform detailed ablation studies to analyze the impact of various attributes.

[1] τ denotes the sampling interval (frames) in the slow path.
[2] α denotes the ratio of the sampling intervals of the slow and fast path.

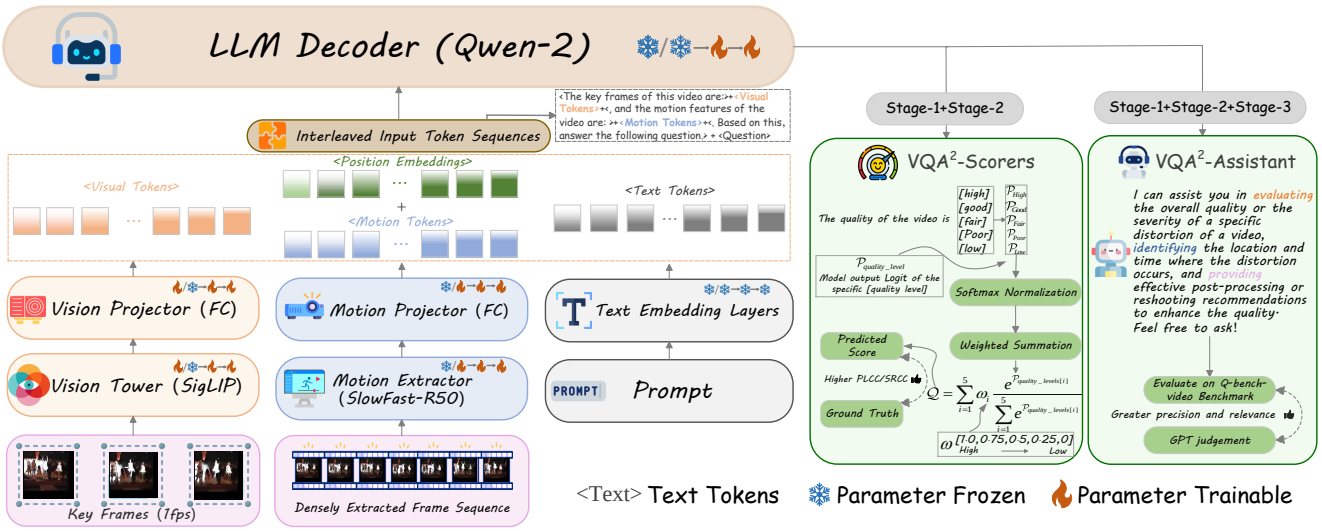

Figure 4: The model structure and training strategy. The model is based on the vanilla *LLaVA-OneVision-Chat (7B)* and *SlowFast-R50*. The training strategy (freeze or unfreeze) across the 3 stages is separated by "→". Specifically, "/" denotes the training strategy used during the "image data training/video data training" in *Stage-1*.

5.1 Experimental Setups

Training Strategies. We strictly follow the training hyperparameters provided in the *LLaVA-OneVision* project. All the models are trained for only one epoch on their respective training data. More details are shown in *supp.*. **SystemPrompt Design.** During the training and evaluation stages, a system prompt is added before all instructions in the dataset. We design the system prompt to be identical in all stages of training. During the evaluation, we employ specially designed system prompts based on the specific task. To thoroughly verify the model's end-to-end performance for the video quality scoring tasks, we do not provide any time information (like length, frame rate, and stalling information) that may need additional extraction about the test video in the system prompt. In the video quality understanding task, to meet the format requirements of the evaluation benchmark, we add time information to the evaluation prompt, including the number of video frames, sampled keyframes, and the sampling interval. The specific system prompt formats in the training and evaluation stages are presented in the *supp.*.

5.2 Evaluation on Quality Scoring Tasks

We assess the quantitative scoring capabilities of *VQA²-UGC-Scorer* and *VQA²-Assistant* on 4 open-source UGC-VQA datasets [19, 44, 50, 62], while evaluating the *VQA²-Streaming-Scorer* on the Waterloo-IV [9], the largest streaming-VQA dataset with 1,350 videos. We select some vanilla LMMs [33, 61] and our base model (with vanilla *LLaVA-ov-chat (7B)* and *SlowFast-R50*) for reference. Additionally, we choose several high-performing UGC-VQA models (*FAST-VQA* [53], *Minimalistic-VQA* [46], *DOVER* [55], *Modular-VQA* [52], *q-align* [57], and *q-instruct* [56]) and Streaming-VQA models (*SQI* [12], *BSQI* [8], and *DSA-QoE* [22]) for further comparison.

