

MLVU: Benchmarking Multi-task Long Video Understanding

Junjie Zhou^{*1,2}, Yan Shu^{*1}, Bo Zhao^{*1,3}, Boya Wu¹, Zhengyang Liang¹, Shitao Xiao¹, Minghao Qin¹, Xi Yang¹, Yongping Xiong², Bo Zhang⁴, Tiejun Huang^{1,5}, Zheng Liu^{†1}

¹ Beijing Academy of Artificial Intelligence, ² Beijing University of Posts and Telecommunications,

³ Shanghai Jiao Tong University, ⁴ Zhejiang University, ⁵ Peking University

{junjiebupt, bozhaonanjing, zhengliu1026}@gmail.com

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, e.g., movies, surveillance footage, egocentric videos, cartoons, game videos, etc., which 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 long video understanding by providing a comprehensive and in-depth analysis of MLLMs.

1. Introduction

Large language models (LLMs) are growing into a general solution for numerous AI tasks [6, 45]. In recent years, it becomes increasingly emphasized to extend LLMs with multi-modal capabilities and thus bring the Multi-modal

LLM, namely, 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, 44]. Along with the development of MLLMs, new benchmarks are continuously created to facilitate comprehensive and in-depth analysis of MLLMs [12, 26, 32, 57].

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 [19, 22, 26, 36, 52], 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, 34], or focus on one specific task, like captioning [52]. 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 [41, 60]. Besides, numerous others are about popular movies and celebrities [13, 27], 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 lengths, ranging from 3 minutes to 2 hours. The average

^{*}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 [50]	1,000	8,564	39.5	✓	✓	✓	✗	✗	✗
TVQA [21]	15,253	15,253	11.2	✓	✗	✗	✗	✗	✗
MSRVTT-QA [52]	2,900	72,821	15.2	✓	✗	✗	✗	✗	✗
MVBBench [26]	3,641	4,000	16.0	✓	✗	✓	✗	✗	✗
Movie101 [58]	101	-	6144	✗	✓	✗	✗	✗	✗
EgoSchema [34]	5,063	5,063	180	✓	✗	✗	✗	✗	✗
MovieChat-1K [41]	130	1,950	500	✓	✓	✗	✗	✓	✗
Video-MME* [13]	900	2,700	1024	✓	✗	✓	✓	✗	✗
LongVideoBench* [49]	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 [41] or well-known narrative elements [17, 27] (**Referring QA**). The first block represents short video understanding benchmarks, and the second block represents long video understanding benchmarks. * denotes work concurrent with MLVU.

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-4o¹ 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, 40, 60]. 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, 29, 63]. Based on the same philosophy, MLLMs have been further developed for video processing using video instruction datasets and specialized video adapters [25, 26, 28, 33, 54, 59]. 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 [27] compresses each video frame

¹<https://openai.com/index/hello-gpt-4o/>

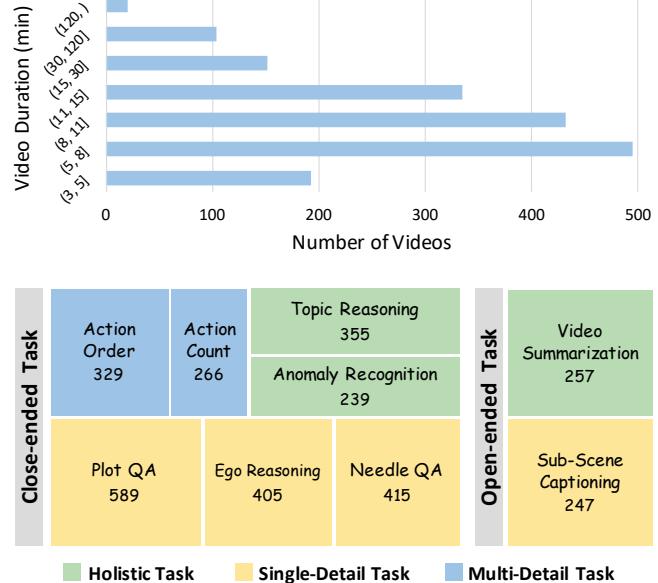
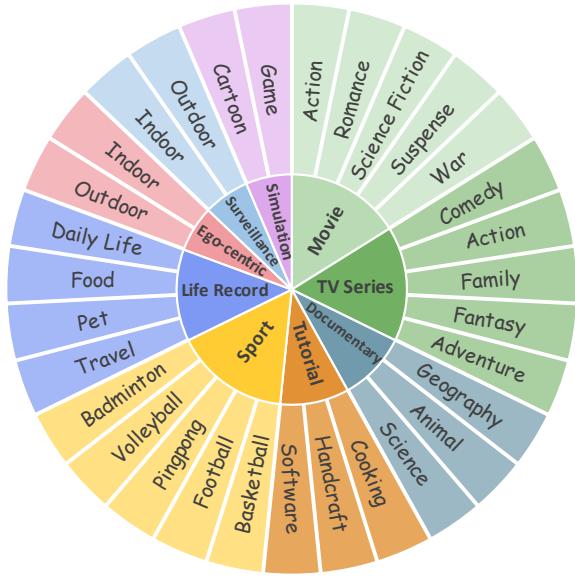


Figure 1. Statistical Overview of our MLVU benchmark. **Left:** Video genres included in MLVU; **Top Right:** Distribution of video duration; **Bottom Right:** Task types and their counts in MLVU.

into two tokens, enabling the model to handle videos several hours long. Methods like MovieChat [41] and MAMLMM [16] introduce specialized memory components for recursive video processing. Furthermore, approaches such as LWM [30], LongVA [60], and Video-XL [40] are designed to extend the context length of MLLMs, facilitating the processing of longer video inputs. Additionally, it is also explored to make selective usage of frames or clips from long videos based on retrievers or agents [38, 46, 53]. 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 [57], 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 [48, 56], action understanding [47, 48], video classification [18], video reasoning [50, 51], and video captioning [35, 52]. Recently, MV-Bench [26] 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 [27] developed a movie question-answering dataset based on MovieNet [17]. 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 [41] 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 [34], 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 annotation², 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 [43] 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.

²Detailed information and annotation guidelines for annotators are presented in Appendix F.

