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

Chaoyou Fu♠, Yuhan Dai[1], Yongdong Luo[2], Lei Li[3], Shuhuai Ren[4]
Renrui Zhang[5], Zihan Wang[6], Chenyu Zhou[2], Yunhang Shen, Mengdan Zhang
Peixian Chen, Yanwei Li[5], Shaohui Lin[6], Sirui Zhao[1], Ke Li
Tong Xu[1] Xiawu Zheng[2], Enhong Chen[1], Rongrong Ji[2], Xing Sun†

[1]USTC, [2]XMU, [3]HKU, [4]PKU, [5]CUHK, [6]ECNU

♠ Project Leader † Corresponding Author

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 in processing sequential visual data is still insufficiently explored, highlighting the absence of a comprehensive, high-quality assessment of their performance. In this paper, we introduce **Video-MME**, the first-ever full-spectrum, **M**ulti-**M**odal **E**valuation 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. 900 videos with a total of 254 hours are manually selected and annotated by repeatedly viewing all the video content, resulting in 2,700 question-answer pairs. With Video-MME, we extensively evaluate various state-of-the-art MLLMs, including GPT-4 series and Gemini 1.5 Pro, as well as open-source image models like InternVL-Chat-V1.5 and video models like LLaVA-NeXT-Video. Our experiments 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, which 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.

Email: {bradyfu24, winfred.sun}@gmail.com

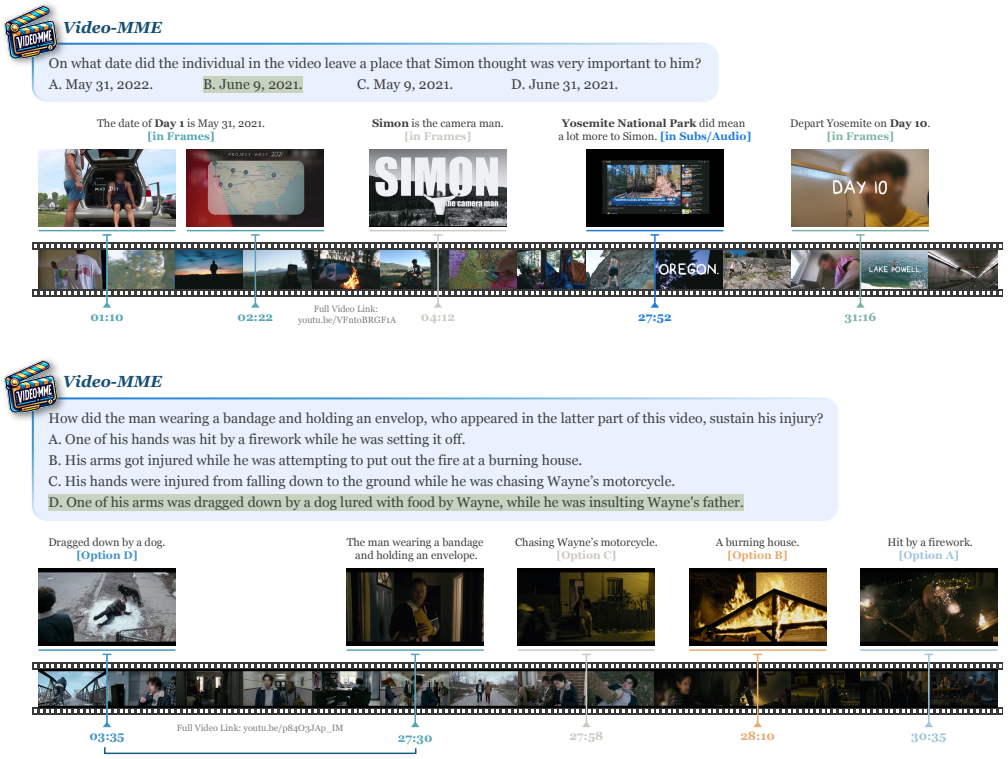

Figure 1: Examples of Video-MME. The ground-truth answer is highlighted in green. In Video-MME, **all data**, including question-answering annotations, videos, subtitles, and audios, are **manually** collected and curated, ensuring diversity and quality.

1 Introduction

The rapid development of multi-modal large language models (MLLMs) in recent years [66, 48, 15, 50, 3, 72] has highlighted their impressive perception and cognitive capabilities across various multimodal benchmarks [13, 67, 42, 73]. These advancements demonstrate the great potential of MLLMs to serve as a foundation that can digest the multi-modal real world [34] 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 [71, 53, 24, 30] have been made to inspire the video understanding potentials of MLLMs with promising results. However, existing video-based benchmarks [29, 24, 45, 39] 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 **M**ulti-**M**odal **E**valuation benchmark crafted 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., astronomy, 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 multi-modal inputs for video understanding.

