

Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis

Chaoyou Fu^{1,2,♦}, Yuhang Dai³, Yongdong Luo⁴, Lei Li⁵, Shuhuai Ren⁶
 Renrui Zhang⁷, Zihan Wang⁸, Chenyu Zhou⁴, Yunhang Shen⁴, Mengdan Zhang⁹
 Peixian Chen⁴, Yanwei Li⁷, Shaohui Lin⁸, Sirui Zhao³, Ke Li⁴, Tong Xu³
 Xiawu Zheng⁴, Enhong Chen³, Caifeng Shan^{1,2,*}, Ran He^{9,*}, Xing Sun^{5,*}

¹State Key Laboratory for Novel Software Technology, Nanjing University

²School of Intelligence Science and Technology, Nanjing University

³State Key Laboratory of Cognitive Intelligence

⁴XMU, ⁵HKU, ⁶PKU, ⁷CUHK, ⁸ECNU, ⁹CASIA

Abstract

In the quest for artificial general intelligence, Multi-modal Large Language Models (MLLMs) have emerged as a focal point in recent advancements. However, the predominant focus remains on developing their capabilities in static image understanding. The potential of MLLMs to process sequential visual data is still insufficiently explored, highlighting the lack of a comprehensive, high-quality assessment of their performance. In this paper, we introduce **Video-MME**, the first-ever full-spectrum, Multi-Modal Evaluation benchmark of MLLMs in Video analysis. Our work distinguishes from existing benchmarks through four key features: 1) **Diversity in video types**, spanning 6 primary visual domains with 30 subfields to ensure broad scenario generalizability; 2) **Duration in temporal dimension**, encompassing both short-, medium-, and long-term videos, ranging from 11 seconds to 1 hour, for robust contextual dynamics; 3) **Breadth in data modalities**, integrating multi-modal inputs besides video frames, including subtitles and audios, to unveil the all-round capabilities of MLLMs; 4) **Quality in annotations**, utilizing rigorous manual labeling by expert annotators to facilitate precise and reliable model assessment. With Video-MME, we extensively evaluate various state-of-the-art MLLMs, and reveal that Gemini 1.5 Pro is the best-performing commercial model, significantly outperforming the open-source models with an average accuracy of 75%, compared to 71.9% for GPT-4o. The results also demonstrate that Video-MME is a universal benchmark that applies to both image and video MLLMs. Further analysis indicates that subtitle and audio

information could significantly enhance video understanding. Besides, a decline in MLLM performance is observed as video duration increases for all models. Our dataset along with these findings underscores the need for further improvements in handling longer sequences and multi-modal data, shedding light on future MLLM development. Project page: <https://video-mme.github.io>.

1. Introduction

The rapid development of MLLMs in recent years [66] has highlighted their impressive perception and cognitive capabilities across various multimodal benchmarks [16, 43, 67]. These advancements show the great potential of MLLMs to serve as a foundation that can digest the multimodal real world and pave the way toward artificial general intelligence. However, current MLLMs and their evaluation primarily focus on static visual data understanding, which fails to capture the dynamic nature of the real world involving complex interactions between objects over time. To approximate real-world scenarios more accurately, it is crucial to explore and assess the capabilities of MLLMs on sequential visual data, such as videos. Many early efforts [18, 33, 55] have been made to inspire the video understanding potentials of MLLMs with promising results. However, existing video-based benchmarks [27, 32, 41, 45] are still limited to thoroughly reveal their performance, such as a lack of diversity in video types, insufficient coverage of temporal dynamics, and the narrow focus on a single modality. These inevitably hinder the all-around evaluation of MLLMs.

To this end, we introduce **Video-MME**, the first-ever comprehensive Multi-Modal Evaluation benchmark crafted

*Corresponding Author. ♦ Project Leader.

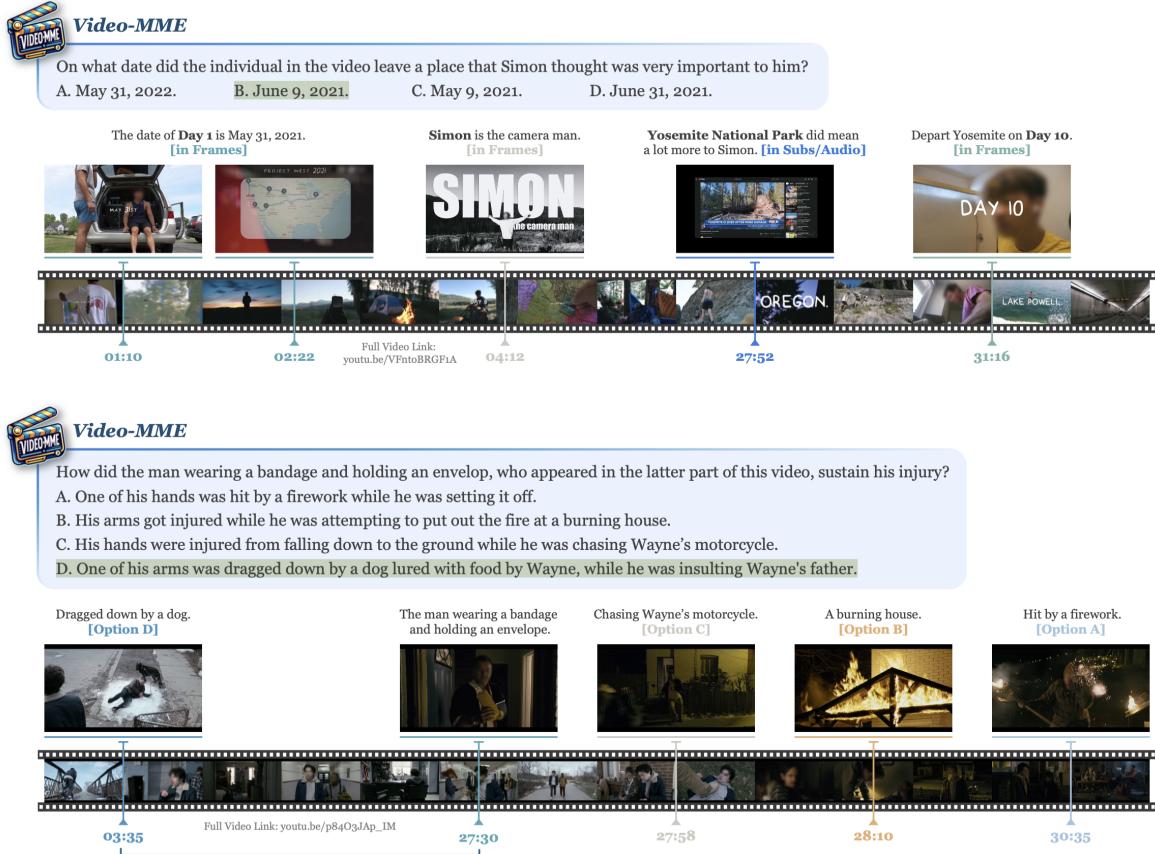


Figure 1. We introduce Video-MME to provide high-quality assessment of MLLMs’ performance, where all the videos and annotations are **manually** collected and curated. These are two highlighted examples of Video-MME with the ground-truth answers in green, which require the model to perform long-term contextual understanding and complex spatial and temporal reasoning. Video-MME is meticulously designed to pose significant challenges, thereby effectively evaluating the compositional video understanding capabilities of MLLMs.