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 E3 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 [31]. 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 [55], 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

Holistic LVU

(a) Topic Reasoning

Q: What is the person in the game doing?
 (A) Fighting with a game boss **(B) Building an automatic farm**
 (C) Exploring a haunted house (D) Designing a character's outfit

(b) Anomaly Recognition

Q: What type of abnormality in this surveillance video?
(A) Fighting (B) Vandalism (C) Robbery (D) Assault

(c) Video Summarization

Prompt: Please summarize the main content of this video.
Standard Answer: The video starts with someone in blue pants entering a bright room, talking to another in a black shirt, and then ...

Single-Detail LVU

(d) Needle Question Answering

Q: What is the man in the video doing on the lake shore during the sunny summer?
 (A) Swimming **(B) Catching the drone**
 (C) Sunbathing (D) Launching the drone

(e) Ego Reasoning

Q: Where was the baking glove before I hung it on the hook?
 (A) On the kitchen count **(B) By the window**
 (C) On the oven (D) In the dishwasher

(f) Plot Question-Answering

Q: What does the cartoon mouse use to hit the cartoon cat?
 (A) Stick **(B) Stone** (C) Vase **(D) Hammer**

(g) Sub-Scene Captioning

Prompt: Please describe how the man in the white suit saved the woman wearing red high heels when she was about to fall due to a twisted ankle...
Standard Answer: The man in the white suit hooked a tree with one foot and used his hand to grab her, preventing her from falling.

Multi-Detail LVU

(h) Action Order

Q: Order these actions from the video: (1) water skiing, (2) playing trombone, (3) making jewelry.
 (A) 1 > 2 > 3 **(B) 1 > 3 > 2** (C) 2 > 1 > 3 (D) 2 > 3 > 1

(i) Action Count

Q: How many times does the action of "carving a pumpkin" occur in this video?
 (A) 0 **(B) 2** (C) 4 **(D) 6**

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.

long videos in SSC are sampled from the Movie101 dataset [58], 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 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 [18], where each clip represents a distinct action. The other one is from the consecutive clips of an action in the ActivityNet-Caption dataset [20]. 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 problem, 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 [18]. 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 [37] 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, 24, 40, 60] 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, 61]. 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 [24], 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’ long video understanding (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

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	–
<i>Image MLLMs</i>													
Otter-I [23]	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 [29]	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
<i>Short Video MLLMs</i>													
Otter-V [23]	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 [54]	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 [25]	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 [59]	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 [26]	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 [28]	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
<i>Long Video MLLMs</i>													
MovieChat [41]	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 [42]	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 [27]	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 [16]	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 [60]	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 [40]	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 [24]	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 [†] [37]	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 LUV tasks, the single-detail LUV tasks, and multi-detail LUV 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.

when addressing multi-detail LUV tasks. Most methods, except for GPT-4o and Video-XL [40], 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.

As a brief conclusion, although today’s MLLMs can deal with some preliminary LUV 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’ LUV abilities are severely constrained by the video length. Moreover, the short video model Video-LLaMA-2 [59] maintains a certain level of LUV 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 [60] and Video-XL [40], maintain consistent performance re-

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.

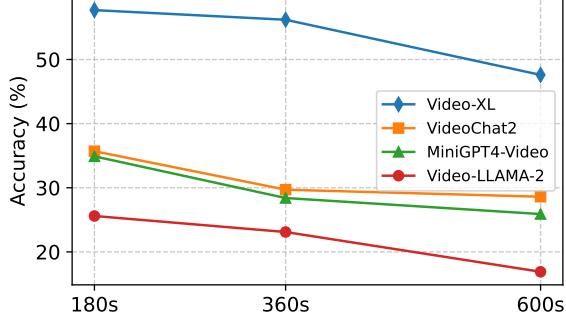


Figure 3. Experimental performance on varying video lengths. The evaluated metric is the average accuracy across five multiple-choice tasks involving local information: NQA, ER, PQA, AC, and AO.

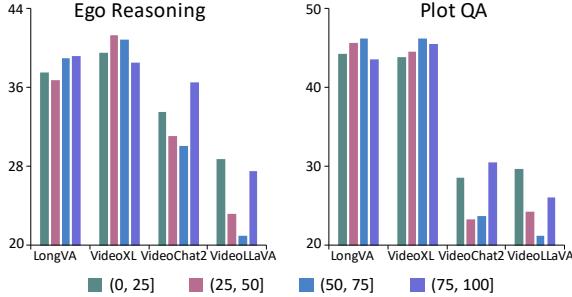


Figure 4. Model performance across different referring clip positions, spanning from the beginning to the end of the entire video.

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.

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

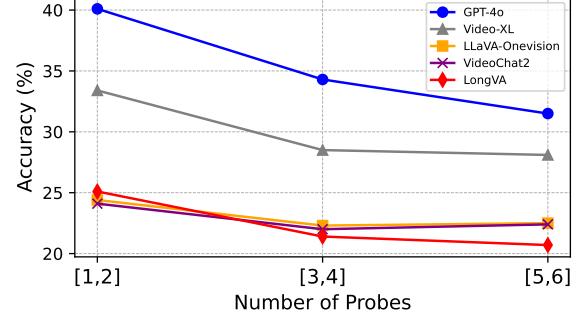


Figure 5. Model performance on the action count (AC) task in relation to the number of probes.

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 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 [57] (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 to 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/clause-3-family>, 2024. 7, 2
- [3] Kirolos Atallah, Xiaoqian Shen, Eslam Abdelrahman, Es-sam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhosseiny. 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, Yuhua 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] 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
- [17] 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
- [18] 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
- [19] 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
- [20] 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
- [21] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tqqa: Localized, compositional video question answering. [arXiv preprint arXiv:1809.01696](#), 2018. 2, 4
- [22] 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
- [23] 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
- [24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 7
- [25] 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
- [26] 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
- [27] 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
- [28] 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
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023. 2, 7
- [30] 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
- [31] 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
- [32] 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
- [33] 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
- [34] 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
- [35] 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
- [36] 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
- [37] OpenAI. Gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. 6, 7, 2
- [38] 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
- [39] 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
- [40] 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
- [41] 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
- [42] 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
- [43] 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
- [44] 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
- [45] 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
- [46] 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
- [47] 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
- [48] 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
- [49] 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
- [50] 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
- [51] 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
- [52] 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
- [53] 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
- [54] 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
- [55] 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
- [56] 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
- [57] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multi-modal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 1, 3, 8
- [58] 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, 6, 1
- [59] 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
- [60] 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
- [61] 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
- [62] 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
- [63] 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