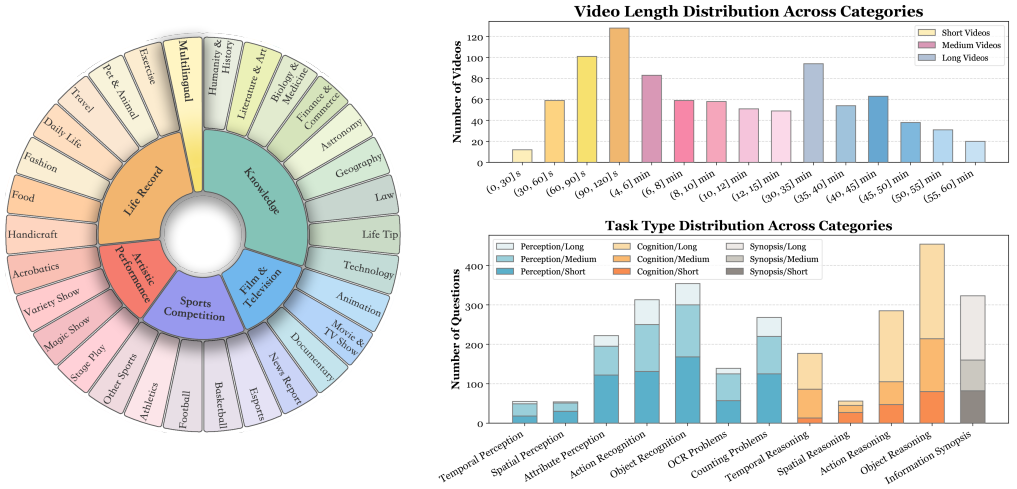

Figure 2: (Left) Video categories. Our benchmark covers 6 key domains and 30 sub-class video types. (Right) Video duration length and question type distributions. Video-MME has a full spectrum of video length and covers different core abilities of MLLMs.

Using Video-MME, we benchmark various state-of-the-art MLLMs, including GPT-4V [48], GPT-4o [49], and Gemini 1.5 Pro [54], alongside open-source image models like InternVL-Chat-V1.5 [9] and video models like LLaVA-NeXT-Video [74]. 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 [31], 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 [9]. 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 filed 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 analysis 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 Video-MME

2.1 Dataset Construction

The dataset construction process of Video-MME consists of three steps: video collection, question-answering annotation, and quality review. The details are as follows.

Video Collection. For a comprehensive coverage of different video types, we first create a domain hierarchy for collecting raw videos from YouTube. We define 6 key domains: Knowledge, Film & Television, Sports Competition, Life Record, and Multilingual, based on popular tendencies on

YouTube. Each domain is further divided into detailed tags, such as football and basketball for sports competition, resulting in a total of 30 fine-grained video classes. The full domain-tag hierarchy and its distribution can be found in the left part of Figure 2. For each class, 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 further investigation. Our final dataset consists of 900 videos with 744 subtitles and 900 audios, spanning various domains with relatively balanced duration lengths, as depicted in the right part of Figure 2.

Question-Answer Annotation. After gathering the raw video data, we annotate it with high-quality question-answer (QA) pairs to evaluate the proficiency of MLLMs in interpreting video content. We employ a multiple-choice QA format to facilitate a straightforward and flexible assessment. The researchers proficient in English with extensive research experience in vision-language learning, perform the annotations. Specifically, they are first asked to watch the whole content of the video, and then to develop 3 corresponding questions, each with 4 potential options by repeatedly watching the video, contributing to 2,700 QA pairs in total. As shown in the bottom right corner of Figure 2, there are a total of 12 task types in the questions, including both perception, reasoning, and information synopsis. Particularly, each QA pair is required to be associated with the video content, avoiding MLLMs being able to answer without looking at the video.

Quality Review. To guarantee the quality of our dataset, we conduct a rigorous manual review process. First, a different annotator is assigned to examine each QA pair to ensure that (i) language expressions are correct and unambiguous; (ii) the question is answerable, and the candidate options and provided golden option are reasonable. Furthermore, to ensure that the questions are challenging enough and require video content as a necessary condition [73], 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 example, the question "What is the biggest achievement of 10 of Argentina in 2022?" that can be directly inferred to the World Cup winner will be filtered out during this process. Questions that do not meet this criterion are returned to the original annotators for revision. By statistics, the accuracy of Gemini 1.5 Pro in the question-only setting is less than 15%. Through our rigorous dataset construction process, we strive to deliver a high-quality, diverse, and well-balanced dataset that will be instrumental for researchers in the field of multi-modal understanding.

