MLVU: Benchmarking Multi-task Long Video Understanding

Junjie Zhou[*1,2], Yan Shu[*2], Bo Zhao[*3,2], Boya Wu[2], Zhengyang Liang[2], Shitao Xiao[2],
Minghao Qin[2], Xi Yang[2], Yongping Xiong[1], Bo Zhang[4], Tiejun Huang[2,5], Zheng Liu[†2]
[1] State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
[2] Beijing Academy of Artificial Intelligence
[3] School of AI, Shanghai Jiao Tong University [4] Zhejiang University [5] Peking University

Abstract

The evaluation of Long Video Understanding (LVU) performance poses an important but challenging research problem. Despite previous efforts, the existing video understanding benchmarks are severely constrained by several issues, especially the insufficient lengths of videos, a lack of diversity in video types and evaluation tasks, and the inappropriateness for evaluating LVU performances. To address the above problems, we propose a new benchmark called MLVU (Multi-task Long Video Understanding Benchmark) for the comprehensive and in-depth evaluation of LVU. MLVU presents the following critical values: 1) The substantial and flexible extension of video lengths, which enables the benchmark to evaluate LVU performance across a wide range of durations. 2) The inclusion of various video genres, such as movies, surveillance, egocentric videos, and cartoons, reflects the models' LVU performances in different scenarios. 3) The development of diversified evaluation tasks, which enables a comprehensive examination of MLLMs' key abilities in long-video understanding. The empirical study with 23 latest MLLMs reveals significant room for improvement in today's technique, as all existing methods struggle with most of the evaluation tasks and exhibit severe performance degradation when handling longer videos. Additionally, it suggests that factors such as context length, image-understanding ability, and the choice of LLM backbone can play critical roles in future advancements. We anticipate that MLVU will advance the research of LVU by providing a comprehensive and in-depth analysis of MLLMs. The code and dataset can be accessed from https://github.com/JUNJIE99/MLVU.

1. Introduction

Large language models (LLMs) are growing into a general solution for numerous AI tasks [6, 48]. In recent years, it becomes increasingly emphasized to extend LLMs with multi-modal capabilities and thus bring the Multimodal LLM (MLLM). Remarkably, it has been made possible for today's MLLMs to perceive information in texts, images, videos, etc., and solve complicated problems in physical environments [1, 47]. Along with the development of MLLMs, new benchmarks are continuously created to facilitate comprehensive and in-depth analysis of MLLMs [12, 27, 33, 34, 60].

However, it remains a great challenge to evaluate the MLLMs' long-video understanding (LVU) performances given the following limitations. Firstly, the majority of existing video understanding benchmarks are made up of short videos [20, 23, 27, 38, 55], whose lengths can be merely a few seconds. As a result, they are insufficient to reflect the MLLMs' long-video understanding capabilities. Secondly, there is a notable lack of diversity in both video genres and evaluation tasks. Existing benchmarks often concentrate on a single video type, such as egocentric videos [15, 36], or focus on one specific task, like captioning [55]. These limitations hinder comprehensive evaluation of LVU capabilities. Last but not least, many previous evaluation tasks are not properly designed for LVU, as they can be solved without using the complex information from long videos. For example, many questions are simply about one single frame in the long videos [44, 63]. Besides, numerous others are about popular movies and celebrities [13, 28], which can be answered directly by MLLMs based on the textual prompts.

Conceptually, MLLMs are expected to handle any type of long video and accomplish any related tasks. Therefore, the evaluation of LVU should emphasize two important properties: *length* and *diversity*. Furthermore, it is crucial that the evaluation tasks are specifically designed to leverage the complex information inherent in long videos, addressing the shortcomings of previous benchmarks. Based on these principles, we propose a novel benchmark called **MLVU** (Mult-task Long Video Understanding Benchmark), which presents the following critical advantages.

- **It makes a substantial extension for the video length.** MLVU is created based on long videos of diversified

[*]Co-first authors
[†]Corresponding author

Benchmarks	#Videos	#QA Pairs	Len. (s)	Close-Ended	Open-Ended	Various Genres	Multi-Level	Multi-Dimension	Referring QA
NExT-QA [53]	1,000	8,564	39.5	✓	✓	✓	✗	✗	✗
TVQA [22]	15,253	15,253	11.2	✓	✗	✗	✗	✗	✗
MSRVTT-QA [55]	2,900	72,821	15.2	✓	✗	✗	✗	✗	✗
MVBench [27]	3,641	4,000	16.0	✓	✗	✓	✗	✗	✗
Movie101 [61]	101	-	6144	✗	✓	✗	✗	✗	✗
EgoSchema [36]	5,063	5,063	180	✓	✗	✗	✗	✗	✗
MovieChat-1K [44]	130	1,950	500	✓	✓	✗	✗	✓	✗
Video-MME* [13]	900	2,700	1024	✓	✗	✓	✓	✗	✗
LongVideoBench* [52]	3,763	6,678	473	✓	✗	✓	✓	✗	✓
MLVU	1,730	3,102	930	✓	✓	✓	✓	✓	✓

Table 1. Comparison of MLVU with existing benchmarks, including the number of videos (**#Videos**), number of QA pairs (**#QA pairs**), average video length (**Len.**), presence of **Close-Ended** tasks, presence of **Open-Ended** tasks, inclusion of various video genres (**Various Genres**), coverage of multiple duration levels (**Multi-Level**), inclusion of multiple dimensions of LVU tasks (**Multi-Dimension**), and questions involving local information with clear referring context rather than direct timestamps [44] or well-known narrative elements [18, 28] (**Referring QA**). The first block represents short video understanding benchmarks, and the second block represents long video understanding benchmarks. * denotes work concurrent with MLVU.

lengths, *ranging from 3 minutes to 2 hours*. The average video length is about 15 minutes, which makes it much longer than most of the existing benchmarks. Additionally, each video is further segmented so that evaluation tasks can be created w.r.t. different video clips (e.g., summarization for the first 3 minutes, the first 6 minutes, and the entire duration of the video). Therefore, it is able to flexibly evaluate the MLLMs' performance across different video lengths.

