Scaling-up Perceptual Video Quality Assessment

Ziheng Jia[1*], Zicheng Zhang[2*], Xiaorong Zhu[1], Chunyi Li[1], Jinliang Han[1], Xiaohong Liu[1], Guangtao Zhai[1,2], Xiongkuo Min[1†]

[1]Shanghai Jiao Tong University
[2]Shanghai Artificial Intelligence Laboratory
jzhws1@sjtu.edu.cn

Abstract

The data scaling law has significantly enhanced large multi-modal models (LMMs) performance across various downstream tasks. However, in the domain of perceptual video quality assessment (VQA), the potential of data scaling remains unprecedented due to the scarcity of labeled resources and the insufficient scale of datasets. To address this, we propose **OmniVQA**, a framework designed to efficiently build high-quality, machine-dominated synthetic multi-modal instruction databases (MIDBs) for VQA. We then scale up to create **OmniVQA-Chat-400K**, the largest dataset in the VQA field concurrently. Our focus is on the technical and aesthetic quality dimensions, with abundant in-context instruction data to provide fine-grained VQA knowledge. Additionally, we build the **OmniVQA-MOS-20K** dataset to enhance the model's quantitative quality rating capabilities. We then introduce a **complementary** training strategy that effectively leverages the knowledge from datasets for different tasks. Furthermore, we propose the **OmniVQA-FG (fine-grain)-Benchmark** to evaluate the fine-grained performance of models. Our results demonstrate that our models achieve state-of-the-art performance in both tasks.

Code — https://github.com/jzhws/Omni-VQA

Introduction

The perceptual video quality assessment (VQA) currently focuses on two key tasks: quality rating and quality understanding. Quality rating refers to assigning a precise score to a video aligning with its human-labeled mean opinion score (MOS), while quality understanding involves providing qualitative feedback and analysis on the video's quality. Recent advancements in large language models (LLMs) and large multi-modal models (LMMs) demonstrate the significant impact of data scaling on various tasks (Wang, Xu, and Ren 2024; Zhang et al. 2024a; Chen et al. 2024b; Zhang et al. 2024b; Islam et al. 2024; Mangalam, Akshulakov, and Malik 2023; Zhang et al. 2025c,b). This motivates a key hypothesis: *scaling up VQA data can potentially improve model performance in perceptual VQA* . Despite this potential, existing VQA datasets may struggle to fully leverage the data scaling effects due to their insufficient data scale (shown in the supplementary materials (*supp.*))). Here, we would like to raise a fundamental question:

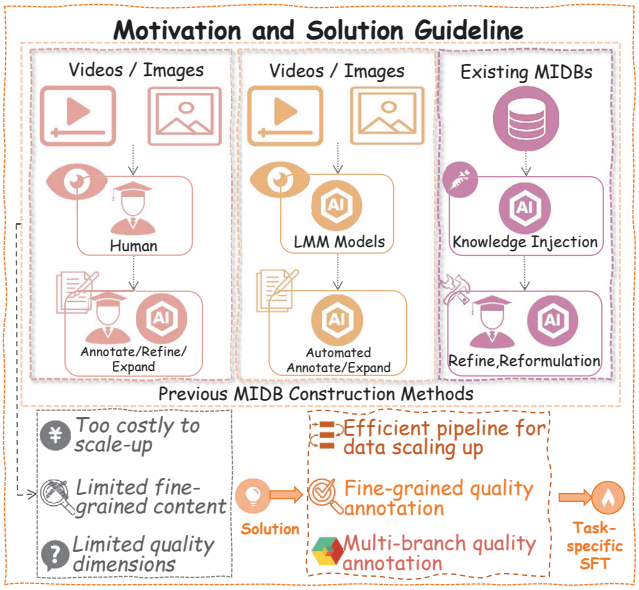

Figure 1: Existing MIDBs for VQA excessively rely on manual annotations or existing datasets and lack comprehensive and diverse annotation dimensions. To address this, we develop the OmniVQA datasets and models.

Why scaling up VQA data is a significant challenge? Perceptual VQA is a task that **mimics human perception**, which inherently requires significant expert-level human involvement. As a consequence, the construction of large-scale datasets becomes not only resource-intensive but also time-consuming. Previous multi-modal instruction databases (**MIDBs**) for VQA like (Wu et al. 2024a; Huang et al. 2024) utilize LLMs, such as *ChatGPT* (Radford et al. 2018), to augment the human annotations. Nonetheless, the overall information gain remains marginal. Furthermore, this annotation methodology is deeply reliant on extensive human resources, thereby hindering its scalability. To tackle these challenges, we introduce **OmniVQA**, an effective framework to build perceptual MIDB for VQA

*These authors contributed equally.
†Corresponding author.

through a machine annotation paradigm with human-in-the-loop, culminating in the creation of the largest MIDB for VQA to date, the **OmniVQA-Chat-400K**. This framework is built with 3 branches that facilitate easy scalability while maintaining high data quality. In addition, we propose **OmniVQA-MOS-20K**, a large-scale human-labeled VQA dataset specially for video quality rating tasks.

Considering the intrinsic interconnection between quality rating and understanding tasks, we further propose a **complementary** training strategy to effectively harness the knowledge in datasets from both tasks, facilitating the training of LMMs. First, we train the model on the dataset on one of the tasks and then finetune it using the remaining data. This approach enables the model to integrate knowledge from both tasks effectively.

Currently, no benchmark is available for fine-grained, in-context video quality understanding. To thoroughly evaluate the models' in-context capabilities, we introduce the **OmniVQA-FG (fine-grain)-Benchmark**, which is meticulously designed to assess the LMM's performance in local-spatiotemporal, fine-grained quality understanding tasks. The motivation guideline of OmniVQA is shown in Fig. 1.