We use *Pearson Linear Correlation Coefficient* (PLCC) and *Spearman Rank Correlation Coefficient* (SRCC) as evaluation metrics. Apart from the *q-align-IQA*, *q-align-onealign*, *q-instruct*, *VQA²-Streaming-Scorer*, and *VQA²-Assistant* which have specific training sets and the training-free method *SQI*, all models in both tasks are trained or optimized on the same datasets (LSVQ (train) for UGC video quality scoring / Waterloo-I, Waterloo-III, and LIVE-NFLX-II for streaming video quality scoring). Specifically, since *Stage-3* involves videos from the LIVE-VQC and LSVQ (1080p), we remove these videos to avoid leakage in all experiments, leaving 88 videos out of 585 in LIVE-VQC and 2,575 videos out of 3,573 in LSVQ (1080p) for evaluation. Tabs. 2 and 3 present the performance of our models and the comparison models on UGC / streaming video quality scoring tasks, respectively.

Experimental results demonstrate that the *UGC-Scorer* and the *Streaming-Scorer* achieve the best performance across most datasets and metrics and rank within the top 3 in almost all of them. Although the *Assistant* slightly lags behind the *UGC-Scorer* in performance, it still delivers a relatively strong scoring performance. This confirms that the *Assistant*, primarily designed for video quality understanding and question answering, can still effectively handle quality scoring tasks, showcasing its versatility.

5.3 Evaluation on Quality Understanding Tasks

We evaluate the capabilities of the *VQA²-Assistant* on video quality understanding tasks using the Q-Bench-Video [67], a comprehensive LMM benchmark for video quality understanding tasks. It contains 1,800 videos and 2,378 multi-type questions. This evaluation encompasses the model's answer accuracy and relevance across 3 question types: binary yes/no questions (*Binary*), multiple-choice (single answer) questions (*Multi.*), and open-ended questions (*Open*). These questions also span 3 quality concerns: video technical quality aspect (*Tech.*), video temporal quality aspect (*Temp.*), and other categories (*Other*), which include AIGC and VAA, etc. The model's responses are evaluated using *GPT* with the same settings in Q-Bench-Video. Specifically, since our training does not

Table 2: Performance on UGC video quality scoring tasks. [red:best; blue: the second best; underline: the third best]

Datasets	LSVQ(1080p) [62]		LSVQ(test) [62]		LIVE-VQC [44]		YT-UGC [19]		KoNViD-1k [50]	
Metrics	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
FAST-VQA [53]	0.765	0.810	0.874	0.878	0.769	0.815	0.725	0.742	0.859	0.857
Minimalist-VQA [46]	0.769	0.818	0.880	0.872	0.765	0.812	0.783	0.799	0.859	0.861
Dover [55]	0.787	0.828	0.888	0.886	0.771	0.819	0.801	0.814	0.890	0.883
Modular-VQA [52]	0.791	0.844	0.894	0.891	0.783	0.825	0.786	0.803	0.878	0.884
mPLUG-Owl-2 [61]	0.398	0.422	0.422	0.434	0.450	0.459	0.437	0.448	0.532	0.532
LLaVA-v1.5 [33]	0.341	0.355	0.441	0.412	0.242	0.302	0.356	0.378	0.435	0.419
q-align-IQA (7B) [57]	0.764	0.842	0.729	0.733	0.724	0.772	0.715	0.723	0.797	0.780
q-align-VQA (7B)	0.758	0.833	0.883	0.882	0.777	0.813	0.811	0.830	0.865	0.876
q-align-onealign (7B)	0.761	0.822	0.886	0.884	0.766	0.826	0.831	0.847	0.876	0.878
q-instruct-mPLUG (7B) [56]	0.602	0.580	0.644	0.640	0.660	0.673	0.601	0.622	0.492	0.520
q-instruct-LLaVA (7B)	0.571	0.562	0.610	0.578	0.685	0.616	0.635	0.667	0.664	0.577
Base	0.527	0.609	0.658	0.638	0.665	0.751	0.626	0.631	0.692	0.691
UGC-scorer	0.782	0.847	0.897	0.885	0.785	0.830	0.814	0.832	0.894	0.884
Assistant	0.760	0.819	0.882	0.856	0.776	0.823	0.854	0.841	0.883	0.844