- **It encompasses a wide variety of video genres.** MLVU includes diverse real-world videos, such as movies, life records, and egocentric videos. Additionally, it features typical simulated videos like games and cartoons. This diversity allows for a comprehensive assessment of MLLMs' performance across various application scenarios.
- **It introduces diversified evaluation tasks tailored for LVU.** MLVU comprises 9 distinct tasks that collectively assess a wide range of MLLMs' LVU capabilities. On one hand, it includes both *multiple-choice and open-ended generation* tasks, reflecting the models' performance in handling different task formats. On the other hand, some tasks are designed to leverage *global information from entire videos*, while others require the use of *specific local information from certain clips*. Moreover, all questions involving local information are annotated with unambiguous context, requiring MLLMs to accurately locate or infer the appropriate clips within long videos.

Table 1 shows that MLVU provides a more comprehensive evaluation of LVU compared to existing and concurrent benchmarks. We extensively investigate 23 popular MLLMs with MLVU, which brings in several critical insights. Firstly, *long-video understanding remains a technically challenging problem for the existing MLLMs*. While GPT-4o [39] achieves the leading performance in the experiment, it only attains an average score of 54.5% in multi-choice tasks. All methods struggle with tasks requiring fine-grained information from entire videos, such as action counting, ordering, and summarization. Secondly, *recent open-source long video MLLMs have made significant strides in LVU* [11, 43, 63]. These advancements have improved the models' capability to process extended visual sequences, thereby closing the gap with leading proprietary models in recent months. Finally, *the empirical results underscore influential factors in LVU*, such as the extension of context length, the improvement of image understanding ability, and the utilization of strong LLM-backbones. In addition to the benchmark's overall conclusion, individual tasks enable fine-grained analysis of MLLMs' performances in each specialized aspects. Therefore, we anticipate the benchmark to assist in improving MLLMs' long-video understanding capabilities by providing insights into their current strengths and weaknesses.

2. Related Work

Multimodal Large Language Models. Multimodal large language models (MLLMs) have attracted significant interest from both academia and industry. Recent advancements in this field have been achieved by integrating LLM backbones with visual encoders and adapters, and fine-tuning the entire architecture through visual instruction tuning [8, 30, 66]. Based on the same philosophy, MLLMs have been further developed for video processing using video instruction datasets and specialized video adapters [26, 27, 29, 35, 57, 62]. However, most existing models are optimized for short videos, typically under one minute, due to the difficulty in establishing sufficient context for longer videos. To address this challenge, researchers have explored compact video representations or extended the context length of MLLMs []. For instance, LLaMa-Vid [28] compresses each video frame

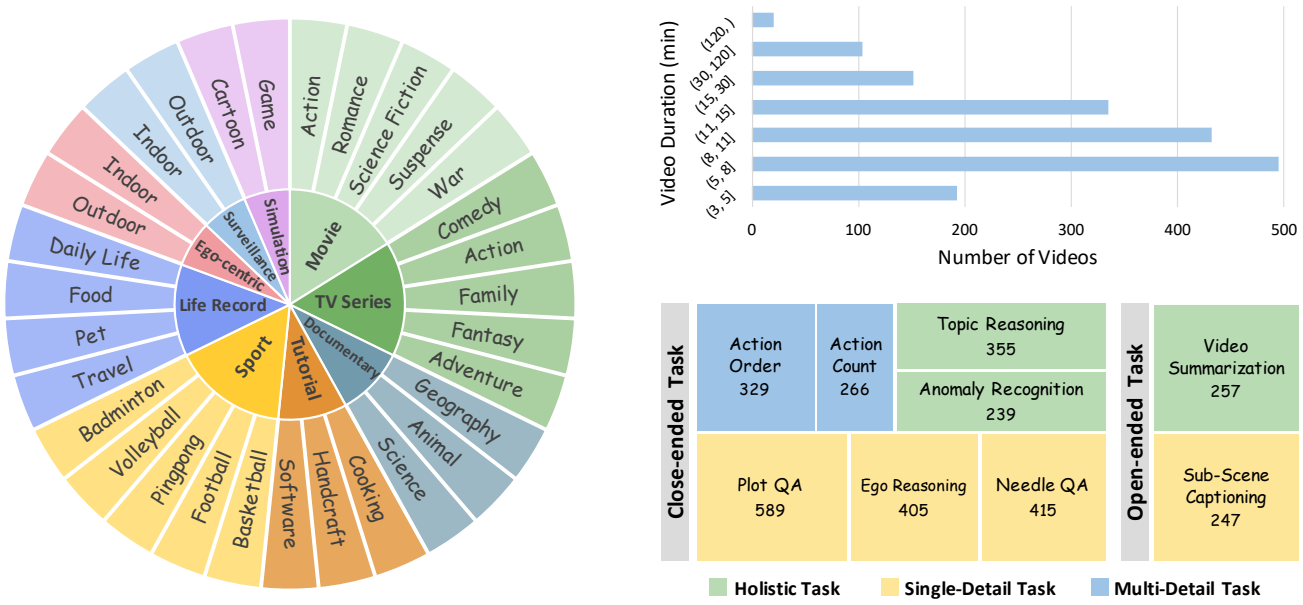

Figure 1. Statistical Overview of our MLVU benchmark. **Left:** Video genres included in MLVU; **Top Right:** Distribution of video duration; **Bottom Right:** Task types and their counts in MLVU.

