LongVQUBench: Benchmarking Long-Term Video Quality Understanding of Vision-Language Models

Arpita Nema, Hanwei Zhu, Xi Zhang, and Weisi Lin

Nanyang Technological University, Singapore
arpita004@e.ntu.edu.sg, {hanwei.zhu, xi.zhang, WSLin}@ntu.edu.sg
Project page: https://longvqubench.github.io.

Abstract. The evaluation of long-term video quality understanding remains an open challenge for large vision-language models (LVLMs). Existing video quality benchmarks predominantly focus on short clips and isolated distortions, overlooking the temporal continuity, cumulative degradation, and reasoning complexity inherent in long-duration content. To address these limitations, we present **LongVQUBench**, a comprehensive benchmark for long-term video quality understanding. LongVQUBench contains over 1,200 diverse videos spanning movies, documentaries, surveillance footage, egocentric recordings, and animated content, accompanied by 1,500 multiple-choice and open-ended questions for validation and testing. To assess perceptual reasoning across different temporal scopes, we introduce three progressively complex evaluation levels: (i) local event quality understanding (LQU) for analyzing localized distortions; (ii) cross-event quality reasoning (CQR) for integrating multiple degraded events; and (iii) global quality understanding (GQU) for holistic perceptual evaluation over extended durations. Furthermore, a needle distortion question-answering (NDQA) paradigm is embedded across all three levels, where spatial or temporal artifacts are sparsely inserted to probe fine-grained detection and reasoning capabilities. Extensive experiments on 14 state-of-the-art LVLMs reveal significant performance degradation with increasing video length and reasoning depth, highlighting their limited capacity for long-range temporal integration and perceptual attribution. We envision LongVQUBench as a foundational step toward the systematic, hierarchical, and explainable evaluation of LVLMs' long-term video quality understanding.

Keywords: Long-term video quality understanding · Video quality benchmark · Large vision-language models

1 Introduction

The rapid progress of large vision-language models (LVLMs) has revolutionized multimodal understanding by integrating visual perception with linguistic reasoning. Recent models such as GPT-5 [43], LLaVA-OneVision-1.5 [2], and

Qwen2.5-VL [5] exhibit strong capabilities in fine-grained recognition, context-aware image description [74], and short- and long-term video question answering [14,54,71], signaling a transition from perception-driven analysis to semantic comprehension. Nevertheless, long-term video quality understanding (LVQU) remains underexplored. Unlike conventional video understanding tasks centered on events and semantics, LVQU requires reasoning over perceptual fidelity, temporal coherence, and cumulative degradation across extended time spans. This is an essential ability for assessing the stability and human-perceived quality of real-world long-form videos, yet still beyond the reach of current LVLMs.

Video quality assessment (VQA) provides a complementary foundation for perceptual modeling by quantifying human judgments of distortions such as compression artifacts [32, 41, 51, 57, 64, 66], transmission errors [22, 39, 42, 61], and temporal flicker [9, 10, 72]. Classical datasets, including LIVE-VQA [42], KoNViD-1k [19], and YouTube-UGC [49], consist mainly of short clips with controlled distortions, which are valuable for low-level modeling but insufficient for capturing the temporal continuity and contextual dynamics of long videos. In contrast, LVLM-oriented benchmarks have emphasized semantic video understanding, such as action recognition and multimodal reasoning [14,15]. Although LongVideoBench [54] extends evaluation to long-form semantics, it remains focused on event comprehension rather than perceptual quality, while Q-Bench-Video [71] explores video-quality reasoning but is restricted to short clips without hierarchical temporal assessment. Therefore, a unified benchmark for long-term video quality understanding, integrating perceptual fidelity, temporal coherence, and multimodal reasoning, is still lacking.

To bridge these gaps, we introduce **LongVQUBench**, a comprehensive benchmark for evaluating the LVQU of LVLMs. The benchmark contains over $1,200$ videos drawn from diverse sources, including films, documentaries, surveillance footage, egocentric recordings, and computer-generated content, spanning durations from a few minutes to nearly two hours. This diversity encompasses a wide range of perceptual distortions, such as lighting drift, scene transitions, codec artifacts, and generative distortions, that emerge over extended viewing periods. To systematically evaluate LVQU across increasing temporal scopes, LongVQUBench defines three hierarchical levels: **local event quality understanding (LQU)**, **cross-event quality reasoning (CQR)**, and **global quality understanding (GQU)**. These levels progressively test a model's capacity to identify localized distortions, integrate perceptual cues across events, and evaluate holistic perceptual integrity and temporal stability.

Beyond this hierarchical structure, LongVQUBench introduces a **needle distortion question-answering (NDQA)** paradigm, in which spatial or temporal artifacts of varying intensities are sparsely embedded throughout long videos. NDQA enables the analysis of fine-grained perceptual sensitivity and challenges LVLMs to reason beyond coarse semantic cues. We evaluate 14 state-of-the-art LVLMs under zero-shot settings. The results show a consistent decline in performance as reasoning depth increases, revealing limitations in temporal localization, distortion attribution, and global quality reasoning. LongVQUBench

establishes the first systematic benchmark for long-term video quality understanding, providing a principled framework to advance perceptual modeling, temporal reasoning, and multimodal integration toward human-level long-form video comprehension.

Before delving into detail, we highlight our main contributions as follows:

- ❑ A comprehensive benchmark, **LongVQUBench**, specifically designed to evaluate the long-term video quality understanding capability of LVLMs across diverse real-world content.
- ❑ A hierarchical evaluation framework, encompassing local, cross-event, and global quality understanding, complemented by the needle distortion question-answering (NDQA) paradigm to probe fine-grained perceptual sensitivity.
- ❑ A large-scale empirical study involving 14 state-of-the-art LVLMs, which exposes fundamental limitations in temporal localization, perceptual attribution, and global reasoning across extended durations.

2 Related Work

2.1 Large Vision-Language Models

Large vision–language models (LVLMs) have significantly advanced multimodal understanding by aligning visual and textual modalities within unified generative frameworks. Foundational models such as CLIP [40] and ALIGN [20] established scalable pretraining paradigms based on contrastive vision–language alignment, while BLIP [27] and BLIP-2 [26] introduced modular strategies that integrate pretrained language models with frozen vision encoders for efficient cross-modal learning. Instruction-tuned LVLMs, including Flamingo [1], GPT-5 [43], LLaVA [34], InstructBLIP [13], and Qwen2.5-VL [5], have achieved strong generalization in vision-language reasoning, demonstrating remarkable progress across captioning, question answering, and grounding tasks.