MLVU: Benchmarking Multi-task Long Video Understanding

Supplementary Material

A. Overview of Appendix

- [B: Evaluation Results on MLVU Dev Set](#)
- [C: Collecting Details of our Universal Long Video Collection \(ULVC\)](#)
- [D: Details of the MLVU Time-Ladder](#)
- [E Detailed Division of Dev and Test Sets in MLVU](#)
- [F: Annotation Details of MLVU](#)
- [G: Details of Baselines and the Evaluation Process.](#)
- [H: Explorations of Video Retrieval Augmented Generation](#)
- [I: More Visualized Examples of MLVU](#)

B. Evaluation Results on MLVU Dev Set

The evaluation results of the baselines on the MLVU dev set are detailed in Table 2. Notably, the multiple-choice questions in the MLVU dev set present four options, whereas the MLVU test set offers six, making the latter more challenging and discriminative.

C. Collecting Details of our Universal Long Video Collection (ULVC)

In the initial stage of our Multi-task Long Video Understanding (MLVU) benchmark creation, we first collected long-form videos from a variety of sources to form our Universal Long Video Collection (ULVC). The entirety of the long videos incorporated into our MLVU benchmark were selected, edited, or synthesized from ULVC.

Specifically, our ULVC includes a diverse set of 986 long videos. This collection features 168 movies from the Movie101 [58] and MovieChat [41] datasets, along with 60 documentaries from MovieChat [41]. It also contains 65 game videos from MineDojo [10], 239 surveillance videos from UCF-Crime [43], and 100 ego-centric videos from Ego4D [15]. Additionally, we independently collected 72 cartoons, 92 TV series, 60 tutorial videos, 60 sports videos, and 70 life records.

It's important to clarify that the quantity of videos in the ULVC does not directly correspond to the number of videos and questions in our MLVU benchmark, which are 1,730 and 3,102 respectively. For example, a two-hour movie from the ULVC might be utilized in its entirety for the Sub-Scene Captioning task, or it could be segmented into several approximately 10-minute clips for the Video Summarization task, or even used as a background video for synthetic video generation. Moreover, a single video could be annotated with multiple questions simultaneously.

D. Details of the MLVU Time-Ladder

As discussed in Section 3.1, most tasks in our MLVU are subject to segment-level annotation. This approach provides us with the flexibility to adjust the length of the video without requiring additional human annotators. Building on this strategy, as mentioned in Section 4.3, we have generated a derivative dataset, *MLVU Time-Ladder*, which includes videos of varying durations - specifically 3, 6, and 10 minutes. This dataset allows us to investigate how video duration impacts LVU task difficulty.

Specifically, during the annotation process of the VS task, we guided annotators to delineate the summarization in accordance with the initial 3 and 6-minute segments. For the PQA and SSC tasks, we requested annotators to identify the segments within the extended video where the pertinent answers are located. In the case of the ego reasoning task, the Ego4D dataset [15] already comprises the intervals where the answers reside. Lastly, for the synthetic tasks of NQA, AO, and AC, we possess the capability to directly generate the necessary video lengths.

E. Detailed Division of Dev and Test Sets in MLVU

Our MLVU comprises a total of 3,102 questions, divided into a dev set with 2,593 questions and a test set with 509 questions. We present the detailed distribution of questions for each task in Table 1.

Task	Dev	Test	Total
Topic Reasoning	264	91	355
Anomaly Recognition	200	39	239
Video Summarization	217	40	257
Needle QA	355	60	415
Ego Reasoning	352	53	405
Plot QA	539	50	589
Sub-Scene Captioning	201	46	247
Action Order	259	70	329
Action Count	206	60	266

Table 1. Detailed Distribution of Questions in the MLVU Dataset Across Dev and Test Sets for Each Task.

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	–	–	25	25	–	25	25	25	–	25	25	25	–
<i>Image MLLMs</i>													
Otter-I [23]	2023-05	16 frm	25.0	25.0	2.18	25.1	25.0	24.9	4.12	13.1	25.2	23.3	3.15
LLaVA-1.6 [29]	2024-01	16 frm	60.6	41.0	2.11	43.1	38.4	41.0	4.35	25.5	25.7	39.3	3.23
Claude-3-Opus [†] [2]	2024-03	16 frm	67.2	43.5	3.11	21.6	40.2	47.8	3.66	18.2	16.7	36.5	3.39
Qwen-VL-Max [†] [4]	2024-01	16 frm	67.4	63.5	2.71	40.3	40.9	43.3	5.21	25.0	14.8	42.2	3.96
<i>Short Video MLLMs</i>													
Otter-V [23]	2023-05	16 frm	24.6	26.0	2.38	28.2	27.6	22.3	4.23	15.1	26.7	24.4	3.31
mPLUG-Owl-V [54]	2023-04	16 frm	28.0	25.0	2.36	24.5	31.8	27.3	5.31	21.2	23.3	25.9	3.84
VideoChat [25]	2023-05	16 frm	33.0	32.0	2.31	27.0	32.1	27.6	5.01	24.3	28.6	29.2	3.66
Video-LLaMA-2 [59]	2024-08	16 frm	54.5	41.5	2.34	39.4	33.5	35.4	5.22	18.5	25.7	35.5	3.78
VideoChat2-HD [26]	2024-06	16 frm	74.6	51.5	2.57	42.0	47.4	43.8	5.04	22.8	29.6	44.5	3.81
Video-LLaVA [28]	2023-11	8 frm	71.6	57.0	2.43	53.2	45.2	48.4	5.25	20.1	35.9	47.3	3.84
ShareGPT4Video [7]	2024-05	16 frm	75.8	51.5	2.52	47.6	43.2	48.4	5.02	34.0	23.3	46.4	3.77
VideoLLaMA2 [9]	2024-06	16 frm	74.6	64.5	2.79	49.9	43.8	45.1	5.18	34.0	27.4	48.5	3.99
<i>Long Video MLLMs</i>													
MovieChat [41]	2023-07	2048 frm	29.5	25.0	2.33	24.2	24.7	25.8	3.23	28.6	22.8	25.8	2.78
Movie-LLM [42]	2024-03	1 fps	30.0	29.0	2.88	29.6	24.7	24.1	5.00	20.5	24.8	26.1	3.94
TimeChat [39]	2023-12	96 frm	23.1	27.0	2.54	24.5	28.4	25.8	4.29	24.7	32.0	30.9	3.42
LLaMA-VID [27]	2023-11	1 fps	50.8	34.5	3.22	30.1	32.7	32.5	5.22	23.9	27.8	33.2	4.22
MA-LMM [16]	2024-04	1000 frm	51.9	35.5	2.12	43.1	38.9	35.8	4.80	25.1	24.3	36.4	3.46
MiniGPT4-Video [3]	2024-04	90 frm	70.9	52.5	2.64	49.0	48.6	44.5	4.07	23.2	23.0	44.5	3.36
LongVA [60]	2024-06	256 frm	83.3	58.5	3.39	69.3	50.0	67.2	5.26	38.6	27.2	56.3	4.33
Video-CCAM [11]	2024-08	96 frm	84.9	66.0	2.84	73.2	60.5	66.1	5.19	42.1	38.4	63.1	4.01
Video-XL [40]	2024-09	256 frm	80.3	54.5	3.25	73.8	57.4	67.9	5.02	68.3	40.3	64.9	4.14
GPT-4o [†] [37]	2024-05	0.5 fps	87.4	74.5	4.90	64.8	57.1	65.1	6.69	56.7	46.3	64.6	5.80