into two tokens, enabling the model to handle videos several hours long. Methods like MovieChat [44] and MA-LMM [17] introduce specialized memory components for recursive video processing. Furthermore, approaches such as LWM [31], LongVA [63], and Video-XL [43] are designed to extend the context length of MLLMs, facilitating the processing of longer video inputs [16, 41]. Additionally, it is also explored to make selective usage of frames or clips from long videos based on retrievers or agents [40, 49, 56]. Despite these progresses, it remains an open problem for MLLMs to effectively handle long videos.

Video Understanding Benchmarks. With the unprecedented interest in MLLMs, the creation of benchmarks for these models has become increasingly emphasized (as advanced by MMMU [60], MME [12], and many other pioneering works). In video understanding, the research community has made significant efforts as well, particularly for short videos. There are specialized benchmarks for temporal perception [51, 59], action understanding [50, 51], video classification [19], video reasoning [53, 54], and video captioning [37, 55]. Recently, MVBench [27] provides a comprehensive short-video benchmark to evaluate general capabilities via question-answering. For long video understanding, people seek to leverage long-form videos, like movies, to create benchmarks. For example, LLaMA-Vid [28] developed a movie question-answering dataset based on MovieNet [18]. Despite using long videos, many questions focus on well-known narrative elements, allowing them to be answered without analyzing the video's content. In contrast, MovieChat [44] avoids specific character names or plot details in its questions. However, since each question provides a specific timestamp, the tasks can be reduced to short-video or image understanding problems. Beyond movies, there are task-specific benchmarks like EgoSchema [36], which presents video reasoning tasks using first-person footage from Ego4D [15]. These specialized benchmarks, however, focus on a single aspect of MLLMs rather than offering a comprehensive analysis of long video understanding. Therefore, it is essential to develop a comprehensive benchmark with carefully designed tasks to effectively evaluate MLLMs' capabilities in understanding long videos.

3. MLVU: Multi-task Long Video Understanding Benchmark

In this section, we start with an overview of MLVU, which highlights its constitution and explains its values over the previous works. Then, we discuss how each evaluation task is constructed in MLVU.

3.1. Overview

MLVU is a multi-task benchmark consisting of 3,102 questions across 9 categories, specifically designed for long video understanding. It is divided into a dev set and a test set, containing 2,593 and 509 questions, respectively. The benchmark is distinguished by the following features.

Diversified Video Categories. MLVU offers a comprehensive collection of videos across various categories (Figure 1 Left). These include typical real-world videos such as movies, documentaries, TV series, egocentric videos, life records, sports, tutorials, and surveillance footage. Additionally, it features significant simulated videos from animated series and game videos.

Substantial Extension of Video Length. MLVU is made up of videos of diversified lengths, spanning from 3 min to

more than 2 hours (Figure 1 Top Right). Besides, each video is further partitioned as incremental segments, e.g., the first 3 min, the first 6 min, and the entire video, where tasks are created for each individual segment. Thus, the MLLMs can be flexibly evaluated across different video lengths.

Diversified Evaluation Tasks. MLVU also provides a diverse array of evaluation tasks, which are closely aligned with the common visual capabilities of MLLMs, such as reasoning, captioning, recognition, perception, and summarization (Figure 1 Bottom Right). All the tasks are tailored for LVU. That is to say, the tasks need to be solved based on the in-depth understanding of video. Some of tasks are to examine whether the global information from the entire video can be effectively utilized (holistic LVU); while others focus on whether the MLLMs can make precise usage of proper local information within the long video (detail LVU). Additionally, both multi-choice and free-form generation tasks are included in MLVU, which help to examine MLLMs' capabilities in handling different task formats.

3.2. Construction of MLVU

The evaluation tasks of MLVU can be categorized into three types: 1) *holistic LVU*, which needs to be solved by making use of the global information from the entire video; 2) *single-detail LVU*, which relies on leveraging one critical plot within the long video; and 3) *multi-detail LVU*, which necessitates the joint utilization of multiple plots within the long video. The construction process of MLVU is discussed w.r.t the above three categories. To facilitate the discussion, we define *ULVC* (Universal Long Video Collection) as the universal collection of long videos from various sources (more details about ULVC are presented in Appendix C).

3.2.1. Holistic LVU

Topic Reasoning (TR). The topic reasoning task requires MLLMs to respond to questions about the principal subject of a long video, as shown with Figure 2 (a). This includes elements such as the video's genre, pivotal events, or primary settings. All questions and answers undergo manual annotation, resulting in a total of 355 questions. TR tasks are formatted as multiple-choice questions, with the model's performance assessed based on accuracy.

Anomaly Recognition (AR). The anomaly recognition task involves identifying the anomalous behavior within a surveillance footage (Figure 2 b). We leverage the surveillance video clips from UCF Crime dataset [46] for this task. The selected video clips are longer than three minutes. We create 239 questions based on the original annotations provided by the dataset. The AR task is also conducted in the multiple-choice format, whose performance is measured by accuracy.