Recent research has extended LVLMs toward multi-image and long-context video understanding. Models like Video-LLaVA [30], LongVA [54], Co-Instruct [56], and mPLUG-Owl3 [60] enhance temporal modeling by processing sequential frames or dynamic clips with interleaved text–video inputs. Similarly, LLaVA-Next [33] and VideoChat [28] improve long-sequence reasoning through efficient frame sampling, recurrent memory fusion, and temporal attention. Despite these advances, most LVLMs remain limited by short-context constraints and lack explicit mechanisms for modeling cumulative perceptual changes over extended durations. This limitation underscores the need for dedicated benchmarks such as **LongVQUBench**, which evaluate perceptual reasoning and temporal quality understanding in realistic long-form video settings.

2.2 Benchmarks for LVU

Benchmarking long-term video understanding (LVU) is crucial for evaluating a model's capacity to capture extended temporal dependencies, multi-event rea-

Table 1: Comparison of **LongVQUBench** with existing benchmarks. Columns report the number of videos (#**Vid.**), number of QA pairs (#**QA**), average video duration in seconds (**Len.**), support for Multiple-choice questions (**MCQ**) and open-ended questions, coverage of diverse genres (**Diverse Genres**), multi-duration evaluation (**Multi-Level**), and capability for video quality understanding (**VQ Underst.**). The upper block lists general video understanding benchmarks, while the lower block focuses on video quality assessment benchmarks.

Benchmarks	#Vid.	#QA	Len. (s)	MCQ	Open-ended	Diverse Genres	Multi-Level	VQ Underst.
Movie101 [63]	101	-	6144	✗	✓	✗	✗	✗
EgoSchema [36]	5,063	5,063	180	✓	✗	✗	✗	✗
MovieChat-1K [44]	1000	13K	500	✓	✓	✗	✗	✗
Video-MME [15]	900	2,700	1024	✓	✗	✓	✓	✗
LongVideoBench [54]	3,763	6,678	473	✓	✗	✓	✓	✗
MLVU [73]	1,730	3,102	930	✓	✓	✓	✓	✗
Q-Bench-Video [71]	1,800	2,378	10	✓	✓	✓	✗	✓
LongVQUBench	1,200	1,500	742.2	✓	✓	✓	✓	✓

soning, and narrative coherence. Early datasets such as ActivityNet [6], Kinetics [23], and Something-Something [17] primarily assess short clips and isolated actions. Subsequent works including Movie101 [63], EgoSchema [36], and MovieChat-1K [44] extend evaluation to narrative and egocentric contexts, while LongVideoBench [54], Video-MME [15], and MLVU [73] advance multi-task and long-context reasoning across diverse domains. Q-Bench-Video [71] further explores video quality understanding but remains limited to short sequences. In contrast, **LongVQUBench** introduces hierarchical, perceptually grounded evaluation across minute-to-hour videos, integrating temporal coherence, multi-level reasoning, and fine-grained perceptual quality assessment (see Table 1).

2.3 VQA Methods

Video quality assessment (VQA) aims to estimate human perceptual judgment of visual fidelity and temporal stability. Classical full-reference (FR) methods, such as PSNR and SSIM [50], model low-level signal fidelity, while perceptually motivated models like VMAF [29] and MOVIE [41], integrate spatial and temporal features. No-reference (NR) approaches, including NIQE [38], BRISQUE [37], and VIDEVAL [46], estimate quality without reference images through handcrafted or learned statistical priors. With deep learning, advanced FR and NR methods such as DeepVQA [24], C3DVQA [58], VSFA [25], Fast-VQA [53], and DOVER [55] leverage perceptual feature representations to achieve stronger correlation with human opinion scores. Recent transformer-based [52, 75, 76] and multimodal approaches [21, 47, 69] further enhance temporal modeling and generalization. While these methods effectively predict short-term quality, they lack the reasoning ability and temporal context understanding required for long-form video analysis. **LongVQUBench** bridges this gap by integrating perceptual

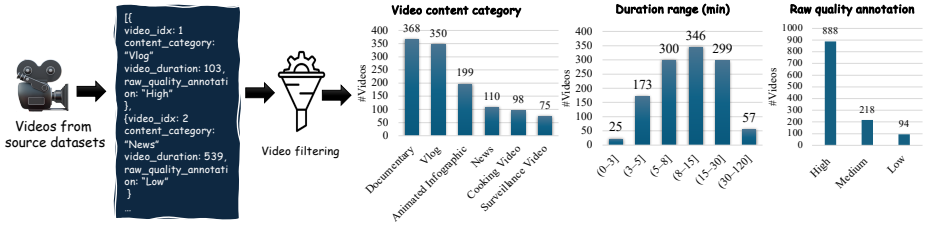


Fig. 1: Long-term duration videos from LongVideoBench [54], MLVU [73], and LongVideoReason [8] are first aggregated and tagged, followed by a filtering and removal process to achieve the target distribution of LongVQUBench.

modeling with language-based reasoning to evaluate long-term video quality understanding in LVLMs.

3 LongVQUBench

This section details the design philosophy, data composition, and evaluation structure of **LongVQUBench**, a large-scale benchmark for assessing the long-term video quality ability of LVLMs. The benchmark is developed to jointly evaluate perceptual fidelity and temporal reasoning across extended durations, establishing a unified testbed for comprehensive analysis.

3.1 Overview

LongVQUBench is designed to examine how LVLMs perceive, reason, and explain video quality over long temporal horizons. Unlike existing short-clip datasets [71], it focuses on the gradual evolution of quality, the accumulation of perceptual degradation, and the reasoning dependencies among temporally distant events. The benchmark is constructed with three design goals: 1) To cover a broad spectrum of long-form videos that reflect diverse real-world and synthetic scenarios; 2) To ensure representative coverage across both video duration and perceptual quality; and 3) To enable structured, hierarchical evaluation of temporal reasoning and fine-grained perceptual sensitivity. An overview of the dataset composition is presented in Figure 1, which summarizes category distributions, duration statistics, and quality-level balance.

3.2 Long-term Video Collection

LongVQUBench contains **1,200 videos** sourced from publicly available datasets and open media collections, including LongVideoBench [54], MLVU [73], and LongVideoReason [8]. Each video spans a duration between several minutes and two hours, enabling analysis of temporal consistency and cross-event degradation in realistic viewing conditions.

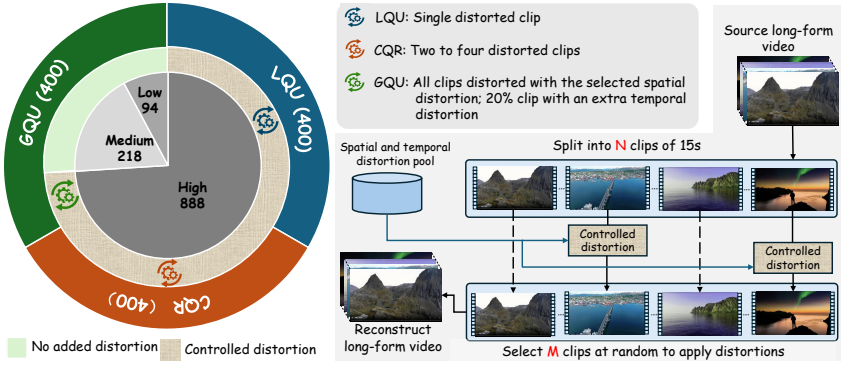


Fig. 2: Left: Distribution of videos across hierarchical evaluation levels along with the proportion of samples subjected to controlled distortions. Right: Illustration of the controlled distortion pipeline. High-quality videos are first segmented into 15-second clips. Controlled distortions are then applied according to predefined distortion pools and configurations (LQU, CQR, GQU). Finally, distorted clips are merged into full-length videos to enable distortion-aware question-answer generation.