Table 2. The overall performances on MLVU dev set, including the holistic LVU tasks (TR: Topic Reasoning, AR: Anomaly Recognition, VS: Video Summary), the single-detail LVU tasks (NQA: Needle QA, ER: Ego Reasoning, PQA: Plot QA, SSC: Sub-Scene Captioning), and multi-detail LVU tasks (AO: Action Order, AC: Action Count). 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.

F. Annotation Details of MLVU

F.1. Topic Reasoning (TR).

The questions and corresponding answers for the TR task were meticulously annotated by human annotators, following the specific guidelines illustrated in Figure 1. We required the annotators to design questions related to the reasoning of the video topic, rather than focusing on the creation of questions about minor details. More visualized examples of TR task can be found in Figure 10.

F.2. Anomaly Recognition (AR).

The anomaly recognition task did not involve manual annotation. We utilized videos exceeding three minutes in duration, extracted from the UCF-Crime dataset [43]. We also modified the original labels to fit a multiple-choice format.

F.3. Video Summarization (VS).

The ground truth data for the VS task were derived from manual annotations. We instructed the annotators to use pronouns instead of specific character names in all annotations. This guideline stemmed from the inherent constraints of most existing MLLMs, which generally lacked the capacity to process audio or subtitles. This made it difficult for

Annotation Guidelines for Topic Reasoning

1. Task Description: Your task is to formulate a question that pertains to the genre and key content of a given long video, and then provide the corresponding answer.

2. Question Requirements:

- Your questions should be centered around the core content of the video, rather than focusing on minor details.
- Suitable topics for questions include the genre of the video, the main events or themes, the primary environmental setting, the depicted weather conditions, and the time period or timeline.

3. Question Format:

- Questions should be structured in a multiple-choice format. Each question should have one correct answer and three plausible, yet incorrect, distractor options.

4. Question Examples (for reference only, not limited):

- What genre does this movie/video fall into?
- Where does the main scene in the video take place?
- What is the main event being narrated in the video?
- What is the protagonist in the video accomplishing?

Figure 1. Annotation Guidelines for the Topic Reasoning Task.

these models to identify specific characters. The annotation instructions and examples provided to the annotators are elaborated in Figure 2. More visualized examples of VS task can be found in Figure 10.

F.4. Needle Question-Answering (NQA).

We leveraged the GPT-4 [1] and the detailed video caption data from the WebVid dataset [5] to facilitate a semi-automated generation of annotated questions and answers for the NQA task. Initially, we selected video clips from WebVid, which we referred to as *needle* clips. The corresponding captions of these needle clips were then fed into GPT-4, which generated question-answer pairs based on the information encapsulated in the captions. The specific prompt provided to GPT-4 is depicted in Figure 3. The generated questions were carefully crafted to focus on a particular detail within the needle clip. These questions were structured to incorporate the maximum number of hints to effectively guide MLLMs in grounding the content of the needle within the context of the longer video. Following this, we randomly selected longer background videos from our ULVC and manually ensured that the scene indicated by the needle's question did not feature in these background videos. The final step involves integrating the needle into the longer video, thereby producing the final needle question video. More visualized examples of NQA task can be found in Figure 11.

F.5. Ego Reasoning (ER).

The video resources, questions, and correct responses used in the ER task were derived from the Natural Language Queries (NLQ) task within the Ego4D dataset [15]. This data was restructured to fit a multiple-choice question format.

F.6. Plot Question-Answering (PQA).

The PQA task's questions and answers were annotated by human annotators, following specific guidelines illustrated in Figure 4. We instructed the annotators to craft questions that probe into the intricate plot details encapsulated within the videos. These questions were designed to encompass both perception and reasoning aspects. We stipulated that both questions and their corresponding answers should avoid the use of specific character names or any objective hints, and should instead utilize pronouns. This approach was strategized to prevent potential information leakage, given that MLLMs often demonstrate a familiarity with the storylines of well-known movies and TV series. Such common-sense knowledge could potentially allow the MLLMs to answer questions correctly without the essential requirement of analyzing the input video.

Nonetheless, the complexity of character interactions and actions in longer videos poses a challenge to conveying plot details using only pronouns and feature descriptions. Previous datasets for plot question answering that avoided the use of character names often resulted in compromised

Annotation Guidelines for Video Summarization

1. Task Description: Your task is to provide a comprehensive summary of the key events occurring within a video clip that ranges from 3 to 15 minutes in length.

2. Annotation Requirements:

- The annotation should encapsulate the principal events portrayed in the video, structured in chronological order.
- Refrain from using specific character names in the annotation. Instead, all characters should be referred to using pronouns and identified by their unique attributes or roles, such as attire, age, profession, etc. For instance, characters could be described as an "elderly individual" (age), a "medical professional" (profession), among others.
- Disregard audio-related information, such as dialogues between characters. The summaries should be derived exclusively from the visual content presented in the video.

3. Annotation Template:

- Initiate your summary by outlining the overall content of the video: the event being narrated or the video's main theme.
- Subsequently, chronologically depict the key events that unfold in the video. The aim is to provide a clear and concise description of the main content, events, and scenes exhibited in the video.