Video Summarization (VS). This task requires MLLMs to summarize the key events in a long video (Figure 2 c). We select the narrative-rich videos from ULVC for this task, including movies, TV series, documentaries, life records, and animated series. There are 257 selected videos in total, whose summaries are manually annotated. During evaluation, the MLLMs are prompted with "Please summarize the main content of this video". We employ GPT-4 to assess the generated summaries by comparing with the annotation results. Details about annotation and evaluation are presented in Appendix F.3 and G.3.

3.2.2. Single-Detail LVU

Needle Question-Answering (NQA). Needle-In-the-Haystack-Search (NIHS) is a popular evaluation task for long-context LLM [32]. Taking the inspiration from NIHS, we create Needle Question-Answering (NQA), shown as Figure 2 (d). In this task, the MLLM is required to answer a question related to a specific segment (referred as *needle*) within a long video (referred as *background video*). The needles are short video clips sampled from WebVid [5] and Clevrer [58], while the background videos are sampled from our ULVC. The needle is randomly inserted into the background video, where a question-answer pair is annotated. By incorporating necessary details, the question can always correspond to the needle without ambiguity. During evaluation, the MLLM needs to infer the location of the needle based on the details provided in the question, and solve the problem on top of the needle's information. The NQA task is structured as multiple-choice, whose performance is measured by accuracy.

Ego Reasoning (ER). Ego-centric videos capture a series of consecutive actions from a first-person perspective. The MLLM needs to reason for a question about a specific behavior in the video, e.g., predicting for the event which is correlated or satisfies a certain causal relationship with the behavior (Figure 2 e). Both videos and QA annotations are collected from the NLQ task of Ego4D [15]. The ER task is structured as multiple-choice, with a total of 405 questions created for this task.

Plot Question-Answering (PQA). In this task, the MLLM needs to reason for questions about a plot in a narrative video, shown as Figure 2 (f). The video is sampled from the movies, TV series, and animated series in our ULVC. There are 589 question-answer pairs created by manual annotation. During annotation, the human annotators are asked to only provide necessary details about the plot but not to suggest any objective hints, e.g., the two characters in the example video are referred as cat and mouse, rather than Tom and Jerry. Therefore, it can prevent the question from being short-cut by the MLLM's common-sense knowledge (more details about PQA can be found in the Appendix F.6).

Sub-Scene Captioning (SSC). In this task, the MLLM needs to generate the caption for a sub-scene in a long video. The long videos in SSC are sampled from the Movie101 dataset [61], while the questions and answers are manually annotated. During annotation, the human annotator is asked to provide a detailed description for the sub-scene as the