Diverse Video Sources. The videos cover multiple domains, including feature films, documentaries, surveillance footage, egocentric recordings, instructional videos, and computer-generated scenes. This diversity captures a wide range of motion patterns, editing styles, and semantic structures, ensuring the benchmark's generality. Such variety also reflects the heterogeneous perceptual challenges faced by LVLMs when analyzing long-form content.

Comprehensive Video Length. As shown in Figure 1, LongVQUBench includes a balanced distribution of video durations. Approximately one quarter of the dataset consists of short videos under 10 minutes, while the remainder spans medium (10 - 30 minutes) and long (30 - 120 minutes) durations. This coverage allows comprehensive analysis of performance trends as the temporal context increases.

Comprehensive Quality Range. To ensure perceptual diversity, the dataset encompasses three quality levels: *high* (H), *moderate* (M), and *low* (L), as illustrated in Figure 1. These levels correspond to varying degrees of compression, lighting instability, motion jitter, and generative distortion. The majority of videos remain of high perceptual quality (H = 888), complemented by moderate (M = 218) and low (L = 94) quality samples. This distribution is intentional: controlled distortions are systematically applied only to high-quality source videos, as shown in Figure 2. Starting with clean visual content enables precise manipulation of distortion type and intensity, allowing us to generate distortion-aware question-answer pairs with reliable ground-truth alignment.

3.3 Benchmark Construction

LongVQUBench is constructed to systematically evaluate the perceptual reasoning capability of LVLMs across long-duration videos. Each video is paired with question-answer (QA) item(s) that assess how well models perceive, reason, and explain video quality under varying perceptual and temporal conditions. The benchmark integrates a hierarchical evaluation framework spanning local, cross-event, and global reasoning levels, complemented by a *Needle Distortion Question-Answering (NDQA)* paradigm for fine-grained sensitivity assessment.

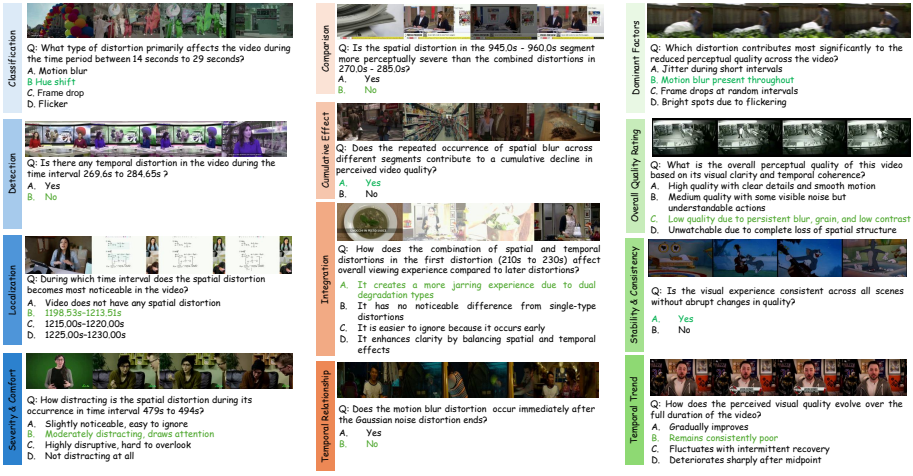
Distortion Configuration. To simulate realistic degradation patterns, our benchmark incorporates a diverse set of *spatial* and *temporal* distortions that serve as the foundation for NDQA construction, as illustrated in Figure 2. The dataset includes 14 spatial and 4 temporal distortion types chosen for their relevance to common video production, transmission, and generative scenarios. Spatial distortions affect frame-level fidelity, while temporal distortions impact motion continuity and global temporal stability. Each distortion is implemented at three controlled intensity levels to ensure perceptual diversity. Detailed distortion types and parameter configurations are provided in the supplementary material.

Needle Distortion Question-Answering (NDQA). The NDQA paradigm employs the aforementioned distortions to probe models' ability to detect and interpret degradations embedded within long videos. In this setting, spatial or temporal distortions of varying amplitudes are introduced without disrupting semantic content or narrative flow. Models are evaluated through two complementary question formats:

- **Multiple-choice questions (MCQ)** include four types: *Yes-or-No*, *What*, *Which*, and *How*. The *Yes-or-No* type examines the presence of perceptual degradations, such as detecting whether flicker or blur occurs within a segment. The *What* type identifies the specific distortion category, while the *How* type quantifies its perceptual strength or temporal extent. The *Which* type requires comparative reasoning, prompting the model to determine which segment or event exhibits more severe degradation. Together, these question types jointly evaluate recognition accuracy, comparative judgment, and sensitivity to perceptual intensity.
- **Open-ended questions**, which require free-form reasoning, prompting models to describe the degradation's nature, location, and perceptual impact.

This dual-format design unifies objective accuracy and interpretive evaluation, forming a balanced framework to measure both perceptual sensitivity and reasoning depth. Building on the NDQA paradigm, the subsequent three levels, *Local Event Quality Understanding (LQU)*, *Cross-Event Quality Reasoning (CQR)*, and *Global Quality Understanding (GQU)*, extend the assessment toward progressively broader temporal and perceptual contexts.

1) Local Event Quality Understanding (LQU): The LQU level evaluates a model's ability to detect, classify, localize, and interpret a single, tempo-



(a) LQU (b) CQR (c) GQU

Fig. 3: LongVQUBench features perceptual quality reasoning questions across multiple temporal scopes: (a) Local Event Quality Understanding (LQU) for analyzing localized distortions; (b) Cross-Event Quality Reasoning (CQR) for integrating multiple degraded events; and (c) Global Quality Understanding (GQU) for holistic perceptual evaluation over extended durations.

rally bounded distortion event within a long video [16, 31]. Each event typically spans 5 to 20 seconds and reflects transient quality degradation phenomena that have become increasingly important in video quality understanding and analysis [21, 71], such as localized blur [21, 35], compression noise [65, 67], luminance fluctuation [10, 12], or flicker [10, 11], which can affect short-term perceptual comfort and visual attention [12, 45]. LQU primarily tests the model's short-term perceptual sensitivity and its capacity to link local distortions with subjective viewing discomfort.