4. Annotation Examples:

- **Cartoon:** This is a video about a cartoon sponge's whimsical adventures. The video begins with a cartoon sponge rushing into a house to converse with a cartoon starfish on a rocking chair. The sponge then heads to a concert hall where he watches a performance, during which a cartoon animal on a throne reprimands a cartoon octopus who continues his act. Later, the cartoon sponge and a cartoon squirrel are seen flying and conversing in the air. The sponge also encounters a cartoon shark preparing to drink coffee and a cartoon lobster sailing on a sponge, after which the lobster chases the sponge away.
- **Movie / TV Series:** This is a video depicting a dramatic narrative. The video starts with a man singing into a microphone, with a few other men playing instruments behind him. The scene changes to someone pushing open a door and walking into a room where others are resting. She then opens another door, enters a room and starts arguing with the singing man, which results in a fight. Next, the woman drives the man away, which results in a car crash. The car then falls off a bridge and gets hit by another car. The screen goes black and then lights up again, revealing a bookshelf filled with books at the end.
- **Documentary:** This is a documentary about forest animals and ecology. The video begins by showing scenes of fish, butterflies, orangutans, and birds in the forest. Then, the video depicts two birds cooperatively building a nest on a rock. As it starts to rain in the forest, a hatchling is born. The two birds catch bugs and frogs in the forest and feed them to the newborn. The camera follows the direction of the flowing river, which converges to form a spectacular waterfall. The video ends with a calm sea and beach, with a large flock of seabirds flying over the sea, hunting for prey close to the water.

Figure 2. Annotation Guidelines for the Video Summarization Task.

question diversity and tended towards generalized queries. We illustrate this through a comparative analysis of TVQA [21], Moviechat [41], and our PQA dataset's question word clouds in Figure 5. While TVQA provides a diverse range of questions, it does so by employing specific character names. In contrast, Moviechat avoids character names, but its questions are frequently overly broad, lack specific plot details, and exhibit diminished diversity. Our PQA dataset successfully navigates these challenges, offering a diverse

range of questions without resorting to the use of character names. More visualized examples of PQA task can be found in Figure 11.

F.7. Sub-Scene Captioning (SSC).

In the development process of the SSC task, we employed human annotators to generate both prompts and standard caption data. The specific guidelines provided to annotators are illustrated in Figure 6. Initially, the annotators identi-

Prompt for Generating Needle Questions

You are a question setter. Your task is to evaluate the participants' ability to capture detailed information from an extremely long video. The participants will receive a lengthy and content-rich video, and you are required to ask a question about a specific piece of information from the video.

I will provide you with a description of the segment that needs to be questioned at the end. Your question must include as much contextual information as possible to help the participants locate the source of the information. The description I provide generally contains multiple clues, and you should ask questions targeting different clues. Your question should be in a multiple-choice format, necessitating the provision of at least four choices, including the correct answer. Depending on the depth of information in the segment description, you can craft between 1 to 3 distinct questions.

Please provide the questions in the JSON format as follows...

Here is the description of the segment that needs to be questioned...

Figure 3. The prompt provided to GPT-4 in the process of creating the question-answer pair for the Needle Question-Answering task.

fied a specific, easily referable sub-scene within a lengthy movie. Subsequently, they crafted a prompt replete with adequate clues to reference this scene, ensuring the uniqueness of these clues throughout the entire film. To prevent any leakage of information, the prompt was designed to exclude any character-specific names or objective hints, instead incorporating rich descriptive details to allude to the plot. Following this, the annotators produced a detailed caption for this sub-scene, and deconstructed the caption into multiple, non-redundant "scoring points" to facilitate quantitative assessment (the details of the evaluation metric can be found in Appendix G.3). More visualized examples of PQA task can be found in Figure 12.

F.8. Action Order (AO).

The videos, questions, and answers for the action order task were all synthetically generated. In order to maintain the high quality of our evaluation data, we adopted a dual-strategy approach. Firstly, we selected actions for the *probe* videos that were not commonly seen in most films, such as making jewelry and water skiing. Secondly, in the selection of background videos, we conducted a cursory review of the video content to further ensure that the actions referenced in the questions were not present in the video. This rigorous methodology ensured the reliability of our data.

F.9. Action Count (AC).

The process of data acquisition and annotation for the action count task closely mirrored that of the action order task. All videos, questions, and answers were synthetically generated.

We employed a strategy consistent with the action order task to ensure the validity and reliability of our evaluation data.

G. Details of Baselines and the Evaluation Process

G.1. Baselines

In this section, we outline the primary baselines evaluated on our MLVU. For image-based MLLMs, most models lack multi-image inference capabilities. Therefore, we select Otter-I, LLaVA-1.6, and InternVL, which have official multi-image implementations. Additionally, we include two proprietary models—Claude-3-Opus and Qwen-VL-Max—that offer APIs for multi-image inference. For the available models, we determine the maximum input frames based on their LLM context length. Claude and Qwen support a maximum of approximately 20 images, so we choose 16 frames to ensure fair comparisons. Regarding video MLLMs, we use default frame sampling strategies. For example, VideoChat2 uniformly samples 16 frames, while LLaMA-Vid samples 1 frame per second. Specifically, GPT-4o can handle up to approximately 500 images at a resolution of 512×512 pixels. Thus, we select a sampling rate of 0.5 fps to accommodate most of our videos.

G.2. Inference Details

We have developed two templates specifically for Multiple-Choice and Generation tasks, as illustrated in Figure 7. Distinct system prompts were designed to accommodate the differences between video-based and image-based MLLMs. Considering the variances in task requirements, we incorporated “option prediction guidance” into the Multiple-Choice template to aid in option extraction. Conversely, in Generation tasks, we do not implement any additional interventions but employ fixed-question guidance to enable models to respond to diverse task questions. In our evaluation, the templates are seamlessly integrated into the evaluation code of open-release models or available API of proprietary models.

G.3. Evaluation Metrics

For the evaluation of Multiple Choice tasks, we directly compute absolute accuracy by matching the predicted option with the ground truth. In Generation tasks, we develop multiple criteria for assessment and employ GPT-4 to rank the alignment between generated texts and the provided answers. As illustrated in Figure 8, we use “Accuracy” and “Relevance” to benchmark Sub-scene Captioning, and “Completeness” and “Reliability” to evaluate the capabilities of Video Summary.