Table 3: Performance on streaming video quality scoring tasks. *Type 1-5* refers to the 5 unique video content types in the Waterloo-IV dataset: *game, documentary, movie, nature, and sports*, each comprising 270 videos. *All* refers to the entire dataset.

Datasets	Type-1		Type-2		Type-3		Type-4		Type-5		Overall	
Metrics	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
SQI [12]	0.778	0.791	0.752	0.763	0.685	0.701	0.642	0.625	0.768	0.783	0.672	0.671
BSQI [8]	0.792	0.813	0.756	0.795	0.707	0.722	0.683	0.724	0.807	0.783	0.707	0.721
DSA-QoE [22]	0.803	0.815	0.767	0.798	0.713	0.722	0.68	0.715	0.813	0.802	0.733	0.744
LLaVA-v1.5	0.542	0.456	0.557	0.613	0.455	0.467	0.384	0.426	0.751	0.817	0.334	0.386
q-align-IQA (7B)	0.683	0.772	0.665	0.697	0.496	0.438	0.694	0.712	0.805	0.813	0.365	0.317
q-align-VQA (7B)	0.805	0.852	0.687	0.765	0.567	0.703	0.664	0.638	0.783	0.825	0.461	0.503
q-align-onealign (7B)	0.774	0.866	0.734	0.781	0.707	0.595	0.493	0.561	0.864	0.805	0.437	0.384
Base	0.322	0.375	0.673	0.697	0.496	0.432	0.391	0.425	0.747	0.766	0.373	0.374
Streaming-scorer	0.857	0.906	0.753	0.782	0.729	0.747	0.709	0.68	0.866	0.827	0.766	0.788

include multi-video analysis, we exclude the questions involving multi-video (564 out of 2, 378). We select a series of high-performing open-source video LMMs [3, 7, 23, 27, 33, 49, 60, 61, 66] and proprietary *GPT series* [1] for comparison. The keyframe sequence input for all models is sampled at 1 fps. Except for the *GPT series*, we set the *greedy search* scheme for model generation, ensuring all results are reproducible. The accuracy of all models on the *test* and *dev* sets of the Q-Bench-Video are presented in Tabs. 4 and 5.

Experimental results indicate that the *Assistant* outperforms the base model across all sub-dimensions in both subsets. In terms of question types, the *Assistant* achieves the most significant improvement on binary questions, surpassing the base model by 19.64% and *GPT-4o* by 8.88% on the *test* set; and it also shows notable gains on what/how questions. As for quality concerns, the *Assistant* achieves substantial improvements in the *Tech.* and *Temp.* sub-dimensions, exceeding the base model by 10.05% and 10.64% in each on the *test* set. This underscores the importance of centering human annotations on quality attributes while incorporating extensive temporal and motion descriptions. Finally, the *Assistant* outperforms *GPT-4o* in *overall* scores on both the *test* and *dev* subsets.

To visualize the application of the quality understanding ability of our *VQA^2-Assistant*, we provided 12 **case study** examples in *supp.*, including 4 **real-world scenario** Q&A instances presented by the Gradio demo.

5.4 Ablation Study

We conduct ablation studies and provide corresponding analyses.
#1: Effects of Pre-training. We remove *Stage-1* and directly train the *UGC-Scorer* using the data from *Stage-2*. Compared to the fully trained model, the resulting model's performance is shown in Tab. 6. Experimental results indicate that the pretraining *Stage-1* plays a crucial role throughout the training process.
#2: Effects of Motion Extraction. We remove the motion extractor and projector, then follow the same training steps to obtain the *Scorers* and *Assistant* models. The models' performance on KonVID-1k and LIVE-VQC, as well as their scores on the *Tech.*, *Temp.*, and *Overall* sub-dimensions in the Q-Bench-Video *dev* subset, are presented in Tab. 7. Experimental results show that motion feature extraction plays a significant role, especially when evaluating the *Temp.* quality concern.