Question Design. Each LQU sample is associated with a question targeting one of five complementary perceptual dimensions:

- *Detection* determines whether a perceptual distortion exists within a given temporal segment.
- *Localization* identifies when the degradation occurs in the video timeline.
- *Classification* specifies the distortion category, such as blur, flicker, or color shift.
- *Severity and Comfort Assessment* estimates how intense, distracting, or perceptually disturbing the distortion appears.
- *Open Reasoning* explains why the observed artifact negatively affects perceived quality, focusing on aspects like motion inconsistency or visual discomfort.

We have shown the sampled questions based on the above design in Figure 3(a). This progressive questioning structure encourages models to move be-

yond binary judgment toward fine-grained perceptual reasoning. Each QA pair is manually validated to ensure visual clarity, temporal precision, and interpretive consistency, enabling objective evaluation of local perceptual sensitivity.

2) Cross-Event Quality Reasoning (CQR): The CQR level examines a model's ability to compare, associate, and integrate multiple distortion events distributed across a long video. Unlike LQU, which focuses on short-term perceptual sensitivity, CQR targets reasoning across extended temporal spans where multiple degradations may occur sequentially or intermittently. This level evaluates whether a model can assess the relative severity of distortions, capture their temporal relationships, and infer how their interactions influence overall perceptual comfort.

Question Design. Each CQR instance is designed to measure the model's capability to conduct multi-event reasoning across one of five complementary dimensions:

- *Comparison* identifies which segment or scene exhibits stronger or more disturbing degradations.
- *Cumulative Effect* assesses how the accumulation of multiple artifacts influences perceptual stability or viewer fatigue.
- *Integration* determines whether the model can synthesize perceptual evidence from multiple segments to form a consistent overall judgment.
- *Temporal Relation* evaluates whether the distortions are temporally correlated, clustered, or independently distributed.
- *Open Reasoning* requires the model to explain how different distortions interact over time, emphasizing contextual reasoning beyond local perception.

Together, these dimensions position CQR as a bridge between local perceptual analysis and holistic quality interpretation. Each annotated QA pair is manually verified to ensure the spatial-temporal correspondence of events and to prevent content bias between compared segments. Several sampled questions are shown in Figure 3(b). This structured formulation provides a controlled setting for testing the model's ability to reason across temporal dependencies and accumulated degradations.

3) Global Quality Understanding (GQU): The GQU level evaluates a model's ability to synthesize an overall perceptual judgment across an entire long video, typically spanning from one minute to two hours. It requires reasoning about temporal trends, cumulative degradations, and perceptual stability over prolonged viewing. In contrast to LQU and CQR, which focus on local distortions or multi-event relations, GQU emphasizes holistic temporal integration, tracking how perceptual quality evolves, stabilizes, or deteriorates over time and how this evolution affects the final perceptual judgment.

Question Design. Each GQU instance aims to measure long-term perceptual coherence through one of five complementary dimensions:

- *Stability Evaluation* assesses the consistency of viewing experience, capturing long-term fluctuations and viewer fatigue.
- *Dominant Factor Identification* determines the principal degradation type or event that most strongly influences the overall judgment.
- *Trend Assessment* identifies whether the perceived quality improves, remains stable, or degrades as the video progresses.
- *Overall Evaluation* estimates the global perceptual quality of entire video, integrating spatial fidelity, temporal smoothness, and aesthetic appeal.
- *Open Reasoning* requires the model to explain why the overall perception aligns with its given judgment, articulating the temporal and perceptual evidence that supports its decision.

Representative questions generated according to the above design are shown in Figure 3(c). This level bridges perceptual aggregation and interpretive reasoning, allowing the evaluation of whether LVLMs can move from event-based assessment to globally consistent quality judgments. All QA items are validated through expert review to ensure reliable temporal coverage and consistent interpretability, thereby establishing a foundation for analyzing holistic perceptual reasoning in long-duration videos.

3.4 Questions & Answers Annotation

All QA pairs in **LongVQUBench** are annotated through a controlled two-stage process to ensure temporal accuracy, semantic clarity, and perceptual consistency. In the first stage, QA items are constructed based on the reasoning dimensions of each level: LQU, CQR, and GQU, by identifying distortion events, marking their temporal boundaries, and formulating questions that probe detection, comparison, and reasoning. In the second stage, each QA item is independently reviewed by multiple experts, with disagreements resolved through consensus to ensure annotation reliability. Both multiple-choice and open-ended formats are adopted: the former targets objective recognition and localization, while the latter evaluates explanatory reasoning, assisted by GPT-based scoring for relevance and completeness [71,73]. This rigorous yet scalable annotation protocol guarantees consistency across perceptual levels and establishes a robust foundation for evaluating long-term video quality understanding. Further details of the annotation procedure are provided in the supplementary material.

4 Results on LongVQUBench

In this section, we first describe the experimental settings, including the participating LVLMs, the evaluation protocol, and the dataset split. We then present quantitative results and analyze the performance of current LVLMs on long-term video quality understanding. More experimental results can be found in the supplementary material.

4.1 Experimental Settings

Benchmark LVLMs. We evaluate a total of **14 LVLMs**, including *3 proprietary models*: GPT-5 [43], Gemini 3 [18], and Qwen-VL-Max [3]; *7 open-source models*: LLaVA-NeXT-Video [33], ShareGPT4Video [7], Qwen3-VL [4], MovieChat [44], LLaVA-Video [70], VQA^2 [21], and Long-RL [8]; and *4 agentic LVLMs*: VideoAgent [48], VideoExplorer [62], LongVT [59], and DeepVideoDiscovery [68]. Notably, VQA^2 is specifically designed for video quality understanding. Together, these models cover a diverse set of architectures, training paradigms, and reasoning mechanisms, enabling a comprehensive evaluation of current LVLM capabilities for long-term video quality understanding. More details of these LVLMs can be found in the supplementary material.

Evaluation Protocol. We evaluate LVLMs using a frame-sampling-based inference pipeline designed for long-duration videos. Given a video $\mathcal{V}$ with duration T, we uniformly sample n frames at a fixed frame rate (FPS). The sampled frame sequence $\{f_1, f_2, \ldots, f_n\}$ is ordered chronologically and provided to the LVLM within a single prompt. The prompt explicitly specifies: (i) the total video duration, (ii) the sampling frame rate (FPS), (iii) the total number of sampled frames, and (iv) a strict output format constraint. The model is instructed to act as an expert in video quality analysis and must select *exactly one* answer from the provided candidate options. Each question is evaluated in a single forward pass without iterative interaction.

Dataset and Split. Experiments are conducted on the **LongVQUBench** dataset, which evaluates long-duration video quality understanding across three hierarchical levels: LQU, CQR, and GQU, each comprising four question-design dimensions. To ensure balanced evaluation across all dimensions, we adopt a **stratified 40% validation / 60% test split**. The split is performed independently within each question dimension to preserve the proportional distribution of samples. The validation set is used to determine the optimal number of sampled frames ($\#frames$). Once the sampling configuration is selected, it is fixed and applied to the held-out test set for final evaluation.