Our contributions are threefold:

- We propose an effective data collection pipeline that supports the creation of large-scale, high-quality MIDBs for VQA, and we then develop the OmniVQA-Chat-400K. We also propose the OmniVQA-MOS-20K, a large-scale video subjective scoring dataset.
- The proposed OmniVQA models achieve state-of-the-art (SOTA) performance by training LMMs with the task-specific complementary training strategy for both quality rating and understanding tasks.
- We introduce the OmniVQA-FG-Benchmark, a comprehensive benchmark designed to evaluate fine-grained spatiotemporal quality understanding performance in both synthetic and real-world scenarios.

Related Works

Perceptual video quality assessment

Perceptual VQA initially focuses on quantitative video quality rating (Min et al. 2024), primarily aiming to fit the human-labeled MOSs in public datasets such as (Ghadiyaram et al. 2017; Sinno and Bovik 2018; Wang, Inguva, and Adsumilli 2019; Ying et al. 2021). Existing works on this task include handcrafted-feature-based methods (Mittal, Soundararajan, and Bovik 2012; Mittal, Moorthy, and Bovik 2012; Tu et al. 2021; Duanmu et al. 2023), deep neural network (DNN)-based approaches (Li et al. 2022; Sun et al. 2022, 2024; Wen et al. 2024; Wu et al. 2023a,b,c), and LMM-based models (Wu et al. 2024b; Ge et al. 2024; Jia et al. 2024). Recently, video quality understanding has become a new emerging research field. *Q-bench-video* (Zhang et al. 2025a) is the first comprehensive benchmark for evaluating the capabilities of LMMs in general video quality understanding tasks. *VQA^2-Assistant* (Jia et al. 2024) is the first LMM capable of video quality understanding and chatting. However, there is almost no work addressing fine-grained tasks such as spatiotemporal local distortions retrieval, thereby leaving significant room for further research.

In the field of LMMs for visual quality assessment, there have been various approaches to constructing MIDBs. They can be broadly divided into 3 categories:

1. **Human annotators as perceivers (mostly used)** (Wu et al. 2024a; Jia et al. 2024; You et al. 2024b; Huang et al. 2024; Zhou et al. 2024): Human annotators directly serve as perceivers of videos/images, and the data is obtained directly through manual annotations and refinement/rewrite by LLMs. The major drawback is the significant cost of human labor and time.
2. **General LMMs as perceiver and annotator (distillation approach)** (You et al. 2024a): General LMMs (e.g., *GPT-4o* (Achiam et al. 2023), *Gemini-1.5* (Team et al. 2024)) directly serve as perceivers and annotators. Its primary limitation lies in the sub-optimal data quality, constrained by distillation from teacher models that are not proprietary within this domain.
3. **Knowledge injection** (Wu et al. 2024b; Chen et al. 2024d; Wu et al. 2024c; Chen et al. 2024a): These approaches involve injecting task-specific knowledge into existing MIDBs to adapt them for new downstream tasks. The deficiency is that the scale of the new MIDB is heavily contingent upon the original MIDB, thus limiting the efficient expansion of the data scale.

These MIDBs also share the following common issues. First, these datasets primarily focus on the annotation of the technical or aesthetic quality of images/videos while lacking comprehensive annotations integrating multiple quality dimensions and factors. Moreover, the majority of these MIDBs (Wu et al. 2024a; Jia et al. 2024; You et al. 2024b; Huang et al. 2024; Zhou et al. 2024; You et al. 2024a; Wu et al. 2024b,c) emphasize an overall description of image/video quality but lack spatiotemporal fine-grained annotation and evaluation. These issues have just become the inspiration for our work.

The OmniVQA-Chat-400K

Driven by the above motivations, we construct the OmniVQA-Chat-400K, currently the most expansive and in-depth perceptual MIDB for VQA. The construction pipeline is illustrated in Fig. 2 and detailed in *supp.* with abundant examples, and the statistical information is also shown in *supp.*. We conduct a user study to evaluate the annotation quality of samples, which is presented in *supp.*.

Candidate video pool construction

We select $100,000$ videos from a large-scale user-generated content (UGC) video dataset (Chen et al. 2024c) to serve as the candidate video pool for subsequent selection. We impose the constraint that the length of all candidate videos must be in the range of $[3, 15)$ seconds. We then utilize 4 state-of-the-art objective video quality rating methods (Wu et al. 2023a,b, 2024b; Sun et al. 2024) to label the objective quality for the candidate videos. To ensure consistency in the scale of scores across different rating methods, the

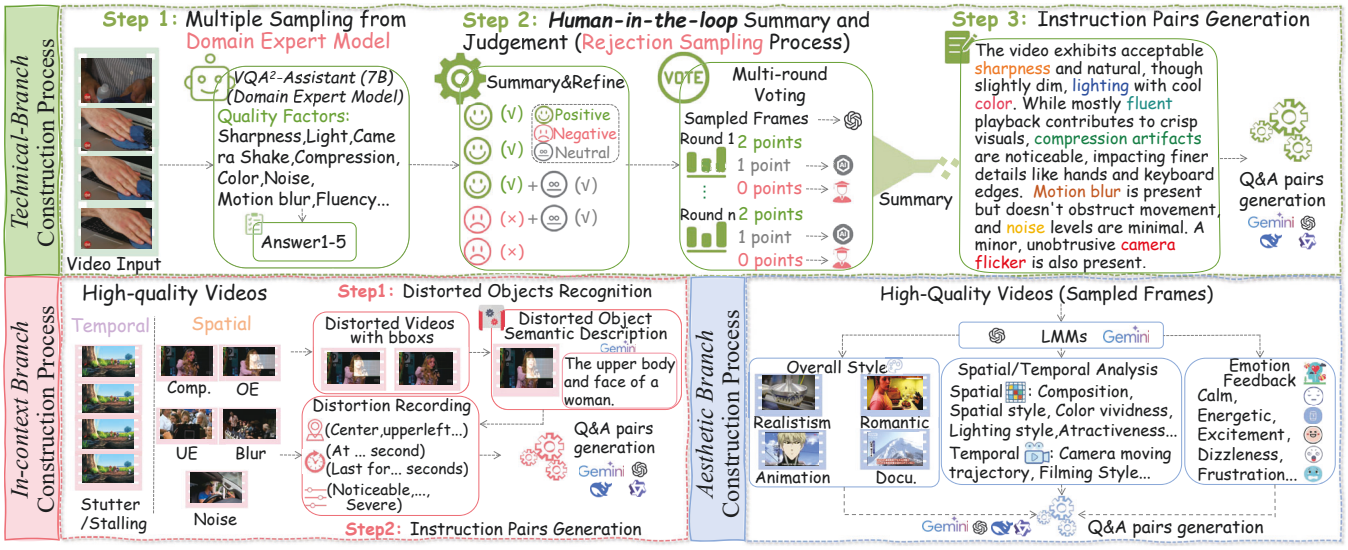

Figure 2: Data construction pipeline of **OmniVQA-Chat-400K**. Comp., OE, UE, and Docu. denote Compression, Over-exposure, Under-exposure, and Documentary.