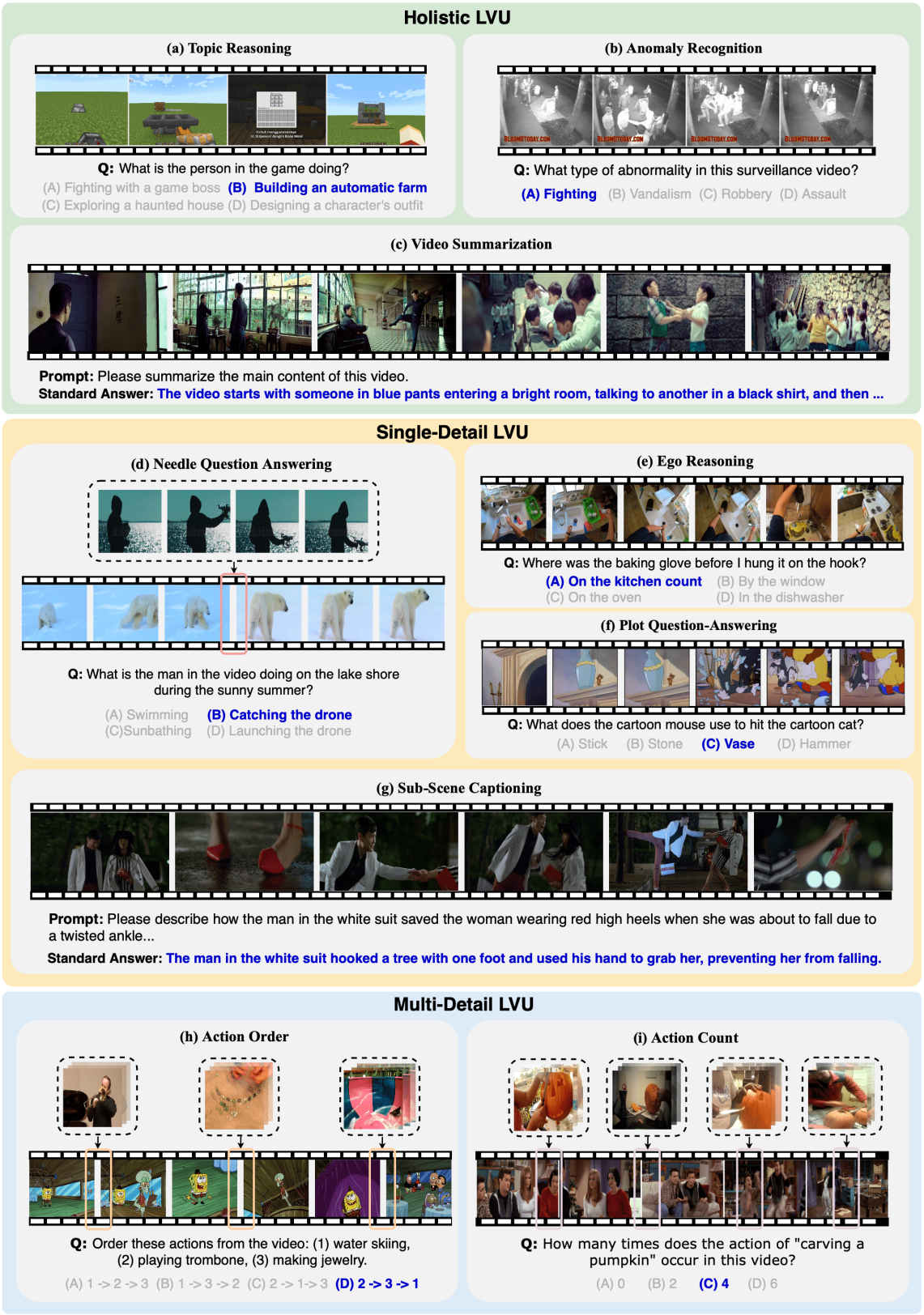

Figure 2. Examples of MLVU. There are nine tasks designed to evaluate the *holistic*, *single-detail*, and *multi-detail LVU* capabilities of MLLMs. The MLLMs are asked to solve the problem (with the ground-truth answers marked in blue) based on the long video input and textual prompt. For multiple-choice questions, we set 4 candidates in the dev set and 6 candidates in the test set.

ground-truth answer. Besides, they need to offer necessary clues in their questions such that the referred sub-scenes can be identified without ambiguity. During evaluation, we employ GPT-4 [1] to measure the quality of caption in comparison with the ground-truth. Details about annotation and evaluation are presented in Appendix F.7 and G.3.

3.2.3. Multi-Detail LVU

Action Order (AO). In this task, the MLLM needs to predict the right order for a sequence of actions (Figure 2 h). The actions are presented by short video clips, called *probes*. The probes are formulated in two different ways. One is made up of clips from the Kinetics dataset [19], where each clip represents a distinct action. The other one is from the consecutive clips of an action in the ActivityNet-Caption dataset [21]. The probes are inserted into a long *background* video, which is sampled from ULVC. There are 329 AO questions in total. The task is structured as a multiple-choice prblem, where the right order is selected from the misleading options provided by the annotator.

Action Count (AC). This task requires the MLLM to count the occurrences of an action within a long video (Figure 2 i). Each action corresponds to multiple short *probe* clips sampled from the Kinetics dataset [19]. The probes of an action are inserted into a long *background* video sampled from ULVC. We also perform manual examination to ensure that the inserted action does not exist in the original background video. A total of 266 evaluation instances have been created. The AC task is structured as a multiple-choice problem, with performance measured by accuracy.

4. Experiments and Analysis

4.1. Settings

We conduct a comprehensive investigation of 23 MLLMs using our MLVU benchmark, encompassing both open-source and proprietary models. The experimental MLLMs are divided into three categories: ***1) Image MLLMs***, primarily fine-tuned using image-related instructions; ***2) Short Video MLLMs***, fine-tuned with short-video related instructions; and ***3) Long Video MLLMs***, optimized for long-video understanding capability. For Image MLLMs, we leverage their multi-image inference capabilities to process segmented frames from original videos. For Video MLLMs, we employ either a uniform sampling strategy or a frame rate sampling strategy for video processing. All models are evaluated based on their official implementations or available APIs, with evaluations conducted in a zero-shot manner. More details about the evaluation are provided in Appendix G.

4.2. Main Results

The overall evaluation results for all investigated MLLMs in the MLVU test set are shown in Table 2 (with dev set results in Appendix B). Individual performances are reported for each task, while average performances are provided for multiple-choice (M-Avg) and generation tasks (G-Avg). From the results, we derive three primary conclusions:

1) The proprietary model GPT-4o [39] achieves optimal performance in our benchmark. It leads in multiple-choice tasks with an M-Avg of 54.5%(within 0-100%) and excels in generation tasks with a G-Avg of 5.87 (within 0.0-10.0), outperforming all other methods.

2) Recent advances in LVU have achieved significant progress, and the gap between open-source long video MLLMs and GPT-4o on close-ended tasks is narrowing. Before June 2024, the best open-source long video MLLMs, MiniGPT4-Video [3], lagged significantly behind GPT-4o. However, recent models [11, 25, 43, 63] have made substantial progress. For instance, LLaVA-Onevision trails GPT-4o by only 2.8% in M-Avg. These models have improved their ability to handle long visual sequences, achieving significant advancements in single-detail (e.g., NQA) and multi-detail (e.g., AC) tasks compared to previous open-source models.

3) Existing methods still struggle to handle most tasks in our benchmark. For instance, GPT-4o only achieves 42.9% in the needle question-answering (NQA) task. In contrast, analogous tasks in the text domain, such as NIHS (Needle-In-the-Haystack-Search) and Passkey Retrieval, are effectively handled by many existing long LLMs [14, 64]. Additionally, GPT-4o shows even less reliability in tasks like ego-reasoning (ER), action ordering (AO), and action count (AC), with most baseline methods performing even worse. These observations indicate that long-video understanding remains a significant challenge for today's MLLMs.

In addition to the primary conclusions from the overall performances, we can also make the following interesting observations about the individual tasks.

4) The close-ended holistic tasks present much higher differentiation than other tasks. These tasks, i.e., topic reasoning (TR) and anomaly recognition (AR), show significant variance in performance across different models. Proprietary MLLMs, like GPT-4o, and superior open-source models, such as InternVL-2 [8], VideoLLaMA2 [9], and LLaVA-OneVision [25], can accurately solve these problems. Meanwhile, many other popular MLLMs still fail to generate meaningful performances. Since these tasks only require an overall understanding of long videos, they can serve as a preliminary indicator of MLLMs' LVU ability.