4.2 Main Results

Validation Results. We first analyze model performance on the 40% validation subset under different frame sampling budgets. Table 2 reports results for open-source and proprietary LVLMs when the number of uniformly sampled frames is capped at 1 FPS. Several observations can be drawn.

1) Increasing frame count yields limited gains. For most models, increasing the number of sampled frames does not consistently improve performance. Large proprietary models such as GPT-5 and Gemini-3 benefit from moderate increases in frame count, with GPT-5 achieving its best performance at 256 frames (74.1 overall) and Gemini-3 peaking at 128 frames (68.9 overall). However, the improvements quickly saturate beyond moderate sampling budgets. In contrast, several video-specialized open-source models achieve their best performance with relatively small inputs, e.g., VQA^2 performs best with only 8 frames

Table 2: Results on the LongVQUBench `val` subset for the long video quality perception ability of LVLMS according to the number of max frames (capped at 1 FPS).

Model	#frames	LQU	CQR	GQU	Total	Model	#frames	LQU	CQR	GQU	Total
GPT-5	8	72.4	77.8	61.2	70.5	LLaVA-Video	**8**	**61.5**	**64.5**	**48.5**	**58.2**
	32	71.5	78.5	63.0	71.0		16	60.2	64.0	47.5	57.2
	64	74.0	80.2	64.5	72.9		30	57.5	62.6	47.0	55.7
	128	75.5	81.6	65.0	74.0		128	N/A	N/A	N/A	N/A
	256	**75.2**	**81.2**	**65.8**	**74.1**		256	N/A	N/A	N/A	N/A
Gemini-3	8	71.2	73.6	59.6	68.1	VQA^2	**8**	**63.2**	**65.5**	**49.5**	**59.4**
	32	67.2	75.0	56.2	66.1		32	59.0	63.5	47.5	56.7
	64	67.0	74.2	58.5	66.6		64	59.5	64.0	52.0	58.5
	128	**71.0**	**76.5**	**59.2**	**68.9**		128	58.0	63.0	50.5	57.2
	256	68.8	75.6	57.6	67.3		256	55.5	61.0	48.0	54.8
Qwen-VL-Max	8	68.2	70.4	48.2	62.3	Long-RL	8	62.5	66.7	50.0	59.7
	32	64.5	70.8	55.0	63.4		32	61.5	67.2	50.0	59.6
	64	**64.5**	**73.5**	**54.5**	**64.2**		**64**	**61.7**	**67.0**	**55.0**	**61.2**
	128	62.0	67.5	53.0	60.8		128	60.5	66.8	52.0	59.8
	250	61.0	65.8	51.2	59.3		256	57.2	66.0	49.2	57.5
LLaVA-NeXT-Video	**8**	**46.6**	**75.2**	**50.2**	**57.3**	ShareGPT4Video	8	30.5	37.2	20.0	29.2
	16	45.8	68.2	42.0	52.0		**16**	**33.5**	**35.0**	**20.5**	**29.7**
	30	45.8	67.8	38.7	50.8		64	32.2	34.5	20.0	28.9
	64	N/A	N/A	N/A	N/A		128	N/A	N/A	N/A	N/A
	128	N/A	N/A	N/A	N/A		256	N/A	N/A	N/A	N/A
Qwen3-VL	8	51.8	59.6	49.5	53.6	MovieChat	8	36.5	38.7	21.2	32.1
	32	54.5	60.0	50.0	54.8		32	37.5	42.5	26.0	35.3
	64	**54.0**	**75.0**	**60.4**	**63.1**		**64**	**35.0**	**43.6**	**30.0**	**36.2**
	128	52.0	58.2	51.0	53.7		128	34.2	42.0	31.5	35.9
	256	49.5	56.5	49.5	51.8		256	33.0	41.5	30.0	34.8

(59.4 overall), and Qwen3-VL peaks at 64 frames (63.1 overall). These results indicate diminishing returns from increasing temporal coverage.

2) Global quality reasoning remains challenging. Across nearly all models, performance follows a consistent pattern where LQU and CQR scores exceed GQU scores. For example, GPT-5 achieves 81.2 on CQR but only 65.8 on GQU, while Gemini-3 obtains 76.5 on CQR versus 59.2 on GQU. This gap suggests that current LVLMs are more capable of detecting localized or cross-event distortions than synthesizing holistic quality judgments over long temporal horizons.

Table 3 presents validation results for agentic LVLMs that employ adaptive keyframe selection strategies. Compared with uniform frame sampling, the effectiveness of adaptive selection varies substantially across methods. DeepVideoDiscovery significantly outperforms other agentic approaches, achieving 71.7 overall accuracy with particularly strong performance on LQU (69.2) and CQR (82.7), approaching the level of leading proprietary LVLMs. In contrast, simpler agent-based systems such as VideoAgent and VideoExplorer achieve substantially lower scores. These results indicate that while adaptive frame selection can be beneficial, its effectiveness strongly depends on the quality of the exploration and reasoning strategy used to identify informative frames.

Test Results. Using the frame configurations selected on the validation set, we report the final accuracy on the 60% held-out test set in Table 4 for multiple-

Table 3: Results on the LongVQUBench `val` subset for the long video quality perception ability of Agentic LVLMs. Keyframes are adaptively selected - details provided in the supplementary material.

Agent	LQU	CQR	GQU	Overall
VideoAgent	34.2	42.5	30.7	35.8
VideoExplorer	44.5	58.5	54.0	52.3
LongVT	43.7	65.0	60.0	56.2
DeepVideoDiscovery	**69.2**	**82.7**	**63.2**	**71.7**

Table 4: `Test` leaderboard of LongVQUBench across 14 LVLMs, organized by hierarchical quality understanding levels and MCQ question dimensions. Abbreviations denote the following terms: #F: Number of Frame; D: Detection; L: Localization; C: Classification; SA: Severity & Comfort Assessment; CMP: Comparison; CE: Cumulative Effect; I: Integration; TR: Temporal Relation; SE: Stability Evaluation; DFI: Dominant Factor Identification; TA: Trend Assessment; OE: Overall Evaluation.