scores are first normalized to the range of $[0, 100)$ and then averaged for each video to determine its objective quality.

The technical branch

The **technical branch** is the primary component of the OmniVQA-Chat-400K. The $55,128$ videos in this branch are randomly selected from the candidate video pool to ensure video content and quality diversity. Each video is annotated across 8 quality factors: **sharpness**, **light**, **compression**, **color**, **noise**, **fluency**, **motion blur**, and **camera shake**. We employ an aggregation strategy, where annotations for each quality factor are initially performed individually and subsequently combined into a video-level description (shown in the **upper** part of Fig. 2).

We propose a novel paradigm based on **rejection sampling (RS)** for the annotation of each quality factor. The process begins by performing multiple coarse annotations using expert LMM for VQA. This process forms the **suggested distribution**. Subsequently, unlike the common practices, we do not only select samples based on certain criteria. Instead, we employ a heuristic method to fully utilize the effective information in all the samples. This method filters and summarizes the N samples obtained through the first sampling with reasoning LLMs, and a voting mechanism using general LMMs is applied to decide the next step. In this case, the general LMMs act as a **judger** rather than an annotator (the deficiencies of *GPT* direct annotation are in *supp.*). We argue that substituting the challenging direct annotation task with the relatively simpler voting decision for general LMMs offers a more practical distillation approach. This method is more effective in guiding the general LMMs toward generating more accurate and definitive responses.

Multiple coarse annotations. We use *VQA^2-Assistant (7B)* (Jia et al. 2024) as the expert model for sampling. We set $N = 5$, meaning that for each quality factor, we pose 5 questions with the same fundamental meaning but varying sentence structures and record the model's responses.

LLM summary. We use *Openai-o1* (Jaech et al. 2024) to conduct the LLM summary process. The information in the responses obtained before is categorized into 3 types: (1) **Positive answers** are those that have similar meanings in the responses and appear in at least 3 of the 5 responses. These answers are then merged into the summary (2)**Negative answers**, which are contradicted in meaning to the positive answers, are excluded from the summary. (3) If any responses contain additional **neutral** information, it should be included in the summary. For example, for the quality factor "**sharpness**" and the question *"How is the sharpness of this video?"*, if the 5 responses are: *"Poor"* (positive), *"Relatively poor"* (positive), *"Poor, with degraded facial details"* (positive with neutral), *"Good, however, the facial details are slightly lost"* (negative with neutral), *"Excellent"* (negative). The summarized response would be: *"The sharpness is poor with degraded human facial details."*

Voting and post-processing. We then use *GPT-4o* to decide the post-processing method. We input the keyframe sequence sampled from the video (1fps) and the prior summary of each quality factor. (Due to the inability of the input keyframe sequences to capture stuttering, we implement a special process for the **fluency** dimension in *supp.*. We then prompt *GPT* to conduct 3 voting rounds for each quality factor, assessing the accuracy and relevance of the provided summaries and assigning a score in $(2, 1, 0)$. Then, the post-processing method is determined and implemented on the given quality factor summaries. If any round of voting results in a score of 0, human experts must intervene to decide between the original input summary and the *GPT*-modified summary for that round. The detailed scoring criteria and the corresponding post-processing methods, including the sub-

jective experiment, are detailed in the *supp.*.

Instruction pairs generation. We use *o1* to merge the annotations of all quality factors, yielding a video-level quality summary. Based on this summary, we ask the model to generate 3 question-answer (Q&A) pairs related to quality factors. To fully leverage the quality summary, we add an extra question for each video: *"Please describe the overall quality of this video, please evaluate as many quality factors as possible."* The video-level quality summary is then provided as the answer to this question.

The in-context branch

The **in-context branch** is designed to augment the model's ability to identify fine-grained local spatiotemporal quality issues in videos. To minimize the influence of inherent distortions, we choose $6,500$ source videos from the candidate pool with objective quality labels above 70. The annotation process is presented in the **lower left** part in Fig. 2.

Synthetically distorted video generation. We manually add local spatiotemporal distortions to the selected videos. Spatial distortions include **overexposure**, **underexposure**, **blur**, **compression artifacts**, and **noise**, while temporal distortion refers to **video stuttering**.

Spatial distortions are randomly added to a rectangular region within the frame, covering $1/4$ of the frame area. The duration of spatial distortion is randomly assigned a value between 1 and 3 seconds, with the starting time of the synthetic distortion also determined as an integer. The intensity of the distortion is categorized into 3 levels: **noticeable**, **relatively severe**, and **severe**. The 5 distortion types are added sequentially for each source video, with the number of distortions fixed at 1. Finally, we record the spatial distortion's start time, duration, distortion type, intensity, and location. For the stuttering, we remove the entire second following a randomly selected integer second and then duplicate the frames from that for an additional second, thereby creating a **frame freeze** effect. Each source video ultimately has 1 to 3 randomly inserted video stutter events. More detailed video generation processes are provided in the *supp.*. We manually filter out the generated videos in which the synthetic spatiotemporal distortion is not visibly perceptible.