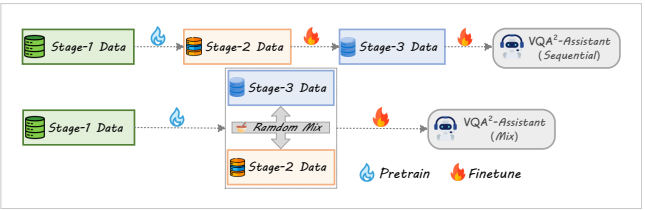

Figure 5: Two different data-combining strategies.

#3: Ablation Study on Different Data-combining Strategies. During the training of the *VQA^2-Assistant*, we alternatively combine the *Stage-2* and *Stage-3* data by randomly mixing them. The data-combining strategies are illustrated in Fig. 5. We compare the performance of the *Mix* and *Sequential* versions of the model on the UGC video quality scoring task and the *test* set of Q-Bench-Video, with the results reported in Tabs. 8 and 9.

Table 4: Evaluation results on the *test* subset of the Q-bench-video. [red:best; blue: the second best; underline: the third best; green: the performance improvement of the trained model compared to the base model]

Categories	Question Types			Quality Concerns			Overall↑
LMMs	Binary↑	Multi.↑	Open↑	Tech.↑	Temp.↑	Other↑	
mPLUG-Owl-2 (7B)	57.72%	42.61%	32.39%	41.99%	46.46%	44.79%	43.99%
LLaVA-v1.5 (7B)	58.39%	50.17%	39.78%	46.31%	48.32%	54.01%	49.23%
mPLUG-Owl3 (7B) [60]	58.05%	57.73%	38.99%	53.04%	51.68%	51.34%	51.27%
LLaVA-ov-chat (7B) [33]	57.05%	53.95%	33.65%	47.60%	45.45%	51.07%	47.85%
InternVL-Chat (7B) [7]	65.77%	55.33%	34.43%	47.04%	50.34%	57.75%	51.43%
VILA1.5 (7B) [23]	58.05%	45.36%	37.26%	45.99%	45.96%	49.33%	46.69%
LLaVA-Next-Video (7B) [66]	65.10%	50.17%	38.21%	48.80%	48.15%	55.75%	50.88%
Qwen2-VL (7B) [49]	50.84%	55.75%	34.49%	46.03%	56.40%	47.28%	46.67%
Qwen2.5-VL (7B) [3]	52.53%	49.48%	38.77%	46.68%	56.53%	42.16%	46.72%
GPT-4V	58.25%	50.17%	44.62%	51.38%	47.79%	51.61%	50.89%
GPT-4o-mini	50.17%	45.99%	37.34%	44.98%	46.60%	33.85%	44.33%
GPT-4o	58.91%	57.80%	43.97%	53.24%	50.79%	56.92%	53.32%
Base	52.86%	53.31%	34.65%	46.68%	46.77%	50.13%	46.61%
Assistant (7B)	67.79%	61.86%	41.82%	56.73%	57.41%	51.34%	56.78%
	+14.93%	+8.55%	+7.17%	+10.05%	+10.64%	+1.21%	+10.17%

Table 5: Evaluation results on the *dev* subset of the Q-bench-video.