Model	#F	LQU				CQR				GQU				Overall
		D	L	C	SA	CMP	CE	I	TR	SE	DFI	TA	OE	
Closed-source LVLMs														
GPT-5	256	**84.6**	71.5	**56.5**	49.0	**71.5**	**87.6**	86.5	83.0	**68.5**	**65.0**	**48.5**	**61.5**	**69.5**
Gemini-3	128	76.5	68.0	54.0	48.0	70.0	85.0	82.0	80.0	67.0	63.0	47.5	60.0	66.8
Qwen-VL-Max	64	70.0	65.5	51.0	44.0	68.0	84.0	80.0	77.0	65.0	60.0	45.5	58.0	64.0
Video LVLMs														
LLaVA-NeXT-Video	8	35.0	58.7	46.7	38.3	61.7	70.3	74.7	73.3	58.3	50.3	58.3	51.7	56.4
ShareGPT4Video	16	26.5	34.6	28.0	32.6	24.6	35.2	34.0	33.0	23.6	28.0	23.5	26.0	29.1
Qwen3-VL	64	46.5	62.5	47.0	46.2	65.0	78.2	75.0	72.1	64.0	60.0	**48.5**	58.5	60.3
MovieChat	64	25.2	41.8	30.0	23.0	36.5	43.6	41.2	39.3	44.0	40.0	25.0	38.0	35.6
LLaVA-Video	8	30.5	57.0	43.0	36.0	66.0	76.0	72.0	68.0	59.0	54.0	39.0	53.0	54.5
VQA^2	8	62.5	53.6	48.0	41.5	60.5	49.0	75.0	72.0	48.2	59.0	44.0	37.5	54.2
Long-RL	64	73.3	74.3	48.3	51.7	38.3	53.3	**93.3**	**83.3**	53.3	63.3	48.3	38.3	59.9
Agentic LVLMs														
VideoAgent	–	25.0	30.5	30.5	28.5	29.5	35.0	47.5	48.0	33.5	40.6	33.0	40.6	35.2
VideoExplorer	–	33.5	49.0	35.0	40.2	52.0	57.0	63.0	60.2	49.5	44.0	42.5	43.5	47.5
LongVT	–	37.5	62.0	55.0	46.5	43.0	56.0	53.0	50.0	41.0	37.5	27.0	37.0	45.5
DeepVideoDiscovery	–	77.3	**83.3**	56.3	**59.7**	66.3	67.3	81.3	73.3	63.3	61.3	46.3	56.3	66.0

choice questions. The Figure 4 shows open-ended question-answer pair in dataset and answers from each of best-performing closed-source LVLM, open-source LVLM, and agentic LVLM. Table 5 reports the relevance and completeness scores of open-ended question answers across the hierarchical levels LQU, CQR, and GQU. The details of the relevance and completeness score of the GPT-based prompt are available in supplementary material. All models are evaluated under a single-pass inference protocol without additional adaptation. The leaderboard provides fine-grained results across hierarchical quality understanding levels, including LQU, CQR, and GQU. Several key observations emerge.

1) Proprietary LVLMs achieve the strongest overall performance. Closed-source models dominate the leaderboard, with GPT-5 achieving the best overall accuracy (69.5), followed by Gemini-3 (66.8) and Qwen-VL-Max (64.0). These models show consistently strong performance across most reasoning dimensions, particularly in cross-event reasoning tasks.

Table 5: `Test` leaderboard of relevance (R) and completeness (C) scores (%) for open-ended questions across hierarchical levels - LQU, CQR, and GQU.

Model	$\mathbf{LQU}_R$	$\mathbf{CQR}_R$	$\mathbf{GQU}_R$	$\mathbf{Overall}_R$	$\mathbf{LQU}_C$	$\mathbf{CQR}_C$	$\mathbf{GQU}_C$	$\mathbf{Overall}_C$
Closed-Source LVLMs								
GPT-5	**88.2**	**89.4**	**89.1**	**88.9**	**45.2**	**47.8**	**44.3**	**45.8**
Gemini-3	86.3	86.1	85.6	86.0	38.3	42.3	38.5	39.7
Qwen-VL-Max	83.4	82.7	84.9	83.7	36.9	40.3	37.2	38.1
Open-Source LVLMs								
LLaVA-NeXT-Video	74.1	78.3	76.6	76.3	36.2	39.5	37.5	37.7
ShareGPT4Video	71.5	76.3	75.2	74.3	35.0	38.2	36.9	36.7
Qwen3-VL	77.2	81.5	79.3	79.3	37.9	40.8	38.8	39.2
MovieChat	72.4	77.5	74.6	74.8	35.4	39.1	36.3	36.9
LLaVA-Video	73.5	78.7	75.3	75.8	36.0	39.4	36.8	37.4
VQA^2	73.0	78.1	74.9	75.3	35.6	39.4	36.6	37.2
Long-RL	76.5	81.0	77.5	78.3	37.4	40.8	37.8	38.7
Agentic LVLMs								
VideoAgent	72.0	74.8	74.2	73.7	35.1	38.6	36.4	36.70
VideoExplorer	76.3	78.2	77.5	77.3	37.4	40.7	37.9	38.67
LongVT	75.2	77.4	76.4	76.3	36.8	40.6	37.3	38.23
DeepVideoDiscovery	80.3	82.6	81.1	81.3	39.4	43.0	39.6	40.67

2) Local event understanding is relatively tractable. LQU tasks achieve the high accuracy compared to GQU across the multiple-choice questions for open-sourced and closed-sourced LVLMs. While proprietary models perform strongly overall, some open-source and agentic systems also demonstrate competitive performance in specific dimensions, such as DeepVideoDiscovery achieving the highest localization accuracy (83.3).

3) Cross-event reasoning highlights the importance of temporal integration. CQR results show substantial variation across models. GPT-5 achieves strong performance in cumulative effect and comparison, while Long-RL achieves the best integration score (93.3), indicating that effective temporal aggregation is crucial for modeling interactions among multiple distortion events.

4) Global quality understanding remains the most challenging level. A consistent performance drop from LQU and CQR to GQU is observed across nearly all models, highlighting the difficulty of synthesizing long-term perceptual evidence into a coherent global quality judgment.

5) Agentic LVLMs show promising but uneven performance. Among agentic approaches, DeepVideoDiscovery achieves competitive performance (66.0 overall), approaching proprietary models and outperforming most LVLMs. Simpler agentic systems lag, showing effective frame exploration and reasoning are critical for agent-based long-video quality understanding.

6) Completeness remains a key challenge despite high relevance. Open-ended question relevance scores remain consistently high across all models, with closed-source models achieving the strongest performance (GPT-5: 88.9%), indicating that most responses effectively address the given questions. However, completeness scores are substantially lower across all tiers, with even the best-performing

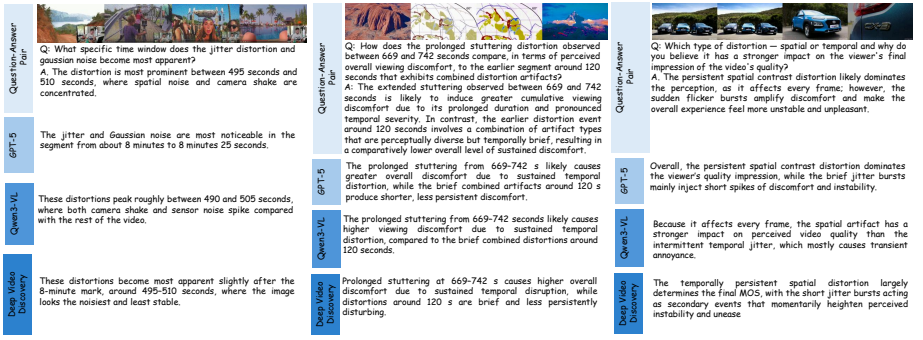


Fig. 4: Open-ended evaluation: Example questions from LongVQUBench across three hierarchical categories (LQU, CQR and GQU), with corresponding answers from each of best-performing closed-source LVLM (GPT-5 [43]), open-source LVLM (Qwen3-VL [4]), and agentic LVLM (DeepVideoDiscovery [68]).

model, GPT-5, reaching only 45.8%. This reflects the inherent difficulty of producing comprehensive answers to open-ended quality assessment questions.