Distorted objects recognition and description. To enhance the LMM's ability to capture and describe the semantics of the spatial distortion regions, we add a highlighted bounding box (bbox) around the distorted rectangle area in the generated videos. The keyframes of the distorted videos with the bbox are then input into the *Gemini-1.5-Pro*, which is tasked with describing the main semantic objects within the bbox with the following criteria: a semantic object is deemed valuable for annotation only if *it is fully contained within the bbox and occupies the primary region, exhibits a distinct contrast from the surrounding area, and remains within the bbox during the distorted period*.

Instruction pairs generation. We then generate 5 instruction Q&A pairs for each spatial and temporal distorted video based on the summarized video distortion information. The Q&A pairs must focus entirely on the spatiotemporal local distortions of the video. If the video information contains the annotated distorted semantic object, at least one of the generated Q&A pairs must be related to this. Similar to the technical branch, each video also includes a summarizing question, which is: *"Please describe the information about the local distortions of the video."* The answer to this question is the summarized video's local distortion information.

The aesthetic branch

The aesthetic branch aims to enhance the diversity and comprehensiveness of the MIDB. To prevent technical distortions from affecting the extraction of aesthetic features, we also select $7,728$ videos with objective quality scores above 70. Then, we directly input keyframes sampled from the videos into *GPT-4o* to annotate the aesthetic quality. Inspired by (Huang et al. 2024), the annotation is conducted from 3 aspects (illustrated in the **lower right** of Fig. 2):

1. **Aesthetic Style**: Label the aesthetic style of the video.
2. **Spatiotemporal Analysis**: Perform a detailed aesthetic analysis from the spatial and temporal perspectives.
3. **Emotional Repercussions**: Provide the emotional experience that the video may evoke in viewers.

The machine annotation for each video is then summarized, and 6 instruction Q&A pairs are derived from this.

Each video also includes a summarizing Q&A pair with the question *"Please give a detailed description of the aesthetic effects of the video."* with all the annotated aesthetic information as the corresponding answer.

The OmniVQA-MOS-20K

To enhance the model's performance on quality rating, we construct a large-scale UGC video subjective scoring dataset comprising $20,000$ videos and over $300,000$ annotated scores. The videos are selected from a distinct candidate pool of $100,000$ videos in (Chen et al. 2024c), separate from the video pool mentioned above. . Each video is annotated with an objective quality label using the same procedure. The video selection aims to ensure a well-balanced distribution of objective quality labels. To this end, we first convert the objective quality levels of all videos to be aligned with the MOSs of LSVQ (train) (Ying et al. 2021). Then, we ensure that the distribution of objective quality levels for the selected videos mirrors that of LSVQ (train). Specifically, for the 5 objective quality levels (determined by their objective label) —excellent (80-100), good (60-80), fair (40-60), poor (20-40), and low (0-20)— the proportion of videos in each quality level must align with LSVQ (train).

During the subjective experiment, we implement a hidden reference supervision strategy where the objective quality level for each video serves as the reference but is not visible to the human annotators. Any human score deviating by two or more levels from the reference is rejected, and rescoring is requested. This method effectively minimizes human annotation bias, especially in scenarios with limited expert-level resources. The distribution of human-labeled scores (averaged for each video) is shown in *supp.*. Finally, we randomly split the dataset into the **OmniVQA-MOS-20K (train)** and

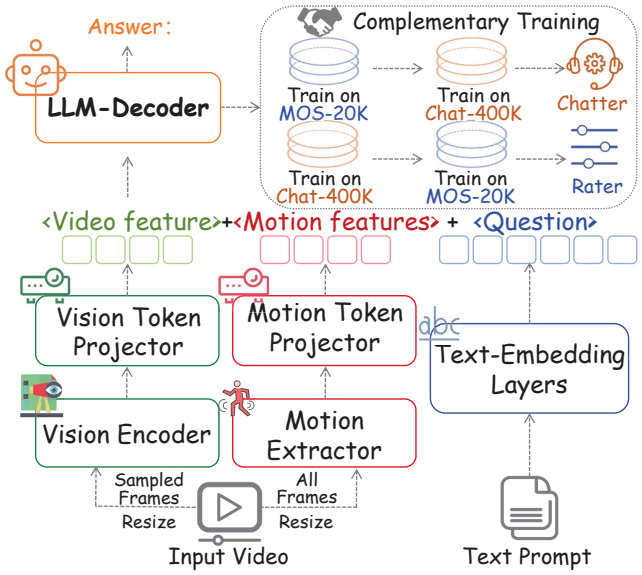

Figure 3: Illustration of OmniVQA models and the complementary training strategy.

the **OmniVQA-MOS-20K (test)** with an $80\% : 20\%$ ratio for subsequent experiments. Since the model training still focuses on text-based tasks, the videos and quality labels in OmniVQA-MOS-20K (train) need to be converted into Q&A pairs. The specific format is as follows: *Q: How would you rate the quality of this video?* The answer is the human-labeled quality rating converted into the quality level.

The OmniVQA-FG-Benchmark

To evaluate the model's performance in fine-grained tasks, we construct the **OmniVQA-FG-Benchmark**. This bench contains $1,200$ Q&A pairs and consists of both machine (*GPT*) and human annotations. The statistical information about the benchmark is given in *supp.*. The entire benchmark focuses on the following 3 core quality concerns:

- **Spatial (S)**: Spatial quality primarily concerns quality issues within specific local regions, particularly focusing on specific areas or semantic objects in the video.
- **Temporal (T)**: Temporal quality concentrates on temporal quality issues, especially issues related to specific timepoints or periods.
- **Spatiotemporal (ST)**: Spatiotemporal quality encompasses both spatial and temporal dimensions, focusing on quality issues that arise in local areas or at semantic objects during specific timepoints or intervals.

Detailed examples are shown in the *supp.*.