2.2 Dataset Statistics

Here, 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 is more uniformly distributed. In the category of long videos, there is a long-tailed distribution where longer videos have fewer samples. 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. In contrast, longer videos mainly feature tasks related to temporal reasoning. Overall, this analysis highlights that our dataset covers a wide range of video durations and various task types, enabling a comprehensive evaluation of temporal understanding.

QA Pair. We demonstrate the language diversity of the questions and answers in our dataset. Table 2 lists the average word count of the textual fields in our dataset. The word counts for questions, options, and answers display notable consistency across different video lengths, suggesting a uniform style of QA pairs in our dataset. On the other hand, the word count for subtitles increases significantly with the length of the videos, e.g., short videos have an average word count of 198.6, while the long video subset have the count up to 6.5K. This trend indicates that longer videos contain more information, as evidenced by the increased volume of subtitles. The analysis reveals that our questions are diverse,

Table 1: Analysis of Certificate Length in seconds. **Avg. V.L.**: average video length, **Med. C.L.**: median certificate length, **Avg. C.L.**: average certificate length.

Video	Avg. V.L.	Med. C.L.	Avg. C.L.
EgoSchema [45]	180	$\sim$ 100	-
Short	82.5	26.0	28.8
Medium	562.7	164.7	160.0
Long	2385.5	890.7	967.7

Table 2: Average word count of different textual fields in Video-MME.

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

and the answers are well-balanced. In addition, the distribution of the four answer 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 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 with extra annotators. As shown in Table 1, 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 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.

Comparison with Previous Benchmarks. We compare the key difference of our dataset with previous benchmarks in Table 3. The first block lists traditional video benchmarks, which typically focus on specific domains such as TV videos or lack a clear hierarchy, making them less suitable for comprehensively diagnosing the limitations of MLLMs. In the middle block, although several benchmarks like TempCompass [39] and MVBench [24] explore multi-level evaluation and source videos from open domains, they still only cover videos with shorter durations. For example, the longest dataset, EgoSchema [45], includes videos up to 180 seconds, leaving the understanding of longer videos unaddressed. Our Video-MME is the first manually annotated benchmark that encompasses open-domain videos with durations ranging from 11 seconds to 1 hour. It evaluates

Table 3: 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 that if a dataset includes multiple task formats, our comparison focuses solely on the multiple-choice segment.

Benchmarks	#Videos	#Clips	Len.(s)	#QA Pairs	Anno.	QA Tokens	Sub. Tokens	Multi-level	Open-domain	Sub.&Aud.
MSRVTT-QA [63]	2,990	2,990	15.2	72,821	A	8.4	✗	✗	✓	✗
MSVD-QA [63]	504	504	9.8	13,157	A	7.6	✗	✗	✓	✗
TGIF-QA [18]	9,575	9,575	3.0	8,506	A&M	20.5	✗	✗	✓	✗
ActivityNet-QA [68]	800	800	111.4	8,000	M	10.2	✗	✗	✗	✗
TVQA [21]	2,179	15,253	11.2	15,253	M	27.8	159.8	✗	✗	✗
How2QA [25]	1,166	2,852	15.3	2,852	M	16.9	31.1	✗	✓	✗
STAR [61]	914	7,098	11.9	7,098	A	19.5	✗	✗	✓	✗
NExT-QA [62]	1,000	1,000	39.5	8,564	A	25.3	✗	✗	✓	✗
MVBench [24]	3,641	3,641	16.0	4,000	A	27.3	✗	✗	✓	✗
Video-Bench [47]	5,917	5,917	56.0	17,036	A&M	21.3	✗	✗	✓	✗
EgoSchema [45]	5,063	5,063	180.0	5,063	A&M	126.8	✗	✗	✗	✗
AutoEval-Video [8]	327	327	14.6	327	M	11.9	✗	✗	✓	✗
TempCompass [39]	410	500	11.4	7,540	A&M	49.2	✗	✗	✓	✗
Video-MME-S	300	300	80.7	900	M	28.7	198.6	✓	✓	✓
Video-MME-M	300	300	515.9	900		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			

different levels of video understanding ability and is supplemented with meta information such as subtitles and audios. This comprehensive approach uniquely positions Video-MME to advance the evaluation and development of MLLMs.