5 Conclusion

We introduce **LongVQUBench**, the first large-scale benchmark designed to evaluate long-term video quality understanding in LVLMs. Unlike existing short-clip quality assessment datasets or long-video semantic benchmarks, our benchmark integrates perceptual fidelity, temporal coherence, and reasoning across extended durations through a hierarchical evaluation framework together with the NDQA paradigm for fine-grained perceptual probing. Extensive experiments on 14 state-of-the-art LVLMs reveal that model performance consistently degrades as temporal span and reasoning complexity increase, and that simply increasing the number of sampled frames does not reliably improve accuracy. While proprietary models achieve the strongest overall performance, even leading systems struggle with global quality reasoning such as stability assessment and dominant factor attribution. These findings highlight that long-term perceptual reasoning remains an open challenge for current LVLMs, motivating future research toward more robust perceptual understanding over long-form videos.

Acknowledgements

This research is partially supported by the Ministry of Education, Singapore, under the funding of MOE-T2EP20123-0006. This work is also supported by gift funding from Amazon Prime Video for research on long-term video quality analysis. The authors would like to thank Alex Mackin and Benoit Vallade of Amazon Prime Video for their technical guidance and feedback on this research.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 3
2. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: LLaVA-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025) 1
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) 11
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025) 11, 15
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025) 2, 3
6. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: ActivityNet: A large-scale video benchmark for human activity understanding. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015) 4
7. Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Lin, B., Tang, Z., et al.: ShareGPT4Video: Improving video understanding and generation with better captions. Advances in Neural Information Processing Systems **37**, 19472–19495 (2024) 11
8. Chen, Y., Huang, W., Shi, B., Hu, Q., Ye, H., Zhu, L., Liu, Z., Molchanov, P., Kautz, J., Qi, X., et al.: Scaling RL to long videos. Advances in Neural Information Processing Systems **38**, 172842–172870 (2026) 5, 11
9. Choi, L.K., Bovik, A.C.: Flicker sensitive motion tuned video quality assessment. In: IEEE Southwest Symposium on Image Analysis and Interpretation. pp. 29–32. IEEE (2016) 2
10. Choi, L.K., Bovik, A.C.: Video quality assessment accounting for temporal visual masking of local flicker. Signal Processing: Image Communication **67**, 182–198 (2018) 2, 8
11. Choi, L.K., Cormack, L.K., Bovik, A.C.: On the visibility of flicker distortions in naturalistic videos. In: International Workshop on Quality of Multimedia Experience. pp. 164–169. IEEE (2013) 8
12. Choi, L.K., Cormack, L.K., Bovik, A.C.: Motion silencing of flicker distortions on naturalistic videos. Signal Processing: Image Communication **39**, 328–341 (2015) 8
13. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems **36**, 49250–49267 (2023) 3
14. Fang, X., Mao, K., Duan, H., Zhao, X., Li, Y., Lin, D., Chen, K.: MMBench-Video: A long-form multi-shot benchmark for holistic video understanding. Advances in Neural Information Processing Systems **37**, 89098–89124 (2024) 2
15. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 24108–24118 (2025) 2, 4

16. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: IEEE International Conference on Computer Vision. pp. 5267–5275 (2017) 8
17. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The "something something" video database for learning and evaluating visual common sense. In: IEEE International Conference on Computer Vision. pp. 5842–5850 (2017) 4
18. Hassabis, D., Kavukcuoglu, K.: A new era of intelligence with Gemini 3. Google Blog (2025), https://blog.google/products-and-platforms/products/gemini/gemini-3/, accessed: Feb. 25, 2026 11
19. Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., Li, S., Saupe, D.: The konstanz natural video database (KoNViD-1k). In: International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–6. IEEE (2017) 2
20. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916 (2021) 3
21. Jia, Z., Zhang, Z., Qian, J., Wu, H., Sun, W., Li, C., Liu, X., Lin, W., Zhai, G., Min, X.: VQA2: Visual question answering for video quality assessment. In: ACM International Conference on Multimedia. pp. 6751–6760 (2025) 4, 8, 11
22. Kanumuri, S., Cosman, P.C., Reibman, A.R., Vaishampayan, V.A.: Modeling packet-loss visibility in MPEG-2 video. IEEE Transactions on Multimedia **8**(2), 341–355 (2006) 2
23. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The Kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 4
24. Kim, W., Kim, J., Ahn, S., Kim, J., Lee, S.: Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In: European Conference on Computer Vision. pp. 219–234 (2018) 4
25. Li, D., Jiang, T., Jiang, M.: Quality assessment of in-the-wild videos. In: ACM International Conference on Multimedia. pp. 2351–2359 (2019) 4
26. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning. pp. 19730–19742 (2023) 3
27. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900 (2022) 3
28. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. Science China Information Sciences **68**(10), 200102 (2025) 3
29. Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A.K., Manohara, M.: Toward a practical perceptual video quality metric. Netflix Technology Blog (Jun 2016), https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652, accessed: Feb. 25, 2026 4
30. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-LLaVA: Learning united visual representation by alignment before projection. In: Conference on Empirical Methods in Natural Language Processing. pp. 5971–5984 (2024) 3
31. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: UniVTG: Towards unified video-language temporal grounding. In: IEEE International Conference on Computer Vision. pp. 2794–2804 (2023) 8