For the **machine annotation**, we select 900 videos (not within the training data) with objective quality labels higher than 70 in the candidate videos pool and manually synthesize varying spatiotemporal distortions at different locations and periods. The annotations concentrate on fixed-form descriptions to evaluate the models' ability to localize spatiotemporal distortions. The questions in the machine-annotated part are all binary and multi-choice (single-answer) questions. For the **human annotation**, we select $1,000$ videos in the candidate video pool with an objective quality score below 70 to ensure that the videos contain abundant quality issues to be annotated. Human annotations exploit the analysis and semantic descriptions of spatiotemporal local quality. It follows a flexible approach, incorporating both multi-choice questions and open-ended questions, ensuring a comprehensive capture and annotation of video quality issues. Annotation examples are shown in *supp.*.

The OmniVQA Models

After constructing the OmniVQA-Chat-400K and OmniVQA-MOS-20K, the *VQA^2-Stage-1* model (Jia et al. 2024) is employed as the base model for supervised fine-tuning (SFT). The base model consists of the *SigLip* (Zhai et al. 2023) vision encoder, the *SlowFast-R50* (Feichtenhofer et al. 2019) motion extractor, and the *Qwen-2* (Yang et al. 2024) LLM model (shown in Fig. 3). The text tokens from the prompt, the vision tokens from the video keyframes, and the motion tokens from the entire video are **interleaved** into a semantically ordered sequence, which is then input into the LLM for text generation. Through different SFT processes, we obtain 2 specialized models: the ***rater***, which focuses on perceptual VQA quantitative rating tasks, and the ***chatter***, which specializes in quality understanding and question-answering tasks.

Complementary training strategy. We posit that the role of the LLM part varies between quantitative rating tasks and quality understanding tasks. In the former, the LLM functions primarily as an effective **regressor**, whereas in the latter, it learns to navigate the complex semantic relationships between different quality factors and modalities. Consequently, we argue that random mixing of training data from these tasks may undermine the LLM's ability to effectively perform on each task, as the divergent training objectives could hinder its capacity to focus on the specific learning goals. However, from a pre-training perspective, these two tasks are perfectly complementary. Firstly, the sequential training process mitigates potential confusion regarding the model's learning objectives. More importantly, the intrinsic relationship between the knowledge of the two tasks suggests that the datasets can provide valuable **prior information** for one another, thus making them well-suited to serve as mutually beneficial pre-training components. The process of complementary training is also depicted in Fig. 3.

Experiments

We conduct a detailed evaluation of our models on video quality rating and understanding tasks to verify the effects of scaled-up datasets and our training strategy. In addition, we perform comprehensive supplementary experiments (detailed in *supp.*) to investigate some key factors.

Datasets	LSVQ(1080p)		LSVQ(test)		LIVE-VQC		KoNViD-1k		YT-UGC		MOS-20K	
Metrics	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Simple-VQA (Sun et al. 2022)	0.760	0.805	0.870	0.868	0.755	0.793	0.826	0.820	0.850	0.845	0.813	0.809
BVQA (Li et al. 2022)	0.747	0.785	0.870	0.861	0.795	0.814	0.795	0.817	0.845	0.847	0.825	0.813
FAST-VQA (Wu et al. 2023a)	0.765	0.793	0.880	0.871	*0.830*	0.822	0.869	0.870	0.828	0.849	0.792	0.783
Dover (Wu et al. 2023b)	0.797	0.821	0.893	0.892	**0.835**	**0.857**	0.885	0.879	0.855	0.861	0.828	0.832
Modular-VQA (Wen et al. 2024)	0.810	0.834	*0.897*	*0.895*	0.803	0.839	0.876	*0.887*	0.862	**0.878**	**0.843**	*0.835*
q-align-VQA (7B) (Wu et al. 2024b)	0.758	0.833	0.883	0.882	0.777	0.813	0.865	0.876	0.811	0.830	0.820	0.831
q-align-onealign (7B)	0.803	0.836	0.888	0.885	0.773	0.829	0.876	0.878	0.831	0.847	0.829	0.826
VQA^2-UGC-Scorer (7B)	0.782	*0.837*	0.897	0.885	0.798	0.830	*0.894*	0.884	0.818	0.827	0.785	0.773
Chatter (7B) (400K)	**0.816**	0.821	0.889	0.856	0.822	0.846	0.882	0.835	*0.859*	0.839	0.810	0.788
Rater (7B)	*0.815*	**0.838**	**0.902**	**0.905**	0.826	*0.855*	**0.895**	**0.900**	**0.872**	*0.873*	*0.837*	**0.837**

Table 1: Performance on quality rating tasks. The **OmniVQA-MOS-20K(test)** is presented as "MOS-20K" in short. [Per column: highest in **bold**, second in *italic*]

Categories	***Q-bench-video-test*** (900 ***questions)***							
LMMs	*Binary*	*Multi.*	*Open*	*Tech.*	*Aes.*	*Temp.*	*AIGC*	*Overall*
mplug-owl3 (7B) (Ye et al. 2024)	56.90%	57.14%	42.88%	53.40%	61.85%	50.34%	45.34%	52.06%
Internvl2 (40B)	52.53%	43.21%	35.13%	42.54%	52.13%	46.43%	42.55%	43.44%
LLaVA-onevision (7B) (Li et al. 2025)	57.58%	48.78%	32.12%	44.98%	50.95%	45.07%	44.72%	45.83%
LLaVA-onevision (72B)	52.19%	54.36%	34.34%	45.62%	54.74%	50.00%	46.58%	46.61%
Qwen2-vl (7B) (Wang et al. 2024)	50.84%	55.75%	34.49%	46.03%	56.40%	50.17%	39.75%	46.67%
Qwen2-vl (72B)	61.62%	**66.90%**	39.24%	55.19%	*63.03%*	52.38%	50.93%	55.44%
Qwen2.5-vl (7B) (Bai et al. 2025)	52.53%	49.48%	38.77%	46.68%	58.53%	46.09%	41.61%	46.72%
Qwen2.5-vl (72B)	62.73%	56.90%	41.31%	52.93%	62.49%	52.15%	46.79%	52.35%
GPT-4o (24-11-20) (Achiam et al. 2023)	60.61%	50.17%	**45.25%**	50.89%	**63.27%**	52.04%	48.45%	51.89%
Gemini-1.5-pro (Team et al. 2024)	56.80%	43.29%	39.26%	44.57%	53.94%	54.64%	44.21%	46.52%
Gemini-2.0-flash	56.23%	46.34%	43.57%	47.73%	58.77%	48.64%	**55.90%**	49.33%
VQA^2-Assistant (7B) (Jia et al. 2024)	*67.12%*	59.93%	39.56%	*55.19%*	56.87%	**57.99%**	43.79%	*55.56%*
Chatter (7B) (400K)	**68.35%**	*63.76%*	*44.46%*	**58.10%**	60.66%	*54.93%*	*52.17%*	**58.50%**

Table 2: Evaluation results on the *Q-bench-video-test*.