Annotation Guidelines for Plot Question-Answering

1. Task Description: Your task is to generate questions and answers based on the plot events depicted in various media, including movie, TV series, and cartoon animations.

2. Question Requirements:

- The questions should target specific details or events within the given video. Both factual and inferential questions are encouraged.
- Avoid using specific character names in the questions. Instead, use pronouns or identify characters by unique attributes or roles (e.g., attire, age, profession).
- Ensure that the plot referred to in your question is unique within the long video. Avoid using vague descriptions that can apply to multiple instances (like "eating"). Instead, refer to unique scenes or add enough details to specify the exact event.

3. Question Format:

- Questions should be structured in a multiple-choice format. Each question should have one correct answer and three plausible, yet incorrect, distractor options.

4. Examples of Questions (for reference only, not limited):

- How does the character in the small boat end up?
- How did the warship and the small boat approach each other?
- Why didn't the old man buy the chicken?
- What mode of transportation did the old man take in the end?
- What was the young woman doing when she drove to the airport?

Figure 4. Annotation Guidelines for the Plot Question-Answering Task.



Figure 5. **Word Cloud Comparison** of questions in TVQA test set (left), MovieChat test set (middle), and our PQA (right). Notably, TVQA's character-specific names require LLMs to recognize characters, risking reliance on pre-existing knowledge. In contrast, MovieChat questions are less diverse. Our PQA addresses these issues, providing enhanced usability and reliability.

H. Explorations of Video Retrieval Augmented Generation

As discussed in Section 4.3, most MLLMs are adversely affected by video length. Drawing inspiration from the use of Retrieval Augmented Generation (RAG) in video understanding, we have developed a zero-shot RAG strategy and seamlessly integrated it into existing MLLMs. Table 3 displays the performance comparison between the baseline models and the models employing our RAG strategy. It is noteworthy that all methods benefit from the RAG strategy in Needle QA, Ego Reasoning, and Plot QA. Conversely, minimal improvement is observed in Action Count, and a

decrease is noted in Action Order and Overall Reasoning. This is primarily because RAG facilitates the retrieval of detail-oriented video clips, which makes models more likely to focus on answer-related cues in specific single-detail reasoning tasks. However, RAG exhibits limited capabilities in multi-detail reasoning and holistic understanding tasks, which require global perception and knowledge aggregation.

The pipeline of our video retrieval augmented generation is illustrated in Figure 9. Initially, a long video is uniformly divided into N video clips, each containing C frames. Subsequently, we employ a robust video feature extraction tool, LanguageBind [62] to extract clip embeddings $F_I \in \mathbb{R}^{N \times d}$, where d represents the dimension of each clip embedding.

Annotation Guidelines for Sub-Scene Captioning

1. Task Description: You are required to provide a detailed **caption** for a specific scene in a long movie and clearly provide a unique **prompt** that can point to this scene.

2. Prompt Requirements:

- The clue in the prompt should direct to a specific and singular scene in the movie.
- Ensure that the prompt does not contain specific character names or movie-specific terms.
- The scene to be described should generally not exceed 1 minute.

3. Caption Requirements:

- Avoid using specific character names in the captions. Instead, use pronouns or identify characters by unique attributes or roles (e.g., attire, age, profession).
- Provide a caption and a list of unique plot details as scoring points, ensuring there's no repetition of details already present in the prompt.

4. Examples:

- Example (1):
 - Prompt: Please describe the situation after the man at the door takes off his hat and throws it away.
 - Caption: The hat flies into the room and is kicked into the large clock by the man in black who stands up.
 - Scoring points: "The hat flies into the room", "is kicked into the large clock", "by the man in black who stands up"
- Example (2):
 - Prompt: Please describe the reaction of the short-haired man when the long-haired man took out the urn.
 - Caption: The short-haired man stood up, held the urn in his hands, and pressed his forehead against the mouth of the urn, unable to hold back his tears.
 - Scoring points: "The short-haired man stood up", "held the urn in his hands", "pressed his forehead against the mouth of the urn", "unable to hold back his tears"

Figure 6. Annotation Guidelines for the Sub-Scene Captioning Task.

We then compute the similarities between F_I and the text embedding F_T , concatenating the top K clips to enhance the model's capability for question-answering. Given that many Video MLLMs are limited to processing only 16 frames, we have adjusted the settings for C and K to accommodate video retrieval in 16-second intervals. As discussed below, the RAG strategy excels in detail-oriented tasks but shows limitations in global understanding tasks. Moreover, it is relatively inefficient, requiring more than one minute to complete the process. Consequently, more effective approaches need to be developed for long video understanding tasks, and we aim to address this in future work.

I. More Visualized Examples of MLVU.

We present additional visualizations of our MLVU annotation examples in Figures 10, 11, and 12.

Multiple-choice

System Prompt:

(Video-MLLMs) Carefully watch this video and pay attention to every detail. Based on your observations, select the best option that accurately addresses the question.

(Image-MLLMs) These frames are from a video. Please examine each frame in the sequence provided to understand the narrative or activities depicted. Based on your observations, select the option that best answers the question.

Question Prompt:

[Question in each multi-choice task]

Only choose the best option. Best option: (

Evaluation:

Accuracy (0-100)



Generation

System Prompt:

(Video-MLLMs) Carefully watch this video and pay attention to every detail. Based on your observations, answer the given questions.

(Image-MLLMs) These frames are from a video. Please examine each frame in the sequence provided to understand the narrative or activities depicted. Based on your observations, answer the given questions.

Question Prompt:

(Video Sum.) Please summarize the main content of this video

(Sub-Scene Cap.) Please describe ...

Evaluation:

GPT Ranking (0-10)



Figure 7. Inference template for our MLVU.

Evaluation Prompt For Sub-Scene Captioning Task

##TASK DESCRIPTION: You are required to evaluate a respondent's answer based on a provided question, some scoring points, and the respondent's answer. You should provide two scores. The first is the accuracy score, which should range from 1 to 5. The second is the relevance score, which should also range from 1 to 5. Below are the criteria for each scoring category.

##ACCURACY Scoring Criteria:

Evaluate the respondent's answer against specific scoring points as follows:

Score 1: The response completely misses the scoring point.

Score 3: The response mentions content related to the scoring point but is not entirely correct.

Score 5: The response accurately addresses the scoring point.