32. Lin, L., Yu, S., Zhou, L., Chen, W., Zhao, T., Wang, Z.: PEA265: Perceptual assessment of video compression artifacts. IEEE Transactions on Circuits and Systems for Video Technology **30**(11), 3898–3910 (2020) 2
33. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/, accessed: Feb. 25, 2026 3, 11
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems **36**, 34892–34916 (2023) 3
35. Liu, Y., Gu, K., Zhai, G., Liu, X., Zhao, D., Gao, W.: Quality assessment for real out-of-focus blurred images. Journal of Visual Communication and Image Representation **46**, 70–80 (2017) 8
36. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023) 4
37. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012) 4
38. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2012) 4
39. Pinson, M.H., Wolf, S.: A new standardized method for objectively measuring video quality. IEEE Transactions on Broadcasting **50**(3), 312–322 (2004) 2
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021) 3
41. Seshadrinathan, K., Bovik, A.C.: Motion tuned spatio-temporal quality assessment of natural videos. IEEE Transactions on Image Processing **19**(2), 335–350 (2009) 2, 4
42. Seshadrinathan, K., Soundararajan, R., Bovik, A.C., Cormack, L.K.: Study of subjective and objective quality assessment of video. IEEE Transactions on Image Processing **19**(6), 1427–1441 (2010) 2
43. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: OpenAI GPT-5 system card. arXiv preprint arXiv:2601.03267 (2025) 1, 3, 11, 15
44. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: MovieChat: From dense token to sparse memory for long video understanding. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 18221–18232 (2024) 4, 11
45. Terzić, K., Hansard, M.: Methods for reducing visual discomfort in stereoscopic 3D: A review. Signal Processing: Image Communication **47**, 402–416 (2016) 8
46. Tu, Z., Wang, Y., Birkbeck, N., Adsumilli, B., Bovik, A.C.: UGC-VQA: Benchmarking blind video quality assessment for user generated content. IEEE Transactions on Image Processing **30**, 4449–4464 (2021) 4
47. Wang, J., Duan, H., Jia, Z., Zhao, Y., Yang, W.Y., Zhang, Z., Chen, Z., Wang, J., Xing, Y., Zhai, G., et al.: LOVE: Benchmarking and evaluating text-to-video generation and video-to-text interpretation. arXiv preprint arXiv:2505.12098 (2025) 4
48. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: VideoAgent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024) 11

49. Wang, Y., Inguva, S., Adsumilli, B.: YouTube UGC dataset for video compression research. In: IEEE International Workshop on Multimedia Signal Processing. pp. 1–5. IEEE (2019) 2
50. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004) 4
51. Wang, Z., Lu, L., Bovik, A.C.: Video quality assessment based on structural distortion measurement. Signal processing: Image communication **19**(2), 121–132 (2004) 2
52. Wen, W., Li, M., Zhang, Y., Liao, Y., Li, J., Zhang, L., Ma, K.: Modular blind video quality assessment. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2763–2772 (2024) 4
53. Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., Lin, W.: FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling. In: European Conference on Computer Vision. pp. 538–554. Springer (2022) 4
54. Wu, H., Li, D., Chen, B., Li, J.: LongVideoBench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024) 2, 3, 4, 5
55. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: International conference on computer vision. pp. 20144–20154 (2023) 4
56. Wu, H., Zhu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Li, C., Wang, A., Sun, W., Yan, Q., et al.: Towards open-ended visual quality comparison. In: European Conference on Computer Vision. pp. 360–377. Springer (2024) 3
57. Xia, J., Shi, Y., Teunissen, K., Heynderickx, I.: Perceivable artifacts in compressed video and their relation to video quality. Signal Processing: Image Communication **24**(7), 548–556 (2009) 2
58. Xu, M., Chen, J., Wang, H., Liu, S., Li, G., Bai, Z.: C3DVQA: Full-reference video quality assessment with 3D convolutional neural network. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4447–4451. IEEE (2020) 4
59. Yang, Z., Wang, S., Zhang, K., Wu, K., Leng, S., Zhang, Y., Li, B., Qin, C., Lu, S., Li, X., et al.: LongVT: Incentivizing "thinking with long videos" via native tool calling. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 33816–33826 (2026) 11
60. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mPLUG-Owl3: Towards long image-sequence understanding in multi-modal large language models. In: International Conference on Learning Representations. vol. 2025, pp. 98891–98913 (2025) 3
61. Yim, C., Bovik, A.C.: Evaluation of temporal variation of video quality in packet loss networks. Signal Processing: Image Communication **26**(1), 24–38 (2011) 2
62. Yuan, H., Liu, Z., Zhou, J., Qian, H., Shu, Y., Sebe, N., Wen, J.R., Dou, Z.: VideoExplorer: Think with videos for agentic long-video understanding. arXiv preprint arXiv:2506.10821 (2025) 11
63. Yue, Z., Zhang, Q., Hu, A., Zhang, L., Wang, Z., Jin, Q.: Movie101: A new movie understanding benchmark. In: Association for Computational Linguistics. pp. 4669–4684 (2023) 4
64. Zhang, X., Wu, X.: Attention-guided image compression by deep reconstruction of compressive sensed saliency skeleton. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 13354–13364 (2021) 2

65. Zhang, X., Wu, X.: Multi-modality deep restoration of extremely compressed face videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2024–2037 (2022) 8
66. Zhang, X., Wu, X.: Lvqac: Lattice vector quantization coupled with spatially adaptive companding for efficient learned image compression. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 10239–10248 (2023) 2
67. Zhang, X., Zhu, H., Zhong, Y., Wang, J., Lin, W.: BADiff: Bandwidth adaptive diffusion model. Advances in Neural Information Processing Systems **38**, 36962–36987 (2026) 8
68. Zhang, X., Jia, Z., Guo, Z., Li, J., Li, B., Li, H., Lu, Y.: Deep Video Discovery: Agentic search with tool use for long-form video understanding. Advances in Neural Information Processing Systems **38**, 89863–89895 (2026) 11, 15
69. Zhang, X., Li, W., Zhao, S., Li, J., Zhang, L., Zhang, J.: VQ-Insight: Teaching VLMs for AI-generated video quality understanding via progressive visual reinforcement learning. In: AAAI Conference on Artificial Intelligence. vol. 40, pp. 12870–12878 (2026) 4
70. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: LLaVA-Video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024) 11
71. Zhang, Z., Jia, Z., Wu, H., Li, C., Chen, Z., Zhou, Y., Sun, W., Liu, X., Min, X., Lin, W., et al.: Q-Bench-Video: Benchmark the video quality understanding of LMMs. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3229–3239 (2025) 2, 4, 5, 8, 10
72. Zheng, Q., Fan, Y., Huang, L., Zhu, T., Liu, J., Hao, Z., Xing, S., Chen, C.J., Min, X., Bovik, A.C., et al.: Video quality assessment: A comprehensive survey. arXiv preprint arXiv:2412.04508 (2024) 2
73. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: MLVU: Benchmarking multi-task long video understanding. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 13691–13701 (2025) 4, 5, 10
74. Zhu, D., Shen, X., Li, X., Elhoseiny, M., et al.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: International Conference on Learning Representations. vol. 2024, pp. 18378–18394 (2024) 2
75. Zhu, H., Chen, B., Zhu, L., Chen, P., Song, L., Wang, S.: Video quality assessment for spatio-temporal resolution adaptive coding. IEEE Transactions on Circuits and Systems for Video Technology **34**(7), 6403–6415 (2024) 4
76. Zhu, H., Chen, B., Zhu, L., Wang, S.: Learning spatiotemporal interactions for user-generated video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology **33**(3), 1031–1042 (2023) 4