Categories	***Question Types***			***Machine Annotated***				***Human Annotated***				***Overall***
LMMs	*Binary*	*Multi.*	*Open*	*Mach. S*	*Mach. T*	*Mach. ST*	*Overall*	*Human S*	*Human T*	*Human ST*	*Overall*	
mplug-owl3 (7B)	25.05%	30.93%	34.74%	21.50%	17.50%	40.00%	24.13%	37.05%	41.82%	41.18%	39.50%	28.88%
Internvl2 (8B)	26.59%	24.29%	24.03%	18.00%	15.50%	16.00%	16.93%	30.22%	32.27%	50.00%	34.33%	25.25%
Internvl2 (40B)	41.13%	23.31%	22.19%	30.75%	26.00%	23.00%	27.33%	31.93%	31.55%	32.20%	31.83%	30.88%
Internvl2.5 (8B)	21.97%	37.38%	21.10%	29.75%	14.50%	28.00%	25.33%	29.14%	27.73%	30.39%	28.83%	28.62%
LLaVA-onevision (7B)	31.60%	29.98%	27.60%	25.00%	17.00%	33.33%	24.53%	33.45%	40.00%	37.25%	36.50%	30.38%
LLaVA-onevision (72B)	36.03%	30.36%	24.03%	30.75%	22.50%	33.33%	29.07%	30.94%	38.64%	40.20%	35.33%	32.00%
Qwen2-vl (7B)	26.59%	23.15%	22.73%	20.25%	18.50%	24.67%	20.67%	26.26%	33.64%	42.16%	31.67%	24.58%
Qwen2-vl (72B)	33.14%	*40.80%*	27.27%	32.50%	26.00%	34.00%	31.07%	39.21%	44.55%	42.16%	41.67%	35.75%
Qwen2.5-vl (7B)	23.12%	24.29%	30.84%	16.50%	5.50%	21.33%	14.53%	33.09%	41.36%	43.14%	37.83%	24.62%
Qwen2.5-vl (72B)	26.20%	24.29%	24.84%	19.00%	7.50%	25.33%	17.20%	32.53%	43.12%	32.61%	36.38%	25.19%
GPT-4o (24-11-20)	48.55%	39.47%	33.12%	48.00%	37.00%	44.67%	44.40%	38.85%	45.91%	42.16%	42.00%	42.58%
VQA^2-Assistant (7B)	*76.11%*	37.19%	*37.99%*	*51.25%*	*60.00%*	*60.00%*	*55.33%*	*43.88%*	**50.45%**	*45.10%*	*46.50%*	*54.12%*
Chatter (7B) (400K)	**81.31%**	**57.31%**	**39.03%**	**56.75%**	**85.50%**	**84.00%**	**69.87%**	**44.52%**	*46.33%*	**56.52%**	**47.01%**	**65.32%**

Table 3: Evaluation results on the OmniVQA-FG-Bench, where "Mach." denotes "Machine", "S" denotes "Spatial", "T" denotes "Temporal", and "ST" denotes "spatiotemporal".

Categories	Quality Rating (SRCC / PLCC)					Overall			Fine-grain		
Training Strategy	*LSVQ(1080p)*	*LSVQ(test)*	*LIVE-VQC*	*KoNViD-1k*	*OmniVQA (test)*	*Tech.*	*Aes.*	*Overall*	*Machine*	*Human*	*Overall*
Direct	0.800 / 0.824	0.880 / 0.878	0.776 / 0.820	0.877 / 0.883	0.819 / 0.820	57.54%	**63.03%**	58.33%	68.00%	50.50%	64.28%
Mix	**0.817** / 0.836	0.898 / 0.896	0.822 / **0.860**	0.887 / 0.898	**0.840 / 0.838**	52.54%	56.32%	52.78%	66.47%	51.17%	62.58%
Complementary	0.815 / **0.838**	**0.902 / 0.905**	**0.826** / 0.855	**0.895 / 0.900**	0.837 / 0.837	**58.10%**	60.66%	**58.50%**	**69.87%**	**56.52%**	**65.32%**

Table 4: Ablation of training strategies. The **Overall** task is evaluated on the *q-bench-video (test)*. [Per column: best in **bold**]

Experimental setups

In all evaluations, we set almost unified system prompts for all LMM models, and the system prompts are illustrated in the *supp.*. We employ the complementary training strategy to train the models. All training is performed with full-parameter-tuning, with only 1 epoch trained on each dataset. The specific hyper-parameter configurations and model structures are also presented in the *supp.*.

Evaluation on quality rating tasks

We compare our models with several DNN-based (Sun et al. 2022; Li et al. 2022; Wu et al. 2023a,b; Wen et al. 2024) and LMM-based (Wu et al. 2024b; Jia et al. 2024) quality rating models on 6 datasets including the OmniVQA-MOS-20K (test). Apart from *q-align*, *VQA^2-UGC-Scorer*, and our models (using complementary training), all models are trained on the merged dataset of OmniVQA-MOS-20K (train) and LSVQ (train) (approximately $43,000$ videos). The evaluation metrics are the *Pearson Linear Correlation Coefficient* (PLCC) and *Spearman Rank Correlation Coefficient* (SRCC). We adopt the quality rating method used in (Wu et al. 2024b; Jia et al. 2024) during testing, which is detailed in the *supp.*. The performance of the models on all datasets is presented in Tab. 1. The experimental results show that the **rater** achieves *Top-3* performance across all 6 datasets. This demonstrates its superior performance in quality rating tasks. Since the **chatter** model is not a proprietary model for the rating task, its performance shows a noticeable decline, but it still indicates acceptable performance.