Categories	Question Types			Quality Concerns			Overall↑
LMMs	Binary↑	Multi.↑	Open↑	Tech.↑	Temp.↑	Other↑	
mPLUG-Owl-2 (7B)	61.79%	37.35%	33.39%	42.67%	51.94%	45.66%	44.27%
LLaVA-v1.5 (7B)	63.79%	46.99%	36.31%	49.33%	45.58%	48.51%	49.34%
mPLUG-Owl3 (7B)	59.14%	57.23%	38.50%	52.00%	54.77%	51.76%	52.21%
LLaVA-ov-chat (7B)	60.47%	53.01%	32.85%	49.42%	52.12%	45.81%	49.39%
InternVL-Chat (7B)	70.43%	49.70%	33.39%	50.25%	51.41%	49.59%	51.65%
VILA1.5 (7B)	58.05%	45.36%	37.26%	45.99%	45.96%	49.33%	46.69%
LLaVA-Next-Video (7B) [66]	69.77%	44.58%	33.39%	49.08%	49.65%	50.00%	49.56%
Qwen2-VL (7B)	57.48%	51.83%	32.78%	51.19%	55.42%	44.78%	47.93%
Qwen2.5-VL (7B)	56.80%	45.43%	39.26%	48.30%	60.59%	45.17%	47.31%
GPT-4V	64.78%	51.20%	43.43%	55.25%	56.18%	50.00%	53.36%
GPT-4o-mini	54.49%	43.98%	38.87%	45.58%	54.42%	47.15%	45.92%
GPT-4o	66.01%	52.56%	42.78%	52.21%	62.59%	50.28%	54.07%
Base	55.78%	45.43%	31.48%	45.25%	50.73%	44.69%	44.62%
Assistant (7B)	75.42%	56.93%	37.78%	61.04%	60.36%	46.87%	56.50%
	+19.64%	+11.50%	+6.30%	+15.79%	+9.63%	+2.18%	+11.88%

Table 6: Comparison of the *UGC-Scorer* performance (SRCC / PLCC) with / without the *Stage-1*. [red: best]

Experiment Types	Intra-dataset=		Cross-dataset =	
Models	*LSVQ(test)*	*LSVQ(1080p)*	*LIVE-VQC*	*KoNViD-1k*
wo-Stage1	0.887 / 0.879	0.751 / 0.812	0.776 / 0.819	0.884 / 0.880
w-Stage1	0.897 / 0.885	0.782 / 0.847	0.785 / 0.830	0.894 / 0.884

Table 7: Comparison of model performance with and without the motion features extractor.

Models	UGC-Scorer		Assistant (*dev* set)		
Version	*LIVE-VQC*	*KoNViD-1k*	*Tech.*↑	*Temp.*↑	*Overall*↑
wo-Motion	0.773 / 0.822	0.873 / 0.865	58.17%	55.85%	54.30%
w-Motion	0.785 / 0.830	0.894 / 0.884	61.04%$_{+2.87\%}$	60.36%$_{+4.51\%}$	56.50%$_{+2.20\%}$

Table 8: Comparison of the *UGC-Scorer's* performance with different data mixture strategies.

Version	*LIVE-VQC*	*LSVQ(1080p)*	*LSVQ(test)*	*KoNViD-1k*
Mix	0.789 / 0.836	0.761 / 0.831	0.883 / 0.873	0.895/0.892
Sequential	0.776 / 0.823	0.760 / 0.819	0.882 / 0.856	0.883 / 0.844

The model trained with the *Mix* strategy further improves scoring performance in several datasets, suggesting that, beyond directly answering video quality-level questions, incorporating detailed quality understanding instructions still contributes to quantitative quality scoring. However, the *Mix* version experiences a notable decline in quality understanding tasks. We attribute this to the scoring-related instructions in *Stage-2*, which typically have a relatively simple focus and format. This impedes the model for handling diverse question types effectively.

Table 9: Comparison of the *Assistant's* performance on the *Q-bench-video-test* with different data mixture strategies.

Version	Question Types			Quality Concerns			Overall↑
	Binary↑	*Multi.*↑	*Open*↑	*Tech.*↑	*Temp.*↑	*Other*↑	
Mix	65.44%	57.04%	39.78%	56.25%	54.55%	47.19%	53.75%
Sequential	67.79%	61.86%	41.82%	56.73%	57.41%	51.34%	56.78%
Improvement	+2.35%	+4.82%	+2.04%	+0.48%	+2.86%	+4.15%	+3.03%

6 Conclusion

We introduce the *VQA^2 Instruction Dataset*, the first large-scale instruction dataset dedicated to video quality assessment through visual question answering, alongside the *VQA^2 series models* built on this dataset. The dataset construction spans 3 stages and includes 157, 755 instruction pairs from diverse video types. Through comprehensive experiments on video quality scoring and quality understanding tasks, our models achieve excellent results in video quality scoring and outperform *GPT-4o* in video quality understanding. The *VQA^2-Assistant* excels in both quality scoring and understanding tasks, effectively fulfilling the demand for model versatility. Our work lays the foundation for the creation of video quality assessment agents.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62271312 and Grant 62132006, and in part by STCSM under Grant 22DZ2229005.