for MLLMs in **Video** analysis. As exemplified in Figure 1, we meticulously curate a dataset of 900 videos across various scenarios, and annotate a set of 2,700 high-quality multiple-choice questions (3 per video) to foster a robust evaluation. As presented in Figure 2, for generalizability, our dataset widely spans 6 visual domains, including Knowledge, Film & Television, Sports Competition, Artistic Performance, Life Record, and Multilingual, with 30 fine-grained categories, e.g., technology, documentary, news report, esports, magic show, and fashion. Importantly, the videos vary significantly in length, ranging from 11 seconds to 1 hour, specifically evaluating the adaptability of MLLMs across varying temporal contexts. Furthermore, Video-MME enriches the assessment by incorporating the associated subtitles and audio tracks, thereby enhancing the analysis of multimodal inputs for video understanding.

Using Video-MME, we benchmark various state-of-the-art MLLMs, including GPT-4V [48], GPT-4o [49], and Gemini 1.5 Pro [55], alongside open-source image models like InternVL-Chat-V1.5 [11] and video models like

LLaVA-NeXT-Video [73]. Our experiments in Table 4 indicate that Gemini 1.5 Pro is the highest-performing commercial model, achieving an average accuracy of 75%. In comparison, open-source MLLMs exhibit substantial gaps compared to commercial models. For instance, the leading open-source model, VILA-1.5 [34], attains an overall accuracy of 59%. These findings suggest there is considerable room for improvement in the open-source community. Our benchmark is also available to advanced image-based models by extending their input to multi-frame images, e.g., Qwen-VL-Max [5] and InternVL-Chat-V1.5 [11]. The accuracies of both the models reach 50%, which is close to that of the video-specific model LLaVA-NeXT-Video, indicating that image understanding is the basis of video understanding and the wide applicability of Video-MME in the field of MLLMs. Further observations in Table 5 indicate that integrating subtitles and audios significantly enhances video comprehension capabilities, e.g., boosting Gemini 1.5 Pro by 6.2% and 4.3% respectively, with the gains being more pronounced for longer videos. A fine-grained analy-

Benchmarks	#Videos	#Clips	Len.(s)	#QA Pairs	Anno.	QA Tokens	Sub. Tokens	Multi-level	Open-domain	Sub.&Aud.
MSRVTT-QA [64]	2,990	2,990	15.2	72,821	A	8.4	X	X	✓	X
MSVD-QA [64]	504	504	9.8	13,157	A	7.6	X	X	✓	X
TGIF-QA [21]	9,575	9,575	3.0	8,506	A&M	20.5	X	X	✓	X
ActivityNet-QA [68]	800	800	111.4	8,000	M	10.2	X	X	X	X
TVQA [24]	2,179	15,253	11.2	15,253	M	27.8	159.8	X	X	X
How2QA [28]	1,166	2,852	15.3	2,852	M	16.9	31.1	X	✓	X
STAR [61]	914	7,098	11.9	7,098	A	19.5	X	X	✓	X
NExT-QA [63]	1,000	1,000	39.5	8,564	A	25.3	X	X	✓	X
MVBench [27]	3,641	3,641	16.0	4,000	A	27.3	X	X	✓	X
Video-Bench [47]	5,917	5,917	56.0	17,036	A&M	21.3	X	X	✓	X
EgoSchema [45]	5,063	5,063	180.0	5,063	A&M	126.8	X	X	X	X
AutoEval-Video [10]	327	327	14.6	327	M	11.9	X	X	✓	X
TempCompass [41]	410	500	11.4	7,540	A&M	49.2	X	X	✓	X
Video-MME-S	300	300	80.7	900		28.7	198.6			
Video-MME-M	300	300	515.9	900	M	32.8	1425.6	✓	✓	✓
Video-MME-L	300	300	2466.7	900		45.6	6515.6			
Video-MME	900	900	1017.9	2,700		35.7	3086.5			

Table 1. **The comparison of various benchmarks** encompasses several key aspects: the total number of videos (**#Videos**), the number of clips (**#Clips**), the average duration of the videos (**Len.**), the number of QA pairs (**#QA Pairs**), the method of annotation (**Anno.**, M/A means the manually/automatic manner), the average number of QA pair tokens (**QA Tokens**), the average number of subtitle tokens (**Sub. Tokens**), whether the videos cover multiple duration levels (**Multi-level**), whether the videos are sourced from a broad range of open domains (**Open-domain**), and whether provide subtitle together with audio information (**Sub.&Aud.**). Video-MME-S/M/L denotes the short/medium/long part. It is important to note our comparison focuses solely on the multiple-choice subset of the datasets.

sis of task types reveals that subtitles and audios are particularly beneficial for videos requiring substantial domain knowledge. We also note a general decline in MLLM performance with increasing video length. This trend suggests that limitations in processing longer video sequences could be a critical bottleneck in the performance of MLLMs.

Finally, we discuss promising avenues for improving the capabilities of MLLMs in processing video content. Potential directions include architectural development for better handling long context inputs and constructing training data focused on complex temporal reasoning scenarios. We expect that our benchmarking, evaluation findings, detailed analysis, and outlined insights will inspire future progress toward more capable and robust MLLMs.