Evaluation on quality understanding tasks

As the primary task of our work, we conduct detailed video quality understanding evaluation experiments, which include both overall and fine-grained tasks. We also include 4 **real-world scenario case studies** in the *supp.* to visualize the functionality of the model.

The overall video quality understanding task is carried out on the *Q-bench-video-test*. Since our training process does not include multi-video comparison analysis, we remove questions involving multi-video quality issues. The evaluation question types include binary questions (*Binary*), multi-choice (single-answer) questions (*Multi.*), and open-ended questions (*Open*). The questions cover different quality concerns, including technical quality (*Tech.*), aesthetic quality (*Aes.*), temporal quality (*Temp.*), and AIGC video quality (*AIGC*). We select some of the latest open-source LMMs with video analysis capabilities (with varying parameter sizes), some proprietary LMMs, and the *VQA^2-Assistant* for comparison. To ensure a fair comparison, the input for each model is only related to its structure. For models without a motion extractor, we input the keyframe sequence obtained by sampling 1 frame per second from the video. For the *VQA^2-Assistant* and our chatter model, we additionally input the whole frame sequence of the video (resized to $(224 * 224)$) to the motion extractor. The experimental results are presented in Tab. 2.

The experimental results show that the **chatter** achieves the best *overall* performance on both the *test* and *dev* subsets, with outstanding performance in the *Tech., Aes., and Temp.* quality concerns. Although it does not outperform some of the most advanced LMMs (*GPT-4o* and *Qwen2-vl (72B)*) on some subcategories, the gap is relatively minimal. This demonstrates that in overall video quality understanding tasks, scaling up synthetic MIDBs with human-in-the-loop can yield superior SFT performance.

For the fine-grained video quality understanding task, we conduct experiments between the **chatter** and the comparison models on the OmniVQA-FG-Bench. Tab. 3 records the performance of each LMM on different subcategories of questions in this benchmark. Models specially trained on VQA tasks (*VQA^2-Assistant (7B)* and **chatter**) achieve significantly superior performance compared to general LMMs. Additionally, our **chatter** outperforms *VQA^2-Assistant* by a significant margin on machine annotation tasks and also achieves better performance on human annotation tasks. This demonstrates the importance of the abundant in-context data in the MIDB for improving the model's performance in spatiotemporal quality understanding tasks.

Discussions

Data scaling effects verification. Additionally, we validate the data-scaling effect by selecting subsets of data (ranging from 100k to 400k, with each subset having an equal distribution across 3 branches) from OmniVQA-Chat-400K for ablation. The performance on Q-bench-video-test and Omni-VQA-FG-Bench are shown in *supp.*. It is evident that the data-scaling effect appears in the 100k-400k range but gradually becomes marginal under 7B parameter size. Mixed training with human-annotated data further improves the performance on the *Q-bench-video (test)*.

Effects of complementary training. We compare 2 other training strategies: the *Direct* strategy, where the model is trained directly on the corresponding dataset for each task, and the *Mix* strategy, in which the 2 datasets are randomly mixed to train one unified model. The experimental results are presented in Tab. 4. The results show that our *Complementary* strategy outperforms the *Direct* strategy on all tasks. While the *Mix* strategy shows no significant difference from the *Complementary* strategy in the quality rating task, it exhibits a clear performance gap in the quality understanding tasks. The results showcase the rationale of the complementary training strategy.

Conclusion

We introduce **OmniVQA-Chat-400K**, the most comprehensive MIDB in the field of perceptual VQA, along with **OmniVQA-MOS-20K**, a large-scale, human-annotated dataset for video quality rating. By leveraging task-oriented, **complementary** training strategy on these datasets, our **chatter** and **rater** models exhibit superior performances in video quality understanding and rating tasks, respectively. Our work presents compelling insights into the feasibility of substituting the **human vision** with **machine vision** for **automated visual quality assessment**.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62522116, Grant 62271312 and Grant 62132006, and in part by STCSM under Grant 22DZ2229005. We also sincerely appreciate the Shanghai Artificial Intelligence Laboratory for computation resources supporting.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Chen, C.; Yang, S.; Wu, H.; Liao, L.; Zhang, Z.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2024a. Q-ground: Image quality grounding with large multi-modality models. In *ACM MM*, 486–495.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2024b. Sharegpt4v: Improving large multimodal models with better captions. In *ECCV*, 370–387.

Chen, T.-S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.-w.; Jeon, B. E.; Fang, Y.; Lee, H.-Y.; Ren, J.; Yang, M.-H.; et al. 2024c. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 13320–13331.

Chen, Z.; Zhang, X.; Li, W.; Pei, R.; Song, F.; Min, X.; Liu, X.; Yuan, X.; Guo, Y.; and Zhang, Y. 2024d. Grounding-IQA: Multimodal Language Grounding Model for Image Quality Assessment. *arXiv preprint arXiv:2411.17237*.

Duanmu, Z.; Liu, W.; Chen, D.; Li, Z.; Wang, Z.; Wang, Y.; and Gao, W. 2023. A bayesian quality-of-experience model for adaptive streaming videos. *ACM TOMM*, 18(3s): 1–24.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *ICCV*, 6202–6211.

Ge, Q.; Sun, W.; Zhang, Y.; Li, Y.; Ji, Z.; Sun, F.; Jui, S.; Min, X.; and Zhai, G. 2024. LMM-VQA: Advancing video quality assessment with large multimodal models. *arXiv preprint arXiv:2408.14008*.

Ghadiyaram, D.; Pan, J.; Bovik, A. C.; Moorthy, A. K.; Panda, P.; and Yang, K.-C. 2017. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE TCSVT*, 28(9): 2061–2077.