3 Experiments

In this section, we evaluate a wide range of MLLMs on our Video-MME benchmark. We first introduce the evaluation settings, and then present 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 information and duration length.

3.1 Settings

We conduct the evaluation on 4 commercial models, i.e., GPT-4V, GPT-4o, Gemini 1.5 Flash, and Gemini 1.5 Pro. Representative open-source video MLLMs including Video-LLaVA, VideoChat2-Mistral, ST-LLM, Chat-UniVi-V1.5, LLaVA-NeXT-Video, and VILA-1.5 are evaluated as well. In addition, we also include advanced image MLLMs, i.e., Qwen-VL-Chat/Max and InternVL-Chat-V1.5, which usually can generalize to multi-image scenarios. We follow their official configurations and try to use more frames[1] for evaluation. A special case is Gemini 1.5 Pro, because it supports extremely long multimodal contexts, so we take frames per second for both short and medium videos. For long videos, we capture a frame every two seconds to ensure the stability of the API. With respect to the setting of adding subtitles, all models except Gemini 1.5 Pro use the subtitles corresponding to the sampled video frames. For example, if you sample 10 frames, take the 10 subtitles that correspond to the time of those 10 frames. Gemini 1.5 Pro uses all subtitles due to the full video frame sampling. Besides, only Gemini 1.5 Pro supports the input of audios by now, whose results are listed in Table 5.

The evaluation adopts the format of "whole video frames + whole subtitles/audios (optional) + question with prompt". We try to use the model's default prompt for multiple-choice questions, but if not we use a common prompt as:

[1]The numbers of the sampled frames are 10 for GPT-4V, 384 for GPT-4o, 8 for Video-LLaVA, 16 for VideoChat2-Mistral, 64 for ST-LLM, 64 for Chat-UniVi-V1.5, 32 for LLaVA-NeXT-Video, 8 for VILA-1.5, 4 for Qwen-VL-Chat/Max, and 10 for InternVL-Chat-V1.5.

Table 4: Performance of MLLMs on Video-MME with short, medium, and long durations, under the settings of "without subtitles" and "with subtitles".

Models	LLM Params	Short (%)		Medium (%)		Long (%)		Overall (%)	
		w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs
Open & Closed-source Image MLLMs									
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 [9]	20B	**60.2**	**61.7**	**46.4**	**49.1**	45.6	46.6	50.7	**52.4**
Open-source Video MLLMs									
Video-LLaVA [30]	7B	45.3	46.1	38.0	40.7	36.2	38.1	39.9	41.6
ST-LLM [36]	7B	45.7	48.4	36.8	41.4	31.3	36.9	37.9	42.3
VideoChat2-Mistral [24]	7B	48.3	52.8	37.0	39.4	33.2	39.2	39.5	43.8
Chat-UniVi-V1.5 [19]	7B	45.7	51.2	40.3	44.6	35.8	41.8	40.6	45.9
LLaVA-NeXT-Video [74]	34B	61.7	65.1	50.1	52.2	44.3	47.2	52.0	54.9
VILA-1.5 [31]	34B	**68.1**	**68.9**	**58.1**	**57.4**	**50.8**	**52.0**	**59.0**	**59.4**
Closed-source MLLMs									
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 [54]	-	78.8	79.8	68.8	74.7	61.1	68.8	70.3	75.0
Gemini 1.5 Pro [54]	-	**81.7**	**84.5**	**74.3**	**81.0**	**67.4**	**77.4**	**75.0**	**81.3**

Table 5: Performance of Gemini 1.5 Pro on 6 major categories of Video-MME. The input modality includes frames only, frames with subtitles, and frames with audios.

Subset	Modality	Category						
		Knowledge	*Film & Television*	*Sports Competition*	*Artistic Performance*	*Life Record*	*Multilingual*	Overall
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)*

> *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:*

Considering the test sample in our benchmark is a multi-choice question with 4 options, we take accuracy as the evaluation metric, the random guess of which is 25%. The accuracy is calculated by matching the output of the model with the real one, without introducing any third party model such as ChatGPT. The example of accuracy calculation can be found on our project page.

3.2 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, achieving an accuracy of 78.8%, while performing the lowest in the Sports Competition category, with an accuracy of 68.6%. As video duration increases, Gemini 1.5 Pro's performance declines (e.g., −6.7% 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, Chat-UniVi-V1.5 achieves the best performance with 40.6%. 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, as exhibited in Figure 3. Nevertheless, there still 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. It is observed that adding subtitles can also help the open-sourced models. For example, the accuracy of LLaVA-NeXT-Video improves from 52% to 54.9%. It is regrettable that none of the open-sourced models support audio input.