2. Related Work

Advancements in MLLMs. Recent advancements in MLLMs have seen notable progress [7, 19, 22, 59, 66]. MLLMs typically comprise three core modules: (i) a vision encoder for visual feature extraction, (ii) a modality alignment module to integrate visual features into the embedding space of the language model, and (iii) an LLM backbone for decoding multi-modal context. CLIP [51] and SigLIP [70] are widely-used for image encoding, while LLaMA [56] and Vicuna [12] serve as popular choices for LLMs. The alignment module varies from simple linear projections [35] to more complex architectures such as Q-Former [13, 25], and gated cross-attention layers substantiated by Flamingo [1, 4]. Additionally, Fuyu-8B [6] introduces a novel framework mapping raw image pixels directly to the LLM embedding space. Regarding MLLMs for processing videos [26, 44, 54], the key difference lies in how

they encode the video into vision tokens compatible with the LLMs. Representative work like Video-LLaMA [71] first uses a ViT [14] with an image Q-Former to encode individual frames and then employs a video Q-Former for temporal modeling. VideoChat2 [27] utilizes a video transformer to encode video features and subsequently implements a Q-Former [25] to compress video tokens. To empower video MLLMs with temporal localization capability [20, 50, 60], TimeChat [53] constructs time-sensitive instruction tuning datasets and encodes timestamp knowledge into visual tokens. However, the potential of MLLMs in processing sequential visual data is still under-explored.

MLLM Benchmarks. Alongside advancements in architecture, significant efforts have been made to improve benchmarking for MLLMs, guiding the development of the next generation of these models. Previous studies have integrated various aspects of evaluation, such as perception and cognitive capabilities, to create comprehensive benchmarks for assessing image MLLMs [16, 40, 67]. As image MLLMs have demonstrated exceptional performance in general perception tasks, benchmarks regarding scientific understanding [31], multi-modal mathematical reasoning [43, 72], and multi-disciplinary [69] capabilities have drawn increasing attention. For video MLLMs, similar efforts have been made to incorporate existing benchmarks [29, 58] for evaluating video understanding [27, 47]. Due to the inherently temporal characteristics of video modalities, specialized benchmarks have been created to evaluate temporal comprehension, underscoring the current limitations of video MLLMs in understanding video content [3, 15, 32, 41, 57, 62].

3. Video-MME

3.1. Dataset Construction

The construction of the Video-MME dataset involves three main steps: video collection, question-answer annotation, and quality review. The details are as follows.

Video Collection. To ensure comprehensive coverage across diverse video types, we first establish a domain hierarchy for sourcing raw videos from YouTube. We identify 6 key domains: Knowledge, Film & Television, Sports Competition, Life Record, and Multilingual, informed by popular trends on YouTube. Each domain is further subdivided into specific tags, such as football and basketball under Sports Competition, yielding a total of 30 finely-grained video categories. The complete domain-tag hierarchy and its distribution are illustrated on the left side of Figure 2. For each category, we collect videos with varying duration lengths, including short (< 2 minutes), medium (4-15 minutes), and long videos (30-60 minutes). Besides, we also obtain corresponding meta-information such as subtitles (if provided) and audios for in-depth investigation. The resulting dataset consists of 900 videos, including 744 videos with subtitles and all 900 with audios, representing a range of domains with a relatively balanced distribution of video durations, as depicted on the right side of Figure 2.

Question-Answer Annotation. Following the collection of raw video data, we annotate the videos with high-quality question-answer (QA) pairs to rigorously assess the proficiency of MLLMs in video content interpretation. We employ a multiple-choice QA format to facilitate a straightforward and flexible assessment and recruit annotators with strong English proficiency and extensive research experience in vision-language learning. Specifically, each annotator is tasked with thoroughly viewing the entire video, then iteratively creating three relevant questions, each accompanied by four candidate options, contributing to a total of 2,700 QA pairs. As shown in the bottom right corner of Figure 2, these questions encompass 12 distinct task types, including perception, reasoning, and information synthesis. Each QA pair is designed to be closely tied to the video content, preventing MLLMs from answering accurately without directly referencing the video.

Quality Review. To guarantee the quality of our dataset, we conduct a rigorous manual review process. First, each QA pair undergoes examination by a different annotator to verify that (i) the language expression is clear and precise; (ii) the question is answerable with logically consistent candidate options, and the designated correct answer is appropriate. Furthermore, to ensure that the questions are sufficiently challenging enough and require video content as a

critical clue to answer [72], we provide the text-only questions to Gemini 1.5 Pro and filter out QA pairs that can be answered solely based on the textual questions. For instance, a question like “What is the biggest achievement of Argentina’s number 10 in 2022?” which can be inferred as the World Cup victory, would be excluded in this phase. Questions that do not meet this criterion are returned to the original annotator for revision. Statistical analysis shows that Gemini 1.5 Pro achieves less than 15% accuracy in the text-only setup, underscoring the robustness of the video content-based requirement. Through our dataset construction process, we strive to deliver a high-quality, diverse, and well-balanced dataset that will serve as a valuable resource for advancing research in multimodal understanding.

3.2. Dataset Statistics

In this section, we present the detailed statistics of our dataset to provide a more comprehensive understanding, including the meta information, QA pairs, certificate lengths, qualitative analysis, and comparison to previous works.

Video & Meta Information. Our dataset comprises a total of 900 videos, 744 subtitles, and 900 audio files. Most videos are accompanied by both subtitles and audios, providing valuable resources for investigating the impact of external information on video understanding performance. The upper right part of Figure 2 illustrates the duration distribution of the collected videos. Specifically, within the short video category, longer videos occupy a larger portion. For medium-length videos, the duration distribution is more uniform, while long videos exhibit a long-tailed distribution, with fewer samples as the duration increases. The bottom right part of Figure 2 shows the distribution of task types. Shorter videos predominantly involve perception-related tasks such as action and object recognition, while longer videos mainly feature tasks related to reasoning. Overall, this analysis highlights that Video-MME covers a wide range of video durations and task types, enabling a comprehensive evaluation of temporal understanding.