Calculate the average score across all scoring points to determine the final accuracy score.

##RELEVANCE Scoring Criteria:

Assess how the respondent's answer relates to the original question:

Score 1: The response is completely off-topic from the question.

Score 2: The response is partially related to the question but contains a significant amount of irrelevant content.

Score 3: The response primarily addresses the question, but the respondent seems uncertain about their own answer.

Score 4: The response mostly addresses the question and the respondent appears confident in their answer.

Score 5: The response is fully focused on addressing the question with no irrelevant content and demonstrates complete certainty.

##INSTRUCTION:

1. Evaluate ACCURACY: First, assess and score each scoring point based on the respondent's answer. Calculate the average of these scores to establish the final accuracy score. Provide a detailed rationale before assigning your score.

2. Evaluate RELEVANCE: Assess the relevance of the respondent's answer to the question. Note that when evaluating relevance, the correctness of the answer is not considered; focus solely on how relevant the answer is to the question. Provide a comprehensive rationale before assigning your score.

3. Output Scores in JSON Format: Present the scores in JSON format as follows...

Evaluation Prompt For Video Summarization Task

##TASK DESCRIPTION:

You are required to evaluate the performance of the respondent in the video summarization task based on the standard answer and the respondent's answer. You should provide two scores. The first is the COMPLETENESS score, which should range from 1 to 5. The second is the RELIABILITY score, which should also range from 1 to 5. Below are the criteria for each scoring category:

##COMPLETENESS Scoring Criteria:

The completeness score focuses on whether the summary covers all key points and main information from the video.

Score 1: The summary hardly covers any of the main content or key points of the video.

Score 2: The summary covers some of the main content and key points but misses many.

Score 3: The summary covers most of the main content and key points.

Score 4: The summary is very comprehensive, covering most to nearly all of the main content and key points.

Score 5: The summary completely covers all the main content and key points of the video.

##CORRECTNESS Scoring Criteria:

The correctness score evaluates the correctness and clarity of the video summary. It checks for factual errors, misleading statements, and contradictions with the video content. If the respondent's answer includes details that are not present in the standard answer, as long as these details do not conflict with the correct answer and are reasonable, points should not be deducted.

Score 1: Contains multiple factual errors and contradictions; presentation is confusing.

Score 2: Includes several errors and some contradictions; needs clearer presentation.

Score 3: Generally accurate with minor errors; minimal contradictions; reasonably clear presentation.

Score 4: Very accurate with negligible inaccuracies; no contradictions; clear and fluent presentation.

Score 5: Completely accurate with no errors or contradictions; presentation is clear and easy to understand.

##INSTRUCTION:

1. Evaluate COMPLETENESS: First, analyze the respondent's answer according to the scoring criteria, then provide an integer score between 1 and 5 based on sufficient evidence.

2. Evaluate CORRECTNESS : First, analyze the respondent's answer according to the scoring criteria, then provide an integer score between 1 and 5 based on sufficient evidence.

3. Output Scores in JSON Format: Present the scores in JSON format as follows...

Figure 8. Detailed prompt for evaluation of generation tasks in MLVU.

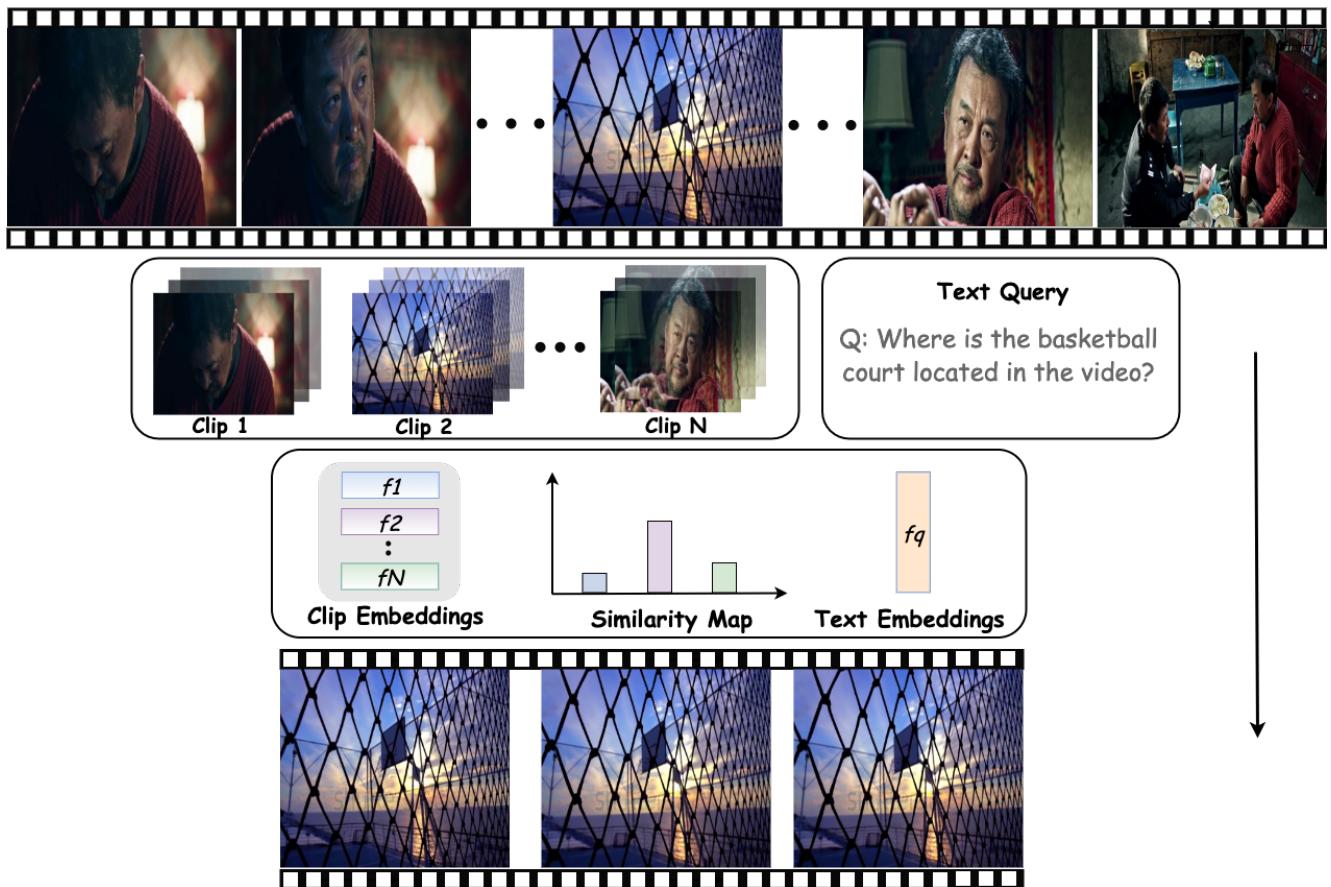


Figure 9. Pipeline of our video retrieval augmented generation strategy.