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. Meanwhile, it also indicates that image understanding is the foundation of video understanding.

We conduct qualitative evaluation (using frames and subtitles) on the two cases in Figure 1. As analyzed in Section 2.2, 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 successfully identifies the events experienced by the target person throughout the video and demonstrates long-distance 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.

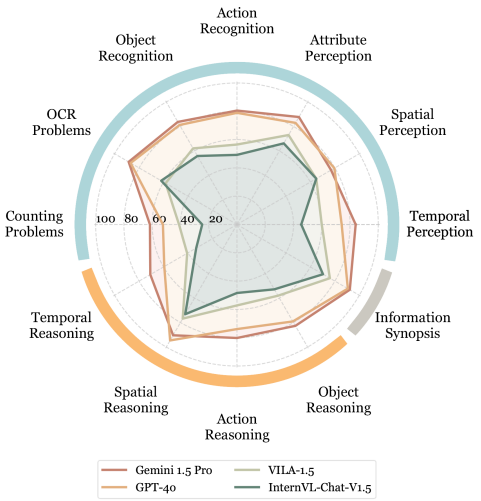

Figure 3: Performance on 12 types of tasks.

3.3 Analysis

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

Could additional modalities benefit the performance? Most of evaluations only take video frames as input, requiring models to answer questions solely based on visual contexts. However, many videos inherently contain extra information from other modalities, such as subtitles and audios. To understand their impact, we vary the combinations of input modalities in the evaluation, and report

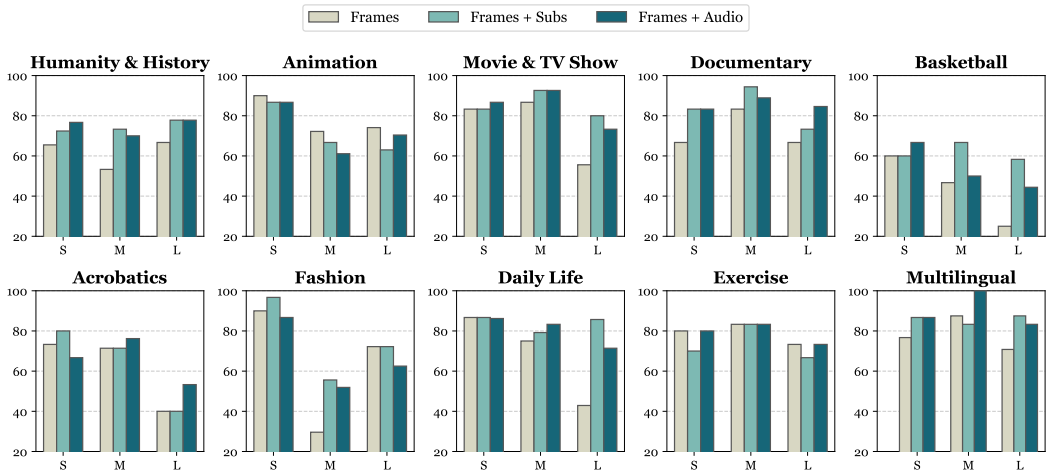

Figure 4: The impact of incorporating additional modality inputs, i.e., subtitles and audios, on the performance of Gemini 1.5 Pro across different types of videos. We only show the results of 10 classes here, and those of the whole 30 classes can be found on our project page.

results in Table 5 and Figure 4. We can draw the following observations. **(1) Introducing subtitles and audios can improve the results.** For example, in the multilingual task shown in Table 5, with the addition of subtitles/audios, Gemini 1.5 Pro achieves +16.7%/+12.5% accuracy on long videos compared to the frame-only setting. This indicates that subtitles and audios provide some necessary information to answer 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 requires the model to utilize subtitles and audio information for accurate responses. **(3) For MLLMs, using subtitles is more effective than audios.** Subtitles are usually transcriptions of audio, primarily capturing speech content, while audios encompass more ambient sounds. As can be seen from Table 5, whether it is short, medium, or long videos, subtitles bring higher improvement than audios. There are some exceptions in Table 5 and Figure 4, such as Multilingual, which may be due to the quality of the subtitles themselves. In addition, the audio also includes some content that subtitles can not express, such as singing and intonation.

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. For example, the accuracy of VILA-1.5 drops by 10% and 17.3% from short to medium and long videos, respectively, while Gemini 1.5 Pro's accuracy decreases by 7.4% and 14.3%. 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. These questions are more challenging than perception and recognition tasks, thus posing 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 [56, 55]. However, many 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 [7]. **(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.