QA Pairs. We present a detailed analysis of language diversity in the questions and answers within our dataset. Table 2 lists the average word counts of the textual fields including questions, options and answers in our dataset. Notably, the word counts display notable consistency across different video lengths, suggesting a standardized format in the QA pairs. In contrast, subtitle word counts increase markedly with video length—e.g., short videos average 198.6 words, while long videos reach up to 6.5K words. This trend indicates longer videos contain more information, as evidenced by the increased volume of subtitles. The analysis reveals that our questions are diverse, and the answers are well-balanced. In addition, the distribution of

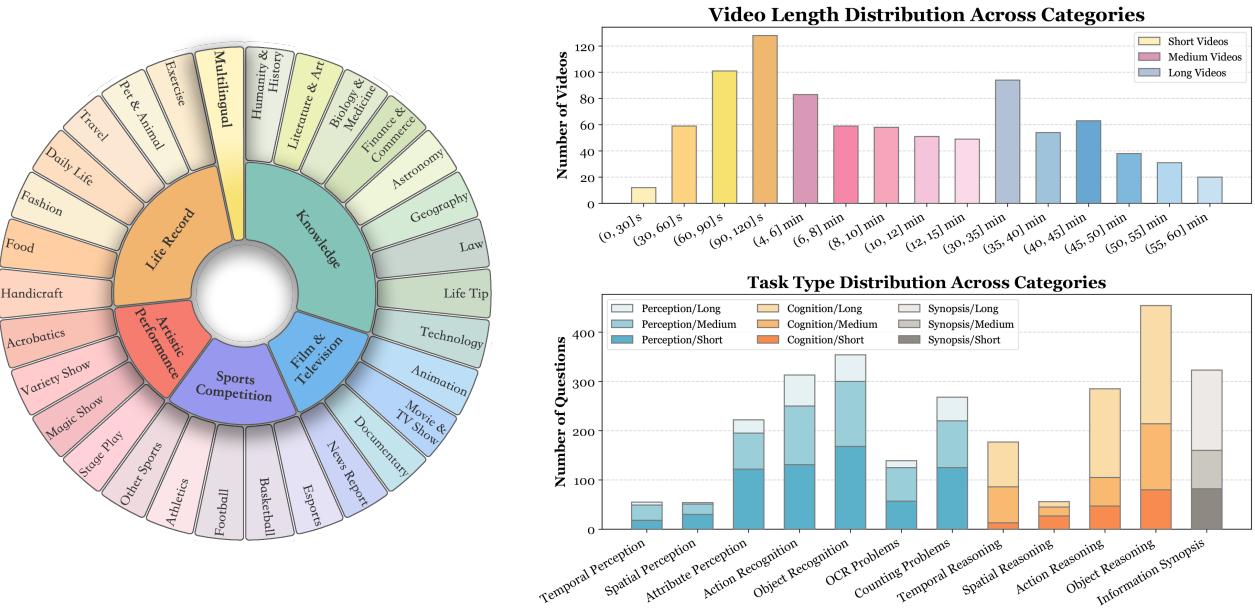


Figure 2. Statistics analysis of Video-MME. (left) Video categories. Our benchmark covers 6 key domains and 30 sub-class video types. (right) Video duration and question type distributions. Video-MME has a full spectrum of video length and covers different core abilities of MLLMs, enabling a comprehensive evaluation of temporal understanding.

Dataset	Question	Options	Answer	Subtitles
Short	11.5	17.2	4.0	198.6
Medium	12.2	20.6	5.0	1425.6
Long	14.5	31.0	7.5	6515.6
All	12.7	22.9	5.5	3086.5

Table 2. Average word counts of different textual fields in Video-MME. The word counts for questions, options, and answers are consistent across video lengths, indicating a standardized QA format and an increasing trend of subtitles information volume.

Video	Avg. V.L.	Med. C.L.	Avg. C.L.
EgoSchema	180	~ 100	-
Short	82.5	26.0	28.8
Medium	562.7	164.7	160.0
Long	2385.5	890.7	967.7

Table 3. Analysis of certificate length in seconds. Avg. V.L.: average video length, Med. C.L.: median certificate length, Avg. C.L.: average certificate length.

the options (A/B/C/D) follows a near-uniform distribution (25.1/27.2/25.3/22.4%), ensuring an unbiased evaluation.

Certificate Length Analysis. Inspired by the recent work EgoSchema [45], we adopt the *certificate length* to analyze the temporal difficulty of the QA pairs. The certificate of

a given video QA pair is defined as the minimum set of sub-clips of the video that are both necessary and sufficient to convince a human verifier that the marked annotation is correct. The certificate length is calculated as the sum of the temporal lengths of the sub-clips identified. We randomly sample 3 videos from each class and calculate the certificate length distribution. As shown in Table 3, our dataset yields a median certificate length of 26s, 164.7s, and 890.7s for short, medium and long videos, respectively. Compared with the certificate length of EgoSchema, our medium and long video subset requires much longer video content digestion to answer the question. To the best of our knowledge, this analysis makes our Video-MME the most challenging Video QA dataset to date.

Qualitative Analysis. Building on our previous analysis, we have established that our proposed benchmark, Video-MME, is both diverse and challenging, making it an exemplary testbed for MLLMs. Figure 1 showcases specific cases from our Video-MME dataset to illustrate this.

In the first example, the model must integrate information from various sources: visual data from video frames (e.g., “Day 1 is May 31, 2021”) and auditory/subtitle content (e.g., referring to “Yosemite National Park”). Moreover, the model is required to perform simple arithmetic operations to determine the exact departure date. This multi-modal and multi-step reasoning highlights the complexity

Models	LLM Params	Short (%)		Medium (%)		Long (%)		Overall (%)	
		w/o subs	w/ subs						
<i>Open & Closed-source Image MLLMs</i>									
Qwen-VL-Chat [5]	7B	46.9	47.3	38.7	40.4	37.8	37.9	41.1	41.9
Qwen-VL-Max [5]	-	55.8	57.6	49.2	48.9	48.9	47.0	51.3	51.2
InternVL-Chat-V1.5 [11]	20B	60.2	61.7	46.4	49.1	45.6	46.6	50.7	52.4
<i>Open-source Video MLLMs</i>									
Video-LLaVA [33]	7B	45.3	46.1	38.0	40.7	36.2	38.1	39.9	41.6
ST-LLM [38]	7B	45.7	48.4	36.8	41.4	31.3	36.9	37.9	42.3
ShareGPT4Video [8]	8B	48.3	53.6	36.3	39.3	35.0	37.9	39.9	43.6
VideoChat2-Mistral [27]	7B	48.3	52.8	37.0	39.4	33.2	39.2	39.5	43.8
Chat-UniVi-V1.5 [23]	7B	45.7	51.2	40.3	44.6	35.8	41.8	40.6	45.9
VITA-1.5 [18]	7B	67.0	69.9	54.2	55.7	47.1	50.4	56.1	58.7
VITA-1.0 [17]	8×7B	65.9	70.4	52.9	56.2	48.6	50.9	55.8	59.2
LLaVA-NeXT-Video [73]	34B	61.7	65.1	50.1	52.2	44.3	47.2	52.0	54.9
VILA-1.5 [34]	34B	68.1	68.9	58.1	57.4	50.8	52.0	59.0	59.4
<i>Closed-source MLLMs</i>									
GPT-4V [48]	-	70.5	73.2	55.8	59.7	53.5	56.9	59.9	63.3
GPT-4o [49]	-	80.0	82.8	70.3	76.6	65.3	72.1	71.9	77.2
Gemini 1.5 Flash [55]	-	78.8	79.8	68.8	74.7	61.1	68.8	70.3	75.0
Gemini 1.5 Pro [55]	-	81.7	84.5	74.3	81.0	67.4	77.4	75.0	81.3