5) It's challenging to deal with tasks that require nuanced understanding of multiple details. Although several MLLMs can handle single-detail LVU tasks to some extent, their performances suffer from catastrophic degradation when addressing multi-detail LVU tasks. Most methods, except for GPT-4o and Video-XL [43], fail entirely in action order (AO) and action count (AC) tasks. Additionally, most approaches struggle with summarization tasks, which require recalling multiple nuanced details from long videos.

Methods	Date	Input	Holistic			Single Detail				Multi Detail		M-Avg	G-Avg
			TR	AR	VS*	NQA	ER	PQA	SSC*	AO	AC		
Full mark	–	–	100	100	10	100	100	100	10	100	100	100	10
Random	–	–	16.7	16.7	–	16.7	16.7	16.7	–	16.7	16.7	16.7	–
Image MLLMs													
Otter-I [24]	2023-05	16 frm	17.6	17.9	2.03	16.7	17.0	18.0	3.90	15.7	16.7	17.1	2.97
LLaVA-1.6 [30]	2024-01	16 frm	63.7	17.9	2.00	13.3	26.4	30.0	4.20	21.4	16.7	27.1	3.10
InternVL-2 [8]	2024-07	16 frm	85.7	51.3	2.55	48.3	47.2	52.0	5.25	32.9	15.0	47.5	3.90
Claude-3-Opus† [2]	2024-03	16 frm	53.8	30.8	2.83	14.0	17.0	20.0	3.67	10.0	6.7	21.8	3.25
Qwen-VL-Max† [4]	2024-01	16 frm	75.8	53.8	3.00	15.0	26.4	4.84	20.0	20.7	11.7	32.2	3.92
Short Video MLLMs													
Otter-V [24]	2023-05	16 frm	16.5	12.8	2.18	16.7	22.6	22.0	4.20	12.9	13.3	16.7	3.19
mPLUG-Owl-V [57]	2023-04	16 frm	25.3	15.4	2.20	6.7	13.2	22.0	5.01	14.3	20.0	16.7	3.61
VideoChat [26]	2023-05	16 frm	26.4	12.8	2.15	18.3	17.0	22.0	4.90	15.7	11.7	17.7	3.53
Video-LLaMA-2 [62]	2024-08	16 frm	52.7	12.8	2.23	13.3	17.0	12.0	4.87	15.7	8.3	18.8	3.55
VideoChat2-HD [27]	2024-06	16 frm	74.7	43.6	2.83	35.0	34.0	30.0	5.14	21.4	23.3	37.4	3.99
Video-LLaVA [29]	2023-11	8 frm	70.3	38.5	20.9	2.30	26.4	26.0	5.06	20.0	21.7	29.3	3.68
ShareGPT4Video [7]	2024-05	16 frm	73.6	25.6	2.53	31.7	45.3	38.0	4.72	17.1	8.3	34.2	3.63
VideoLLaMA2 [9]	2024-06	16 frm	80.2	53.8	2.80	36.7	54.7	54.0	5.09	42.9	16.7	48.4	3.95
Long Video MLLMs													
MovieChat [44]	2023-07	2048 frm	18.7	10.3	2.30	23.3	15.1	16.0	3.24	17.1	15.0	16.5	2.77
Movie-LLM [45]	2024-03	1 fps	27.5	25.6	2.10	10.0	11.3	16.0	4.93	20.0	21.7	18.9	3.52
LLaMA-VID [28]	2023-11	1 fps	20.9	23.1	2.70	21.7	11.3	16.0	4.15	18.6	15.0	18.1	3.43
MA-LMM [17]	2024-04	1000 frm	44.0	23.1	3.04	13.3	30.2	14.0	4.61	18.6	13.3	22.4	3.83
MiniGPT4-Video [3]	2024-04	90 frm	64.9	46.2	2.50	20.0	30.2	30.0	4.27	15.7	15.0	31.7	3.39
LongVA [63]	2024-06	256 frm	81.3	41.0	2.90	46.7	39.6	46.0	4.92	17.1	23.3	42.1	3.91
Video-CCAM [11]	2024-08	96 frm	79.1	38.5	2.65	45.0	52.8	56.0	4.49	24.3	26.7	46.1	3.57
Video-XL [43]	2024-09	256 frm	78.0	28.2	3.40	50.0	41.5	46.0	5.02	48.6	31.7	46.3	4.21
LLaVA-Onevision [25]	2024-08	32 frm	83.5	56.4	3.75	46.7	58.4	58.0	5.09	35.7	23.3	51.7	4.42
GPT-4o† [39]	2024-05	0.5 fps	83.7	68.8	4.94	42.9	47.8	57.1	6.80	46.2	35.0	54.5	5.87

Table 2. The overall performances on MLVU test set, including the holistic LVU tasks, the single-detail LVU tasks, and multi-detail LVU tasks. Date: the release date of the MLLM. M-Avg: the average performance of multiple-choice tasks; G-Avg: the average performance of generation tasks (marked by $*$). Two input strategies are used by the MLLMs in evaluation: Uniform Sampling (**N frm**), which evenly samples N frames from the video; Frame Rate Sampling (**N fps**), which samples N frames per second. † denotes proprietary models.

As a brief conclusion, although today's MLLMs can deal with some preliminary LVU tasks, it remains a tough challenge to achieve an in-depth understanding of nuanced information within long videos.