4 Discussions

Our evaluation using Video-MME has revealed several critical insights into the current MLLMs and highlighted areas for future improvement. Here, we further discuss 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. Innovative approaches to context extension, both architectural and infrastructural, are essential. For instance, exploring techniques like ring attention [35], as investigated by large world models [34], 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 [56, 11]. In essence, improving long context modeling ability is crucial for the next generation of MLLMs to understand long sequential world dynamics effectively.

Building Datasets with Complex Temporal Understanding. Our evaluation also highlights the demand for temporal reasoning oriented instruction-tuning datasets, considering that traditional video datasets with short video inputs, such as MSRVTT-QA and ActivityNet-QA. Although there have been efforts to construct high-quality datasets involving complex temporal reasoning over long videos [45, 23], the availability of such datasets is still insufficient compared to the text only [40, 46] and image datasets [27, 65]. The long-tailed nature of this data makes it challenging to acquire. Efforts towards better annotation paradigms such as human-in-the-loop frameworks [29] and automatic data synthesizing explorations are crucial [37]. Developing such datasets can better leverage advanced architectural innovations to provide MLLMs with sufficient training supervision for a robust understanding of the temporal dimension of videos.

5 Related Work

Advancements in MLLMs. Recent advancements in MLLMs have seen notable progress [66, 14]. 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 [52] and SigLIP [70] are widely-used for image encoding, while LLaMA [58] and Vicuna [10] serve as popular choices for LLMs. The alignment module varies from simple linear projections [33, 75] to more complex architectures such as Q-Former [22, 11], and gated cross-attention layers substantiated by Flamingo and IDEFICS [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 [23, 71, 43, 57, 32, 44, 19, 20, 64], 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 [12] with an image Q-Former to encode individual frames and then employs a video Q-Former for temporal modeling. VideoChat2 [24] utilizes a video transformer to encode video features and subsequently implements a Q-Former [22] to compress video tokens. To empower video MLLMs with temporal localization capability [17, 51, 60], TimeChat [56] constructs time-sensitive instruction tuning datasets and encodes timestamp knowledge into visual tokens. VTimeLLM [16] proposes a LLaVA-like three-stage training method. However, the potential of MLLMs in processing sequential visual data is still under-explored. Therefore, we introduce Video-MME for full-spectrum, multi-modal evaluation of MLLMs in video analysis.

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 [13, 67, 38]. As image MLLMs have demonstrated exceptional performance in general perception tasks, benchmarks regarding scientific understanding [28], multi-modal mathematical reasoning [41, 73], and multi-disciplinary [69] capabilities have drawn increasing attention. For video MLLMs, similar efforts have been made to incorporate existing benchmarks [59, 26] for evaluating video understanding [24, 47]. Given the temporal nature of video modalities, specific benchmarks have

been developed to address temporal understanding, highlighting the limitations of current video MLLMs in comprehending video content [29, 39].

In this work, we introduce a new high-quality video understanding benchmark, Video-MME. Compared to previous benchmarks, Video-MME includes a diverse set of videos of varying durations, supplemented with external modalities such as audios and subtitles. Additionally, it features human-annotated multi-level QA pairs, providing a comprehensive assessment framework for MLLMs. Our results indicate that open-source MLLMs still have a large gap with closed models. Our analysis and discussion further shed lights on the future development of MLLMs.

6 Conclusion

In this paper, we have introduced **Video-MME**, the first comprehensive multi-modal benchmark designed to evaluate MLLMs for video tasks. Our benchmark incorporates a diverse range of video types, varying temporal durations, and multiple data modalities, all annotated with high-quality, expert-labeled QA pairs. Our extensive evaluation of state-of-the-art MLLMs, including commercial and open-source models, highlights significant performance differences. Commercial models, particularly Gemini 1.5 Pro, demonstrate superior performance compared to open-source variants. The integration of subtitles and audio tracks significantly enhances video understanding, especially for longer videos. However, we observe a general decline in performance as video duration increases. These findings underscore the need for further advancements in handling longer multi-modal data. We hope Video-MME will inspire future research and development in improving the capabilities of MLLMs.

Acknowledgments

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

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv preprint*, 2022.

[2] C. An, F. Huang, J. Zhang, S. Gong, X. Qiu, C. Zhou, and L. Kong. Training-free long-context scaling of large language models. *ArXiv preprint*, 2024.

[3] Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024.

[4] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv preprint*, 2023.

[5] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv preprint*, 2023.

[6] R. Bavishi, E. Elsen, C. Hawthorne, M. Nye, A. Odena, A. Somani, and S. Taşırlar. Introducing our multimodal models, 2023.