Table 4. **Performance of MLLMs on Video-MME.** The results are evaluated on short, medium, and long durations, under the settings of “without subtitles” and “with subtitles”. In the “with subtitles” setting, subtitles are selected to correspond to the sampled video frames.

and high quality of our dataset. The second example involves a question placed towards the end of a video, with the provided answer options dispersed across different segments of the video. This necessitates a comprehensive understanding of the entire content, which can be as long as 30 minutes. These highlighted cases underscore that our Video-MME dataset is meticulously designed to pose significant challenges, thereby effectively evaluating the compositional video understanding capabilities of MLLMs.

4. Experiments

In this section, we evaluate a variety of MLLMs on our Video-MME benchmark. We begin by outlining the evaluation settings, followed by the quantitative results for both open-source and closed-source models. Finally, we present case studies to provide an intuitive understanding, and investigate the effect of the modality and duration.

4.1. Settings

We conduct the evaluation on 4 commercial models, i.e., GPT-4V [48], GPT-4o [49], Gemini 1.5 Flash [55], and Gemini 1.5 Pro [55]. Representative open-source video MLLMs including Video-LLaVA [33], VideoChat2-Mistral [27], ST-LLM [38], ShareGPT4Video [8], Chat-UniVi-V1.5 [23], LLaVA-NeXT-Video [73], VITA-1.0 [17], VITA-1.5 [18], and VILA-1.5 [34] are evaluated as well. In addition, we also include advanced image MLLMs, i.e., Qwen-VL-Chat/Max [5] and InternVL-Chat-

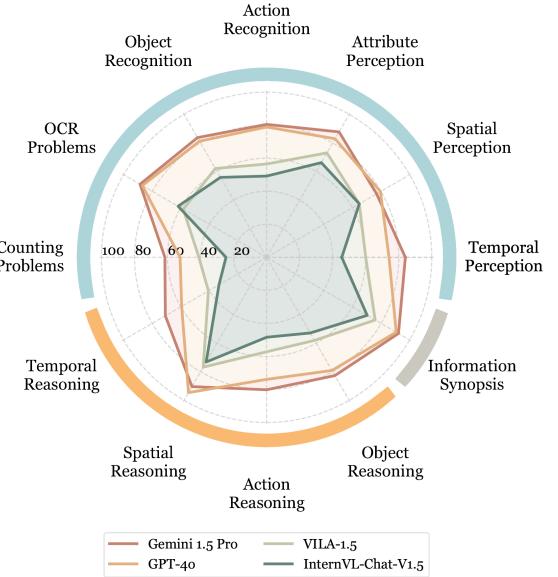


Figure 3. **Comparison between four selected MLLMs on Video-MME across different tasks.** The light blue in the outer circle stands for perception tasks and the orange represents reasoning tasks. As can be observed, counting problems are a joint bottleneck of current multimodal models in video understanding.

V1.5 [11], which usually can generalize to multi-image scenarios. We follow their official configurations and try to use more frames for evaluation. The accuracy is computed by directly comparing the model’s output with

Subset	Modality	Category						Overall
		Knowledge	Film & Television	Sports Competition	Artistic Performance	Life Record	Multilingual	
Short	Frames	78.3	80.8	76.7	86.7	88.1	76.7	81.7
	+ Subs	83.3 (+4.9)	86.7 (+5.8)	79.3 (+2.7)	87.6 (+1.0)	86.6 (-1.5)	86.7 (+10.0)	84.5 (+2.8)
	+ Audio	81.4 (+3.1)	87.5 (+6.7)	78.7 (+2.0)	86.7 (-)	85.6 (-2.4)	86.7 (+10.0)	83.6 (+1.9)
Medium	Frames	70.2	81.2	68.7	84.3	73.4	87.5	74.3
	+ Subs	83.3 (+13.1)	84.9 (+3.7)	76.2 (+7.5)	85.3 (+1.0)	76.8 (+3.4)	83.3 (-4.2)	81.0 (+6.7)
	+ Audio	80.2 (+9.9)	83.9 (+2.6)	72.1 (+3.4)	84.3 (-)	76.8 (+3.4)	100.0 (+12.5)	79.5 (+5.2)
Long	Frames	73.4	70.1	58.3	63.3	65.1	70.8	67.4
	+ Subs	83.0 (+9.6)	71.4 (+1.3)	77.5 (+19.2)	70.0 (+6.7)	74.6 (+9.5)	87.5 (+16.7)	77.4 (+10.1)
	+ Audio	81.1 (+7.7)	73.2 (+3.1)	72.6 (+14.3)	63.3 (-)	66.7 (+1.6)	83.3 (+12.5)	73.6 (+6.2)
Overall	Frames	74.1	77.9	68.6	78.8	77.4	78.2	75.0
	+ Subs	83.2 (+9.2)	81.8 (+3.9)	77.7 (+9.1)	81.5 (+2.7)	80.3 (+2.9)	85.9 (+7.7)	81.3 (+6.2)
	+ Audio	80.9 (+6.8)	82.4 (+4.5)	74.6 (+6.1)	78.8 (-)	78.0 (+0.6)	89.7 (+11.5)	79.4 (+4.3)

Table 5. **Performance of Gemini 1.5 Pro across six major categories in the Video-MME benchmark.** Evaluation includes three input modality settings: frames only, frames with subtitles, and frames with audio, highlighting the impact of multimodal inputs on model performance within each category. As one can observe, subtitles and audios contribute essential information for understanding the video.

the ground-truth answer, without utilizing any external models, such as ChatGPT.