4.3. Further Analysis

6) Longer videos are more challenging for MLLMs. We evaluate MLLMs' performances across various video lengths. For this purpose, we introduce a derivative dataset alongside MLVU, called *MLVU Time-ladder*. In this dataset, the same kinds of evaluation tasks are created for videos of variant lengths, including 180s, 360s, and 600s (more details presented in Appendix D). As shown in Figure 3, the performances of all models tend to decline as the video length grows, which indicates that the existing MLLMs' LVU abilities are severely constrained by the video length. Moreover, the short video model Video-LLaMA-2 [62] maintains a certain level of LVU ability at 3 minutes, but its performance approaches random results at 10 minutes.

7) The performance of recent advanced long video MLLMs remains robust regardless of the position of the referring clip within the long video. In single-detail tasks, the referring clip denotes the specific segment of the long video that is referenced or inferred to answer a question. As shown in Figure 4, we categorize clip positions into four intervals and assess model performance on two single-detail tasks: ego reasoning (ER) and plot question-answering (PQA). Recent long video MLLMs, such as LongVA [63] and Video-XL [43], maintain consistent performance regardless of the referring clip's position within the video. Conversely, short video MLLMs are more sensitive to clip location. This indicates that recent advancements in long video MLLMs enhance both reliable clue retrieval and effective reasoning from extended visual sequences.

Impact of Context Length			Impact of IU			Impact of LLM		
Model	Context Len.	M-Avg	Model	MMMU (Val)	M-Avg	Model	LLM	M-Avg
MGV	16	24.2	Otter-I	32.2	17.1	VLM2	Vicuna-7B	13.3
	90	31.7↑7.5	LLaVA-1.6	35.8	27.1↑10.0		Vicuna-13B	18.8↑5.5
GPT-4o	16	45.8	GPT-4V	58.1	43.3	MGV	LLaMA-7B	20.6
	256	54.5↑8.7	GPT-4o	63.8	45.8↑2.5		Mistral-7B	31.7 ↑11.1

Table 3. Detailed discussions about the impact from context length, image understanding (IU) ability, and LLM Backbone. For the IU impact experiment, we used 16-frame uniform sampling for both GPT-4V and GPT-4o. MGV: MiniGPT4-Video, VLM2: Video-LLaMA-2.

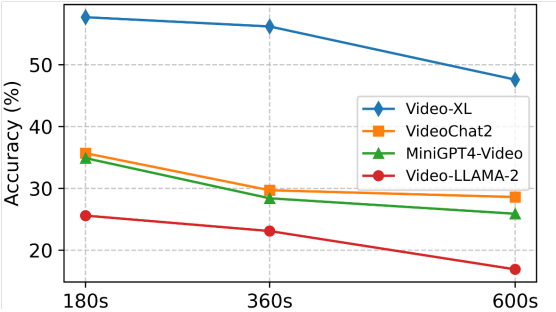

Figure 3. Experimental performance on varying video lengths. The evaluated metric is the average accuracy across five multiple-choice tasks involving local information: NQA, ER, PQA, AC, and AO.

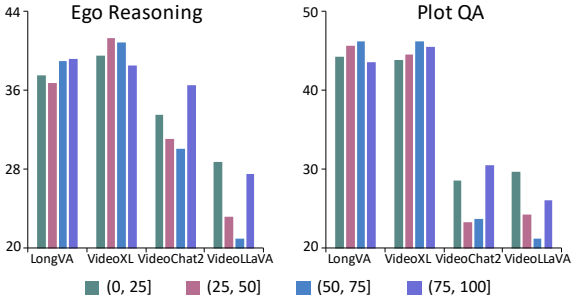

Figure 4. Model performance across different referring clip positions, spanning from the beginning to the end of the entire video.

8) The challenge of multi-detail tasks increases with the number of details. We analyzed model performance on the action count (AC) task by grouping questions based on the number of probes (which correspond to details) and evaluating the average performance within these groups. As shown in Figure 5, performance significantly declines across all models as the number of probes increases. This indicates that current MLLMs face substantial difficulties comprehending and processing multiple details simultaneously, highlighting a critical area for future improvement in long video understanding capabilities.

9) Context Length, Image-Understanding ability, and the choice of LLM Backbones are key factors in LVU performance. As shown in Table 3, we conducted ablation experiments on several factors affecting MLLMs, using M-Avg as the evaluation metric. First, we examined the models' handling of different context lengths. Specifically, we increased MiniGPT4-Video's input from 16 to 90 frames

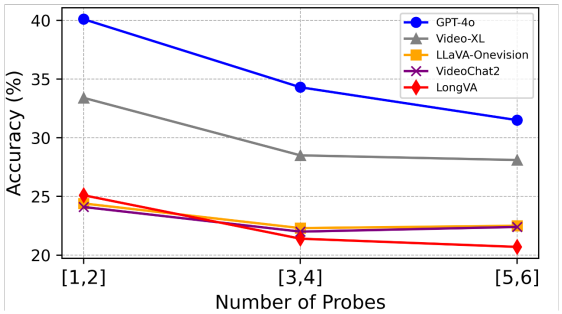

Figure 5. Model performance on the action count (AC) task in relation to the number of probes.

and GPT-4o's input from 16 to 256 frames (as shown on the left side of Table 3). Both models showed consistent performance improvements with longer input lengths. To assess the impact of image understanding (IU) capabilities, we referred to the results from MMMU [60] (presented in the middle of Table 3). It is evident that MLLMs' LVU performance generally aligns with their IU performance in MMMU. Finally, we compared MLLMs using different backbones (depicted on the right side of Table 3). The findings indicate that LVU performance improves with larger (Vicuna-13B vs. Vicuna-7B) and more advanced backbones (Mistral-7B vs. Llama-2-7B). These observations indicate that LVU is the result of multiple complex factors, with the ability to perceive longer videos and effectively utilize the perceived information being crucial for the improvement of LVU.