[7] S. Chen, L. Li, S. Ren, R. Gao, Y. Liu, X. Bi, X. Sun, and L. Hou. Towards multimodal video paragraph captioning models robust to missing modality. *ArXiv preprint*, 2024.

[8] X. Chen, Y. Lin, Y. Zhang, and W. Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *ArXiv preprint*, 2023.

[9] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *ArXiv preprint*, 2024.

[10] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[11] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv preprint*, 2023.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[13] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv preprint*, 2023.

[14] C. Fu, R. Zhang, H. Lin, Z. Wang, T. Gao, Y. Luo, Y. Huang, Z. Zhang, L. Qiu, G. Ye, et al. A challenger to gpt-4v? early explorations of gemini in visual expertise. *ArXiv preprint*, 2023.

[15] Gemini Team. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, 2023.

[16] B. Huang, X. Wang, H. Chen, Z. Song, and W. Zhu. Vtimellm: Empower llm to grasp video moments. *ArXiv preprint*, 2023.

[17] D.-A. Huang, S. Liao, S. Radhakrishnan, H. Yin, P. Molchanov, Z. Yu, and J. Kautz. Lita: Language instructed temporal-localization assistant. *ArXiv preprint*, 2024.

[18] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017.

[19] P. Jin, R. Takanobu, C. Zhang, X. Cao, and L. Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *ArXiv preprint*, 2023.

[20] R. Jung, H. Go, J. Yi, J. Jang, D. Kim, J. Suh, A. S. Lee, C. Han, J. Lee, J. Kim, J.-Y. Kim, J. Kim, K. Park, L. Lee, M. Ha, M. Seo, A. Jo, E. Park, H. Kianinejad, S. Kim, T. Moon, W. Jeong, A. Popescu, E. Kim, E. Yoon, G. Heo, H. Choi, J. Kang, K. Han, N. Seo, S. Nguyen, R. Won, Y. E. Park, A. Giuliani, D. Chung, H. Yoon, J. Le, J. Ahn, J. Lee, M. Saini, M. Sanders, S. Lee, S. Kim, and T. Couture. Pegasus-v1 technical report. *ArXiv preprint*, 2024.

[21] J. Lei, L. Yu, M. Bansal, and T. Berg. TVQA: Localized, compositional video question answering. In *EMNLP*, 2018.

[22] J. Li, D. Li, S. Savarese, and S. C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[23] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao. Videochat: Chat-centric video understanding. *ArXiv preprint*, 2023.

[24] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *ArXiv preprint*, 2023.

[25] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, 2020.

[26] L. Li, J. Lei, Z. Gan, L. Yu, Y.-C. Chen, R. Pillai, Y. Cheng, L. Zhou, X. E. Wang, W. Y. Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. *ArXiv preprint*, 2021.

[27] L. Li, Y. Yin, S. Li, L. Chen, P. Wang, S. Ren, M. Li, Y. Yang, J. Xu, X. Sun, L. Kong, and Q. Liu. M^3IT: A large-scale dataset towards multi-modal multilingual instruction tuning. *ArXiv preprint*, 2023.

[28] L. Li, Y. Wang, R. Xu, P. Wang, X. Feng, L. Kong, and Q. Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *ArXiv preprint*, 2024.

[29] S. Li, L. Li, S. Ren, Y. Liu, Y. Liu, R. Gao, X. Sun, and L. Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. *ArXiv preprint*, 2023.

[30] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan. Video-llava: Learning united visual representation by alignment before projection. *ArXiv preprint*, 2023.

[31] J. Lin, H. Yin, W. Ping, Y. Lu, P. Molchanov, A. Tao, H. Mao, J. Kautz, M. Shoeybi, and S. Han. Vila: On pre-training for visual language models. *ArXiv preprint*, 2023.

[32] H. Liu, Q. Fan, T. Liu, L. Yang, Y. Tao, H. Huang, R. He, and H. Yang. Video-teller: Enhancing cross-modal generation with fusion and decoupling. *ArXiv preprint*, 2023.

[33] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *ArXiv preprint*, 2023.

[34] H. Liu, W. Yan, M. Zaharia, and P. Abbeel. World model on million-length video and language with ringattention. *ArXiv preprint*, 2024.

[35] H. Liu, M. Zaharia, and P. Abbeel. Ring attention with blockwise transformers for near-infinite context. In *ICLR*, 2024.

[36] R. Liu, C. Li, H. Tang, Y. Ge, Y. Shan, and G. Li. St-llm: Large language models are effective temporal learners. *ArXiv preprint*, 2024.