4.2. Quantitative Results

Performance of Commercial Models. As one of the pioneering commercial large models integrated with video comprehension capabilities, Gemini 1.5 Pro has achieved the best performance among its peers on Video-MME. As depicted in Table 4, with video frames as input alone, Gemini 1.5 Pro attains an accuracy of 75%, surpassing GPT-4V and GPT-4o by 15.1% and 3.1%, respectively. Table 5 shows the fine-grained performance of Gemini 1.5 Pro. Among the 6 major video categories, Gemini 1.5 Pro performs the best in Artistic Performance while performing the lowest in the Sports Competition category. As video duration increases, Gemini 1.5 Pro’s performance declines (e.g., –14.3% from short to long videos), highlighting the model’s weakness in capturing long-range temporal relationships. Nevertheless, Gemini 1.5 Pro’s performance on long videos still surpasses almost all open-source models, except for VILA-1.5, on short videos, demonstrating its superior capabilities. In addition to visual frame input, Gemini 1.5 Pro’s support for additional modalities, including subtitles and audios, provides opportunities for further performance improvement. For example, Table 5 displays that using audios can increase accuracy by 6.2% for long videos, and the improvement in the multilingual category even reaches 16.7%. We can also see that the effect of subtitles and audios is different in these six categories. These motivate future research to develop versatile models that can support a wider range of modality inputs.

Performance of Open-sourced Models. As shown in Table 4, among the 7B models, VITA-1.5 achieves the best performance with 56.1%. VILA-1.5 with 34B LLM

achieves an accuracy of 59%, demonstrating its stronger capabilities, especially in the tasks of spatial reasoning, attribute perception, and information synopsis (Figure 3). Nevertheless, there remains a significant gap between VILA-1.5 and Gemini Pro 1.5, particularly in counting problems, action recognition, and temporal perception, indicating substantial room for improvement. Regarding modality, adding subtitles consistently improves the performance of open-source models. Unfortunately, most of the open-source models do not support audio input, so audio evaluation is omitted.

Apart from video MLLMs, we also evaluate the performance of image MLLMs on Video-MME. Table 4 reveals that image-based Qwen-VL-Max and InterVL-Chat-V1.5 attain comparable performance to LLaVA-NeXT-Video, demonstrating their superior generalization capacity on sequential data, and the universality of Video-MME in both image and video MLLMs. It also indicates that image understanding is the foundation of video understanding.

4.3. Analysis

We conduct further analysis to explore the factors influencing the video understanding performance, e.g., additional modality information and video duration.

Could additional modalities benefit the performance? Most evaluations use only video frames as input, requiring models to answer questions based solely on visual context. However, many videos inherently include additional information from other modalities, such as subtitles and audios. To understand their impact, we vary the combinations of input modalities in the evaluation. We can draw the following observations from the results in Table 5. **(1) Introducing subtitles and audios can improve the results.** For instance, in the multilingual task presented in Table 5, Gemini

1.5 Pro achieves a +16.7% and +12.5% accuracy improvement on long videos with the addition of subtitles and audios, respectively, compared to the frame-only setting. This suggests that subtitles and audios contribute essential information for answering the questions. **(2) Subtitles and audios provide greater assistance in understanding long videos compared to short videos.** For example, in Table 5, compared to only using frames, the addition of subtitles improves the model’s performance by 2.8% on short videos and by 10.1% on long videos. This is because test samples of long videos include more challenging reasoning questions, which require the model to utilize subtitles and audio information for accurate responses. **(3) For MLLMs, using subtitles is more effective than audios.** Subtitles are typically transcriptions of speech, focusing on verbal content, while audio includes additional ambient sounds. As shown in Table 5, subtitles consistently provide greater improvement than audio across different durations of videos. Multilingual can be regarded as an exception, possibly due to the quality of the subtitles.

How MLLMs are robust to varied video duration? In Table 4, we respectively compare the performance of different models on short, medium, and long videos. As video duration increases, both open-sourced and commercial models exhibit a significant decline in performance. There are three main reasons for the performance decline. **(1) Increased proportion of difficult tasks.** As shown in the bottom right corner of Figure 2, test samples for long videos contain a higher proportion of reasoning questions, which poses a greater challenge to the model’s capabilities. **(2) Increased sparsity in frame sampling, leading to a reduction in effective input information.** Ideally, for videos of varying lengths, models should sample video frames at a fixed fps to ensure consistent information density in frame sequences [52, 53]. However, existing open-sourced models fix the number of input frames, e.g., 8 frames, resulting in excessively sparse information density as the video length increases. This sparsity prevents the model from retaining all useful visual semantics, hindering accurate predictions. Introducing additional modalities, e.g., subtitles, can effectively supplement the missing information [9]. **(3) Increased difficulty in long context understanding.** Although Gemini 1.5 Pro correspondingly increases the number of sampled frames in the long video, there is still a significant performance degradation. Understanding the long context of either single-modality (LLM) or multi-modality (MLLM) is always a great challenge.

5. Discussions

Our evaluation using Video-MME has revealed several critical insights into the current MLLMs and highlighted areas

for future improvement. In this section, we provide more discussions on potential future directions.

Improving Long Context Modeling Capabilities of MLLMs.

One of the significant challenges identified in our evaluation is the decline in performance as video duration increases. For open-source models, the restricted input frames can become an information bottleneck for understanding the full content of long videos, which advocates for innovative approaches of both architectural and infrastructural context extension. For instance, exploring techniques like ring attention [37], as investigated by large world models [36], and training-free context extension methods could be beneficial [2]. Additionally, developing architectures such as a temporal Q-Former to adaptively identify key frames in the video or compress video tokens to reduce computational overhead based on the questions posed is also worth exploring [13, 53]. In essence, improving long context modeling ability is crucial for the next generation of MLLMs to understand long sequential world dynamics.

Building Datasets with Complex Temporal Understanding.

Our evaluation emphasizes the need for instruction-tuning datasets focused on temporal reasoning, particularly given the prevalence of traditional video datasets with short inputs, such as MSRVTT-QA and ActivityNet-QA. Although there have been efforts to construct high-quality datasets involving complex temporal reasoning in long videos [26, 45], the availability of such datasets remains limited compared to those for text [42, 46] and image [30, 65]. The long-tailed distribution of this data poses challenges for acquisition. Advancements in annotation methods, such as human-in-the-loop frameworks [32] and explorations into automatic data synthesis are crucial [39]. Developing these datasets will enable better use of architectural innovations, providing MLLMs with sufficient training supervision for robust temporal understanding in videos.

6. Conclusion

In this paper, we have introduced Video-MME, which is designed to evaluate MLLMs on video understanding tasks. Our benchmark incorporates a diverse range of video types, temporal durations, and data modalities with high-quality, expert-annotated QA pairs. Our extensive evaluation underscores the need for further advancements in handling longer multimodal data and we hope Video-MME will inspire future research and development in improving the capabilities of MLLMs.