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. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *CVPR*. 2425–2433.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[4] Christos G Bampis and Alan C Bovik. 2018. Feature-based prediction of streaming video QoE: Distortions, stalling and memory. *Elsevier Signal Processing: Image Communication* 68 (2018), 218–228.

[5] Christos G Bampis, Zhi Li, Ioannis Katsavounidis, Te-Yuan Huang, Chaitanya Ekanadham, and Alan C Bovik. 2021. Towards perceptually optimized adaptive video streaming-a realistic quality of experience database. *IEEE TIP* 30 (2021), 5182–5197.

[6] Christos George Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik. 2017. Study of temporal effects on subjective video quality of experience. *IEEE TIP* 26, 11 (2017), 5217–5231.

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821* (2024).

[8] Zhengfang Duanmu, Wentao Liu, Diqi Chen, Zhuoran Li, Zhou Wang, Yizhou Wang, and Wen Gao. 2023. A bayesian quality-of-experience model for adaptive streaming videos. *ACM TOMM* 18, 3s (2023), 1–24.

[9] Zhengfang Duanmu, Wentao Liu, Zhuoran Li, Diqi Chen, Zhou Wang, Yizhou Wang, and Wen Gao. 2020. The Waterloo Streaming Quality-of-Experience Database-IV. *IEEE Dataport* (2020).

[10] Zhengfang Duanmu, Kede Ma, and Zhou Wang. 2017. Quality-of-experience of adaptive video streaming: Exploring the space of adaptations. In *ACM MM*. 1752–1760.

[11] Zhengfang Duanmu, Abdul Rehman, and Zhou Wang. 2018. A quality-of-experience database for adaptive video streaming. *IEEE TBC* 64, 2 (2018), 474–487.

[12] Zhengfang Duanmu, Kai Zeng, Kede Ma, Abdul Rehman, and Zhou Wang. 2016. A quality-of-experience index for streaming video. *IEEE JSTSP* 11, 1 (2016), 154–166.

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *CVPR*. 6202–6211.

[14] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075* (2024).

[15] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010* (2023).

[16] Deepti Ghadiyaram, Janice Pan, and Alan C Bovik. 2017. A subjective and objective study of stalling events in mobile streaming videos. *IEEE TCSVT* 29, 1 (2017), 183–197.

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

[18] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. 2024. VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation. In *EMNLP*. 2105–2123.

[19] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 2017. The Konstanz natural video database (KoNViD-1k). In *QoMEX*. 1–6.

[20] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP* 29 (2020), 4041–4056.

[21] Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. 2024. AesExpert: Towards Multi-modality Foundation Model for Image Aesthetics Perception. *arXiv preprint arXiv:2404.09624* (2024).

[22] Ziheng Jia, Xiongkuo Min, Wei Sun, and Guangtao Zhai. 2024. Continuous and overall quality of experience evaluation for streaming video based on rich features exploration and dual-stage attention. *IEEE TCSVT* (2024).

[23] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. 2023. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *CVPR*. 10041–10051.

[24] Jonathan Kua, Grenville Armitage, and Philip Branch. 2017. A survey of rate adaptation techniques for dynamic adaptive streaming over HTTP. *IEEE Communications Surveys & Tutorials* 19, 3 (2017), 1842–1866.

[25] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125* (2023).

[26] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. 2022. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE TCSVT* 32, 9 (2022), 5944–5958.

[27] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* (2024).

[28] Chunyi Li, Haoning Wu, Hongkun Hao, Zicheng Zhang, Tengchaun Kou, Chaofeng Chen, Lei Bai, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. 2024. G-Refine: A General Quality Refiner for Text-to-Image Generation. *arXiv preprint arXiv:2404.18343* (2024).