5. Conclusion

This paper presents MLVU, a novel benchmark for the assessment of long video understanding. With several critical innovations: the substantial extension of video lengths, the inclusion of various video genres, and the development of diversified LVU-oriented evaluation tasks, the new benchmark is able provide a comprehensive and in-depth analysis for MLLMs' long-video understanding performance. The empirical study on MLVU reveals LVU remains a technically challenging problem for today's state-of-the-art MLLMs. Future advancements may call for the joint optimization of complex factors, such as context length, image understanding ability, and even LLM backbones. We anticipate this benchmark will facilitate future research in long-video understanding of MLLMs.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 1, 6, 3

[2] Anthropic. Claude 3. https://www.anthropic.com/news/claude-3-family, 2024. 7, 2

[3] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. arXiv preprint arXiv:2404.03413, 2024. 6, 7, 2

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023. 7, 2

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1728–1738, 2021. 4, 3

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020. 1

[7] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. arXiv preprint arXiv:2406.04325, 2024. 7, 2

[8] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821, 2024. 2, 6, 7

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

[10] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. Advances in Neural Information Processing Systems, 35: 18343–18362, 2022. 1

[11] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. arXiv preprint arXiv:2408.14023, 2024. 2, 6, 7

[12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023. 1, 3

[13] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075, 2024. 1, 2

[14] Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context. arXiv preprint arXiv:2402.10171, 2024. 6

[15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18995–19012, 2022. 1, 3, 4

[16] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. arXiv preprint arXiv:2411.14794, 2024. 3

[17] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. arXiv preprint arXiv:2404.05726, 2024. 3, 7, 2

[18] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pages 709–727. Springer, 2020. 2, 3

[19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017. 3, 6

[20] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. Complex video reasoning and robustness evaluation suite for video-lmms. arXiv preprint arXiv:2405.03690, 2024. 1

[21] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision, pages 706–715, 2017. 6

[22] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696, 2018. 2, 4

[23] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125, 2023. 1

[24] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. CoRR, abs/2305.03726, 2023. 7, 2

[25] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chun-

yuan Li. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024. 6, 7

[26] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355, 2023. 2, 7

[27] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. arXiv preprint arXiv:2311.17005, 2023. 1, 2, 3, 7

[28] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. arXiv preprint arXiv:2311.17043, 2023. 1, 2, 3, 7

[29] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122, 2023. 2, 7

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2023. 2, 7

[31] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. arXiv preprint arXiv:2402.08268, 2024. 3

[32] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12:157–173, 2024. 4

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281, 2023. 1

[34] Yexin Liu, Zhengyang Liang, Yueze Wang, Muyang He, Jian Li, and Bo Zhao. Seeing clearly, answering incorrectly: A multimodal robustness benchmark for evaluating mllms on leading questions. arXiv preprint arXiv:2406.10638, 2024. 1

[35] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424, 2023. 2

[36] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems, 36, 2023. 1, 2, 3

[37] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In Proceedings of the IEEE/CVF international conference on computer vision, pages 2630–2640, 2019. 3

[38] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. arXiv preprint arXiv:2311.16103, 2023. 1

[39] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 2, 6, 7

[40] Junting Pan, Ziyi Lin, Yuying Ge, Xiatian Zhu, Renrui Zhang, Yi Wang, Yu Qiao, and Hongsheng Li. Retrieving-to-answer: Zero-shot video question answering with frozen large language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 272–283, 2023. 3

[41] Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, and Yanfeng Wang Weidi Xie. Towards universal soccer video understanding. arXiv preprint arXiv:2412.01820, 2024. 3

[42] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. arXiv preprint arXiv:2312.02051, 2023. 2

[43] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. arXiv preprint arXiv:2409.14485, 2024. 2, 3, 6, 7

[44] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. arXiv preprint arXiv:2307.16449, 2023. 1, 2, 3, 7, 4

[45] Zhende Song, Chenchen Wang, Jiamu Sheng, Chi Zhang, Gang Yu, Jiayuan Fan, and Tao Chen. Moviellm: Enhancing long video understanding with ai-generated movies. arXiv preprint arXiv:2403.01422, 2024. 7, 2

[46] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6479–6488, 2018. 4, 1, 2

[47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023. 1

[48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023. 1

[49] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. arXiv preprint arXiv:2403.10517, 2024. 3

[50] Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. Advances in Neural Information Processing Systems, 36, 2023. 3

[51] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2), 2021. 3

[52] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. arXiv preprint arXiv:2407.15754, 2024. 2

[53] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9777–9786, 2021. 2, 3

[54] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. arXiv preprint arXiv:2306.14899, 2023. 3

[55] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5288–5296, 2016. 1, 2, 3

[56] Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Retrieval-based video language model for efficient long video question answering. arXiv preprint arXiv:2312.04931, 2023. 3

[57] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178, 2023. 2, 7

[58] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In International Conference on Learning Representations, 2019. 4

[59] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 9127–9134, 2019. 3

[60] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502, 2023. 1, 3, 8

[61] Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. Movie101: A new movie understanding benchmark. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4669–4684, 2023. 2, 4, 1

[62] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023. 2, 7

[63] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. arXiv preprint arXiv:2406.16852, 2024. 1, 2, 3, 6, 7

[64] Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. ∞ bench: Extending long context evaluation beyond 100k tokens. arXiv preprint arXiv:2402.13718, 2024. 6

[65] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. arXiv preprint arXiv:2310.01852, 2023. 6

[66] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In The Twelfth International Conference on Learning Representations, 2023. 2