Acknowledgments

This work was funded by National Natural Science Foundation of China under Grant 62441234. We appreciate the

efforts made by Yu Bai, Fangyuan Liu, Yigeng Jiang, and Zezhong Wu in the construction of the benchmark.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 3
- [2] Chen An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. Training-free long-context scaling of large language models. *ArXiv preprint*, 2024. 8
- [3] Kirolos Atallah, Chenhui Gou, Eslam Abdelrahman, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding. *ArXiv preprint*, 2024. 3
- [4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv preprint*, 2023. 3
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv preprint*, 2023. 2, 6
- [6] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Soman, and Sagnak Tasirlar. Fuyu-8b: A multimodal architecture for ai agents. URL: <https://www.addept.ai/blog/fuyu-8b>, 2023. 3
- [7] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 2023. 3
- [8] 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*, 2024. 6
- [9] Sishuo Chen, Lei Li, Shuhuai Ren, Rundong Gao, Yuanxin Liu, Xiaohan Bi, Xu Sun, and Lu Hou. Towards multimodal video paragraph captioning models robust to missing modality. *ArXiv preprint*, 2024. 8
- [10] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *ArXiv preprint*, 2023. 3
- [11] 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*, 2024. 2, 6
- [12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhang-hao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yong-hao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3
- [13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv preprint*, 2023. 3, 8
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3
- [15] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *ArXiv preprint*, 2024. 3
- [16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv preprint*, 2023. 1, 3
- [17] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *ArXiv preprint*, 2024. 6
- [18] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Hetting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *ArXiv preprint*, 2025. 1, 6
- [19] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2024. 3
- [20] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. *ArXiv preprint*, 2024. 3
- [21] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. 3
- [22] Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2024. 3
- [23] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *ArXiv preprint*, 2023. 6
- [24] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *EMNLP*, 2018. 3
- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3

- [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*, 2023. 3, 8
- [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*, 2023. 1, 3, 6
- [28] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, 2020. 3
- [29] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. *ArXiv preprint*, 2021. 3
- [30] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. M³IT: A large-scale dataset towards multi-modal multilingual instruction tuning. *ArXiv preprint*, 2023. 8
- [31] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *ArXiv preprint*, 2024. 3
- [32] Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. *ArXiv preprint*, 2023. 1, 3, 8
- [33] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning unified visual representation by alignment before projection. *ArXiv preprint*, 2023. 1, 6
- [34] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *ArXiv preprint*, 2023. 2, 6
- [35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv preprint*, 2023. 3
- [36] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *ArXiv preprint*, 2024. 8
- [37] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. In *ICLR*, 2024. 8
- [38] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. *ArXiv preprint*, 2024. 6
- [39] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models. *ArXiv preprint*, 2024. 8
- [40] 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*, 2023. 3
- [41] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *ArXiv preprint*, 2024. 1, 3
- [42] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *ArXiv preprint*, 2023. 8
- [43] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv preprint*, 2023. 1, 3
- [44] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Ming-Hui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *ArXiv preprint*, 2023. 3
- [45] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2024. 1, 3, 5, 8
- [46] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022. 8
- [47] 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*, 2023. 3
- [48] OpenAI. GPT-4V(ision) system card, 2023. 2, 6
- [49] OpenAI. GPT-4o system card, 2024. 2, 6
- [50] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yuetong Zhuang, and Siliang Tang. Momenator: Advancing video large language model with fine-grained temporal reasoning. *ArXiv preprint*, 2024. 3
- [51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3
- [52] Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou. TESTA: Temporal-spatial token aggregation for long-form video-language understanding. In *EMNLP*, 2023. 8
- [53] 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*, 2023. 3, 8
- [54] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tianbo Ye, Yang Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. *ArXiv preprint*, 2023. 3
- [55] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, 2024. 1, 2, 6

- [56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv preprint*, 2023. 3
- [57] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark. *ArXiv preprint*, 2024. 3
- [58] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 3
- [59] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 2023. 3
- [60] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos. *ArXiv preprint*, 2024. 3
- [61] Bo Wu and Shoubin Yu. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS*, 2024. 3
- [62] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *ArXiv preprint*, 2024. 3
- [63] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 3
- [64] D. Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yuetong Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 3
- [65] Zhiyang Xu, Trevor Ashby, Chao Feng, Rulin Shao, Ying Shen, Di Jin, Qifan Wang, and Lifu Huang. Vision-flan:scaling visual instruction tuning. *ArXiv preprint*, 2023. 8
- [66] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 2024. 1, 3
- [67] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv preprint*, 2023. 1, 3
- [68] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuetong Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019. 3
- [69] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *ArXiv preprint*, 2023. 3
- [70] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 3
- [71] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv preprint*, 2023. 3
- [72] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ArXiv preprint*, 2024. 3, 4
- [73] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 2, 6

Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis

Supplementary Material

7. Detailed Experimental Settings

Models. We conduct a comprehensive evaluation on four commercial models and nine representative open-source video-based multimodal large language models. To further demonstrate the adaptability of our benchmark to multi-image scenarios, we also include three widely utilized image-based MLLMs as part of the evaluation. The complete list of models evaluated is provided below.

- **Commercial MLLMs:** GPT-4V, GPT-4o, Gemini 1.5 Flash, and Gemini 1.5 Pro.
- **Open-source MLLMs:** Video-LLaVA, ST-LLM, ShareGPT4Video, VideoChat2-Mistral, VILA-1.5, Chat-UniVi-V1.5, VITA-1.0, VITA-1.5, LLaVA-NeXT-Video.
- **Advanced Image MLLMs:** Qwen-VL-Chat/Max and InternVL-Chat-V1.5.

Frame Extraction. A standard approach involves extracting a sequence of frames from the video and interpreting the resulting multi-image inputs. For Gemini 1.5 Pro which supports extremely long multimodal contexts, we sample frames at 1 frame per second for short and medium videos, and at 1 frame every 2 seconds for long videos to ensure API stability. For all other models, frame extraction adheres to their respective official guidelines, uniformly sampling a specified number of frames from the video. The specific numbers of sampled frames are as follows: 10 frames for GPT-4V, 384 for GPT-4o, 8 for Video-LLaVA, 16 for VideoChat2-Mistral, 16 for ShareGPT4Video, 64 for ST-LLM, 64 for Chat-UniVi-V1.5, 32 for VITA-1.0, 16 for VITA-1.5, 32 for LLaVA-NeXT-Video, 8 for VILA-1.5, 4 for both Qwen-VL-Chat and Qwen-VL-Max, 10 for InternVL-Chat-V1.5.