Model	Settings	Needle QA	Ego Rea.	Plot QA	Action Or.	Action Co.	Anomaly Rec.	Topic Rea.
LLaVA-B	-	43.1	38.4	41.0	25.5	25.7	41.0	60.6
	C=2,K=8	50.7	45.7	49.7	26.3	26.7	40.8	59.8
	C=4,K=4	53.5	43.5	50.6	25.9	29.6	39.9	58.5
LLaVA-R	C=8,K=2	55.2	42.6	50.3	25.1	30.1	40.6	59.5
	-	52.7	43.5	54.4	32.8	23.8	67.0	78.8
	C=2,K=8	77.2	52.6	61.4	30.1	36.4	57.9	69.2
InternVL-B	C=4,K=4	76.3	51.4	59.9	29.3	36.9	58.3	69.4
	C=8,K=2	77.8	48.9	61.6	31.7	33.0	60.2	62.3
	-	39.4	33.5	35.4	18.5	25.7	41.5	54.5
InternVL-R	C=2,K=8	61.4	42.6	38.8	17.4	17.5	35.7	48.5
	C=4,K=4	58.9	42.6	39.1	17.8	23.8	36.0	49.3
	C=8,K=2	62.0	38.4	36.2	25.5	18.0	38.5	51.0
Video-LLaMA-B	-	42.0	47.4	43.8	22.8	29.6	51.5	74.6
	C=2,K=8	72.1	53.7	55.5	21.6	30.1	45.8	68.2
	C=4,K=4	72.4	55.4	53.4	22.4	31.1	45.3	68.9
Video-LLaMA-R	C=8,K=2	73.8	53.1	55.3	22.0	31.6	46.6	69.7
	-	49.0	48.6	44.5	23.2	23.0	52.5	70.9
	C=2,K=8	60.6	44.3	47.4	23.2	23.7	42.8	60.9
MiniGPT4-Video-B	C=4,K=4	60.3	44.6	46.9	26.3	23.8	42.6	60.7
	C=8,K=2	56.3	44.6	46.6	27.4	24.8	45.0	47.5
	-	49.0	48.6	44.5	23.2	23.0	52.5	70.9

Table 3. Quantitative results on video Retrieval Augmented Generation. “model-B” and “model-R” denote Baseline and RAG models respectively. We evaluate two image MLLMs and three video MLLMs in different settings.

Topic Reasoning



Question: What type of film is this?

- (A) Mystery (B) Action (C) Comedy **(D) Romance**



Question: What is this video about?

- (A) A person in the game taking care of pets **(B) A person in the game building a structure by the lake**
 (C) A person in the game planting trees by the lake (D) A documentary about humans and nature



Question: Where is the main setting of the video?

- (A) Desert (B) Grassland **(C) Outside the house** (D) Inside the house

Video Summarization



Prompt: Please summarize the main content of this video.

The video begins with two men talking in a dimly lit room. After one of the men leaves, he enters another house where an elderly woman is present. They engage in conversation, and the elderly woman appears sad. In another scene, two women are talking, and one of them takes car keys and leaves. She arrives at another location and talks with a woman and a man. Subsequently, one of the women makes a phone call.



Prompt: Please summarize the main content of this video.

The video starts with a man singing into a microphone, with a few other men playing instruments behind him. The scene changes to someone pushing open a door and walking into a room where others are resting. She then opens another door, enters a room and starts arguing with the singing man, which results in a fight. Next, the woman drives the man away, which results in a car crash. The car then falls off a bridge and gets hit by another car. The screen goes black and then lights up again, revealing a bookshelf filled with books at the end.

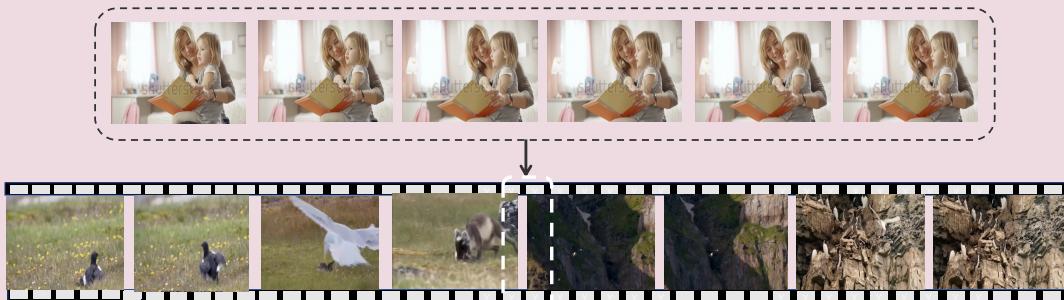
Figure 10. More Examples of Topic Reasoning and Video Summarization Tasks.

Needle Question-Answering



Question: Where is the senior businessman having a serious conversation on the cell phone?

- (A) In a park (B) By the sea shore (C) In his office (D) At a restaurant



Question: What are the little girl and her grandmother doing together?

- (A) Watching TV (B) Playing a game **(C) Reading a children's book** (D) Eating dinner

Plot Question Answering



Question: What happened after the person with the yellow stripe arrived at the camp?



Question: What color is the table lamp in the background of the scene where a man and a women are chatting?

- (A) Black (B) White (C) Green (D) Yellow

Figure 11. More Examples of Needle Question Answering and Plot Question Answering Tasks.

Sub-Scene Captioning



Prompt: Please describe the situation after the woman in red walked to the window of the bridal shop.

Answer: The woman in red took a picture with her camera. As the photo slowly slid out, she looked down at it.



Prompt: Please describe the process of a man alone in a room looking for a camera.

The man raises his cue stick to find the angle, then turns around and walks to a statue where he finds the camera.

Figure 12. More Examples of Sub-Scene Captioning.