[29] Dingquan Li, Tingting Jiang, and Ming Jiang. 2019. Quality assessment of in-the-wild videos. In *ACM MM*. 2351–2359.

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. 19730–19742.

[31] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*. 22195–22206.

[32] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *QoMEX*. 1–3.

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *NIPS* 36 (2024).

[34] Hongbo Liu, Mingda Wu, Kun Yuan, Ming Sun, Yansong Tang, Chuanchuan Zheng, Xing Wen, and Xiu Li. 2023. Ada-dqa: Adaptive diverse quality-aware feature acquisition for video quality assessment. In *ACM MM*. 6695–6704.

[35] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. 2018. End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks.. In *ACM MM*. 546–554.

[36] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *ICML*. 22631–22648.

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

[38] Anish Mittal, Michele A Saad, and Alan C Bovik. 2015. A completely blind video integrity oracle. *IEEE TIP* 25, 1 (2015), 289–300.

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

[40] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. 2016. CVD2014—A database for evaluating no-reference video quality assessment algorithms. *IEEE TIP* 25, 7 (2016), 3073–3086.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. 8748–8763.

[42] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

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

[44] Zeina Sinno and Alan Conrad Bovik. 2018. Large-scale study of perceptual video quality. *IEEE TIP* 28, 2 (2018), 612–627.

[45] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. 2022. A deep learning based no-reference quality assessment model for ugc videos. In *ACM MM*. 856–865.

[46] Wei Sun, Wen Wen, Xiongkuo Min, Long Lan, Guangtao Zhai, and Kede Ma. 2024. Analysis of video quality datasets via design of minimalistic video quality models. *IEEE TPAMI* (2024).

[47] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. 2021. UGC-VQA: Benchmarking blind video quality assessment for user generated content. *IEEE TIP* 30 (2021), 4449–4464.

[48] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. 2021. RAPIQUE: Rapid and accurate video quality prediction of user generated content. *IEEE OJSP* 2 (2021), 425–440.

[49] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing

vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).
[50] Yilin Wang, Sasi Inguva, and Balu Adsumilli. 2019. YouTube UGC dataset for video compression research. In *MMSP*. 1–5.
[51] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. 2021. Rich features for perceptual quality assessment of UGC videos. In *CVPR*. 13435–13444.
[52] Wen Wen, Mu Li, Yabin Zhang, Yiting Liao, Junlin Li, Li Zhang, and Kede Ma. 2024. Modular Blind Video Quality Assessment. In *CVPR*. 2763–2772.
[53] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2022. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *ECCV*. 538–554.
[54] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. 2023. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE TPAMI* (2023).
[55] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *CVPR*. 20144–20154.
[56] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. 2024. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *CVPR*. 25490–25500.
[57] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. 2023. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090* (2023).
[58] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. 2025. Towards open-ended visual quality comparison. In *ECCV*. 360–377.
[59] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).
[60] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840* (2024).
[61] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*. 13040–13051.
[62] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. 2021. Patch-vq:'patching up'the video quality problem. In *CVPR*. 14019–14029.
[63] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. 2023. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. *arXiv preprint arXiv:2312.08962* (2023).
[64] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *CVPR*. 11975–11986.
[65] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 2023. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *CVPR*. 14071–14081.
[66] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video Instruction Tuning With Synthetic Data. *arXiv preprint arXiv:2410.02713* (2024).
[67] Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. 2025. Q-Bench-Video: Benchmarking the Video Quality Understanding of LMMs. In *CVPR*.
[68] Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. 2024. Q-bench: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE TPAMI* (2024).
[69] Zicheng Zhang, Wei Wu, Wei Sun, Danyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. 2023. MD-VQA: Multi-dimensional quality assessment for UGC live videos. In *CVPR*. 1746–1755.
[70] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Baixuan Zhao, Xiaohong Liu, and Guangtao Zhai. 2024. Quality Assessment in the Era of Large Models: A Survey. *arXiv preprint arXiv:2409.00031* (2024).
[71] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding. *arXiv preprint arXiv:2406.04264* (2024).