Subtitle Utilization. In the subtitle-enabled setting, all models utilize subtitles corresponding to the timestamps of the sampled video frames. For instance, if 10 frames are sampled from a video, the 10 subtitles that correspond to the respective timestamps of those frames are selected. This approach ensures accurate coherent synchronization between the visual and textual multimodal input for evaluation.

Evaluation. The evaluation follows the format: “entire video frames + complete subtitles/audios (optional) + question with prompt.” Whenever possible, the model’s default prompt is utilized for multiple-choice questions. If unavailable, a standardized prompt is employed as follows:

*This video’s subtitles are listed below: [Subtitles]
Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option. [Question] The best answer is:*

The accuracy is computed by extracting the model’s output using regular expressions and comparing it directly with the ground-truth answer, without relying on external judges like commonly-used ChatGPT.

8. Additional Analysis

How do MLLMs perform on the two highlighted cases in Figure 1? We conduct qualitative evaluation (using frames and subtitles) on the two cases in Figure 1. As analyzed in Section 3.2, these two cases comprehensively examine the model’s capabilities in OCR, attribute perception, object recognition, and long-range temporal reasoning, making them highly challenging. **For the date-related question in Case 1**, Video-LLaVA identifies the date (May 31st) from the frame at 01:10 and subtitles, but fails to perform reasoning based on context and incorrectly determines the year of the event, leading to the erroneous selection of option A. The remaining open-sourced models miscalculate the date 10 days after May 31st during the reasoning process, resulting in the incorrect choice of option C. **For the event-related question in Case 2**, Video-LLaVA, VideoChat2, and ST-LLM incorrectly associate the target person with nearby events, resulting in the selection of incorrect options A or C. In contrast, LLaVA-NeXT-Video and Gemini 1.5 Pro accurately track the events involving the target individual across the entire video, showcasing robust long-range temporal modeling capabilities. They correctly link the target person’s injury at 03:35 with his reappearance at 27:30, identifying the true cause of the injury (option D). In summary, the questions in our benchmark pose significant challenges to the models, which motivates MLLMs to advance both their perception and reasoning capabilities.

Could additional modalities benefit the performance across categories? Figure 4 presents the results of Gemini 1.5 Pro across the 30 subcategories of Video-MME, under the testing modes of frames, frames + subtitles, and frames + audio. The results indicate that subtitles and audio positively contribute to video understanding in multimodal large models. However, the extent of improvement provided by these modalities varies across different domains.

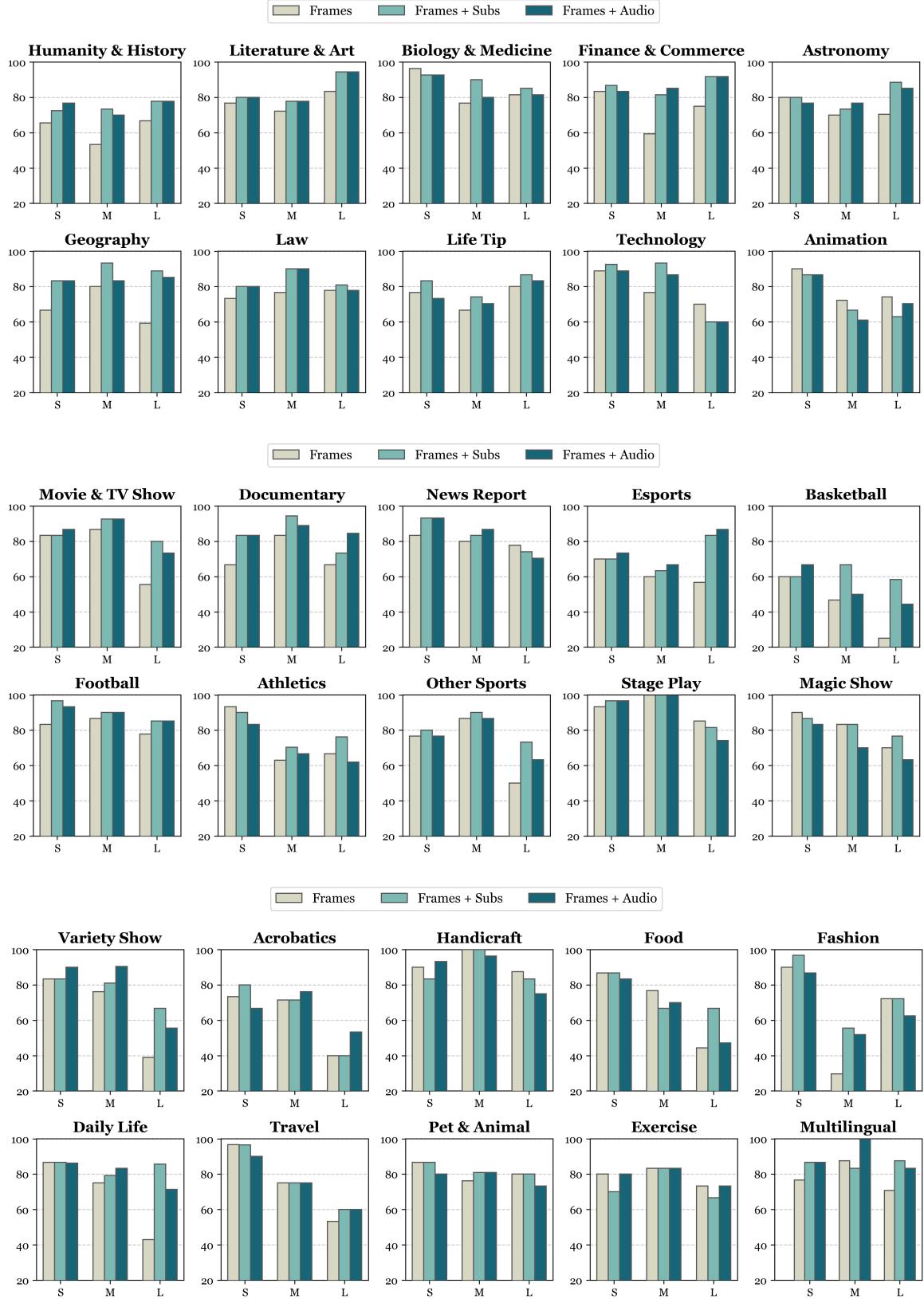


Figure 4. Evaluation results of Gemini 1.5 Pro across different video subcategories.