[37] R. Liu, J. Wei, F. Liu, C. Si, Y. Zhang, J. Rao, S. Zheng, D. Peng, D. Yang, D. Zhou, et al. Best practices and lessons learned on synthetic data for language models. *ArXiv preprint*, 2024.

[38] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? *ArXiv preprint*, 2023.

[39] Y. Liu, S. Li, Y. Liu, Y. Wang, S. Ren, L. Li, S. Chen, X. Sun, and L. Hou. Tempcompass: Do video llms really understand videos? *ArXiv preprint*, 2024.

[40] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *ArXiv preprint*, 2023.

[41] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv preprint*, 2023.

[42] P. Lu, H. Bansal, T. Xia, J. Liu, C. yue Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv preprint*, 2023.

[43] R. Luo, Z. Zhao, M. Yang, J. Dong, M.-H. Qiu, P. Lu, T. Wang, and Z. Wei. Valley: Video assistant with large language model enhanced ability. *ArXiv preprint*, 2023.

[44] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv preprint*, 2023.

[45] K. Mangalam, R. Akshulakov, and J. Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2024.

[46] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022.

[47] M. Ning, B. Zhu, Y. Xie, B. Lin, J. Cui, L. Yuan, D. Chen, and L. Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *ArXiv preprint*, 2023.

[48] OpenAI. GPT-4V(ision) system card, 2023.

[49] OpenAI. GPT-4o system card, 2024.

[50] A. Ormazabal, C. Zheng, C. d. M. d'Autume, D. Yogatama, D. Fu, D. Ong, E. Chen, E. Lamprecht, H. Pham, I. Ong, et al. Reka core, flash, and edge: A series of powerful multimodal language models. *ArXiv preprint*, 2024.

[51] L. Qian, J. Li, Y. Wu, Y. Ye, H. Fei, T.-S. Chua, Y. Zhuang, and S. Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. *ArXiv preprint*, 2024.

[52] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[53] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, 2024.

[54] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. P. Lillicrap, J.-B. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, I. Antonoglou, R. Anil, S. Borgeaud, A. M. Dai, K. Millican, E. Dyer, M. Glaese, T. Sottiaux, B. Lee, F. Viola, M. Reynolds, Y. Xu, J. Molloy, J. Chen, M. Isard, P. Barham, T. Hennigan, R. McIlroy, M. Johnson, J. Schalkwyk, E. Collins, E. Rutherford, E. Moreira, K. W. Ayoub, M. Goel, C. Meyer, G. Thornton, Z. Yang, H. Michalewski, Z. Abbas, and e. Nathan Schucher. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, 2024.

[55] S. Ren, S. Chen, S. Li, X. Sun, and L. Hou. TESTA: Temporal-spatial token aggregation for long-form video-language understanding. In *Findings of EMNLP*, 2023.

[56] S. Ren, L. Yao, S. Li, X. Sun, and L. Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *ArXiv preprint*, 2023.

[57] E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, X. Guo, T. Ye, Y. Lu, J.-N. Hwang, and G. Wang. Moviechat: From dense token to sparse memory for long video understanding. *ArXiv preprint*, 2023.

[58] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv preprint*, 2023.

[59] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019.

[60] Y. Wang, X. Meng, J. Liang, Y. Wang, Q. Liu, and D. Zhao. Hawkeye: Training video-text llms for grounding text in videos. *ArXiv preprint*, 2024.

[61] B. Wu and S. Yu. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS*, 2024.

[62] J. Xiao, X. Shang, A. Yao, and T. Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021.

[63] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017.

[64] L. Xu, Y. Zhao, D. Zhou, Z. Lin, S. K. Ng, and J. Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *ArXiv preprint*, 2024.

[65] Z. Xu, T. Ashby, C. Feng, R. Shao, Y. Shen, D. Jin, Q. Wang, and L. Huang. Vision-flan:scaling visual instruction tuning. *ArXiv preprint*, 2023.

[66] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. A survey on multimodal large language models. *ArXiv preprint*, 2023.

[67] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv preprint*, 2023.

[68] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019.

[69] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *ArXiv preprint*, 2023.

[70] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[71] H. Zhang, X. Li, and L. Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv preprint*, 2023.

[72] R. Zhang, J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In *ICLR*, 2023.

[73] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, P. Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ArXiv preprint*, 2024.

[74] Y. Zhang, B. Li, h. Liu, Y. j. Lee, L. Gui, D. Fu, J. Feng, Z. Liu, and C. Li. Llava-next: A strong zero-shot video understanding model, 2024.

[75] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv preprint*, 2023.