Huang, Y.; Sheng, X.; Yang, Z.; Yuan, Q.; Duan, Z.; Chen, P.; Li, L.; Lin, W.; and Shi, G. 2024. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *ACM MM*, 5911–5920.

Islam, M. M.; Ho, N.; Yang, X.; Nagarajan, T.; Torresani, L.; and Bertasius, G. 2024. Video recap: Recursive captioning of hour-long videos. In *CVPR*, 18198–18208.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jia, Z.; Zhang, Z.; Qian, J.; Wu, H.; Sun, W.; Li, C.; Liu, X.; Lin, W.; Zhai, G.; and Min, X. 2024. VQA ̃2: Visual Question Answering for Video Quality Assessment. *arXiv preprint arXiv:2411.03795*.

Li, B.; Zhang, W.; Tian, M.; Zhai, G.; and Wang, X. 2022. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE TCSVT*, 32(9): 5944–5958.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2025. Llava-onevision: Easy visual task transfer. *TMLR*.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NIPS*, 36: 46212–46244.

Min, X.; Duan, H.; Sun, W.; Zhu, Y.; and Zhai, G. 2024. Perceptual video quality assessment: A survey. *SCIS*, 67(11): 211301.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE SPL*, 20(3): 209–212.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Sinno, Z.; and Bovik, A. C. 2018. Large-scale study of perceptual video quality. *IEEE TIP*, 28(2): 612–627.

Sun, W.; Min, X.; Lu, W.; and Zhai, G. 2022. A deep learning based no-reference quality assessment model for ugc videos. In *ACM MM*, 856–865.

Sun, W.; Wen, W.; Min, X.; Lan, L.; Zhai, G.; and Ma, K. 2024. Analysis of video quality datasets via design of minimalistic video quality models. *IEEE TPAMI*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Tu, Z.; Wang, Y.; Birkbeck, N.; Adsumilli, B.; and Bovik, A. C. 2021. UGC-VQA: Benchmarking blind video quality assessment for user generated content. *IEEE TIP*, 30: 4449–4464.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Y.; Inguva, S.; and Adsumilli, B. 2019. YouTube UGC dataset for video compression research. In *IEEE MMSP*, 1–5.

Wang, Z.; Xu, G.; and Ren, M. 2024. LLM-Generated Natural Language Meets Scaling Laws: New Explorations and Data Augmentation Methods. *arXiv preprint arXiv:2407.00322*.

Wen, W.; Li, M.; Zhang, Y.; Liao, Y.; Li, J.; Zhang, L.; and Ma, K. 2024. Modular blind video quality assessment. In *CVPR*, 2763–2772.

Wu, H.; Chen, C.; Liao, L.; Hou, J.; Sun, W.; Yan, Q.; Gu, J.; and Lin, W. 2023a. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE TPAMI*, 45(12): 15185–15202.

Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2023b. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 20144–20154.

Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2023c. Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach. In *ACM MM*, 1045–1054.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Xu, K.; Li, C.; Hou, J.; Zhai, G.; et al. 2024a. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *CVPR*, 25490–25500.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2024b. Q-ALIGN: teaching LMMs for visual scoring via discrete text-defined levels. In *ICML*, 54015–54029.

Wu, H.; Zhu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Li, C.; Wang, A.; Sun, W.; Yan, Q.; et al. 2024c. Towards open-ended visual quality comparison. In *ECCV*, 360–377.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *ICLR*.

Ying, Z.; Mandal, M.; Ghadiyaram, D.; and Bovik, A. 2021. Patch-vq:'patching up'the video quality problem. In *CVPR*, 14019–14029.

You, Z.; Gu, J.; Li, Z.; Cai, X.; Zhu, K.; Dong, C.; and Xue, T. 2024a. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842*.

You, Z.; Li, Z.; Gu, J.; Yin, Z.; Xue, T.; and Dong, C. 2024b. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *ECCV*, 259–276.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *ICCV*, 11975–11986.

Zhang, B.; Liu, Z.; Cherry, C.; and Firat, O. 2024a. When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. In *ICLR*.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024b. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zhang, Z.; Jia, Z.; Wu, H.; Li, C.; Chen, Z.; Zhou, Y.; Sun, W.; Liu, X.; Min, X.; Lin, W.; et al. 2025a. Q-Bench-Video: Benchmarking the Video Quality Understanding of LMMs. *CVPR*.

Zhang, Z.; Wang, J.; Guo, Y.; Wen, F.; Chen, Z.; Wang, H.; Li, W.; Sun, L.; Zhou, Y.; Zhang, J.; Yan, B.; Jia, Z.; Xiao, J.; Tian, Y.; Zhu, X.; Zhang, K.; Li, C.; Liu, X.; Min, X.; Jia, Q.; and Zhai, G. 2025b. AIBench: Towards trustworthy evaluation under the 45° law. *Displays*, 103255.

Zhang, Z.; Wang, J.; Wen, F.; Guo, Y.; Zhao, X.; Fang, X.; Ding, S.; Jia, Z.; Xiao, J.; Shen, Y.; Zheng, Y.; Zhu, X.; Wu, Y.; Jiao, Z.; Sun, W.; Chen, Z.; Zhang, K.; Fu, K.; Cao, Y.; Hu, M.; Zhou, Y.; Zhou, X.; Cao, J.; Zhou, W.; Cao, J.; Li, R.; Zhou, D.; Tian, Y.; Zhu, X.; Li, C.; Wu, H.; Liu, X.; He, J.; Zhou, Y.; Liu, H.; Zhang, L.; Wang, Z.; Duan, H.; Zhou, Y.; Min, X.; Jia, Q.; Zhou, D.; Zhang, W.; Cao, J.; Yang, X.; Yu, J.; Zhang, S.; Duan, H.; and Zhai, G. 2025c. Large Multimodal Models Evaluation: A Survey. *SCIS*.

Zhou, Z.; Wang, Q.; Lin, B.; Su, Y.; Chen, R.; Tao, X.; Zheng, A.; Yuan, L.; Wan, P.; and Zhang, D. 2024. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*.