MMBench-Video: A Long-Form Multi-Shot Benchmark for Holistic Video Understanding

Kangrui Mao[1,2] Haodong Duan[2] Yuanhan Zhang[2,3] Songyang Zhang[2]
Dahua Lin[2] Kai Chen[2]
[1]Shanghai Jiao Tong University [2]Shanghai AI Laboratory
[3]Nanyang Technological University
karrymao@sjtu.edu.cn duanhaodong@pjlab.org.cn yuanhan002@e.ntu.edu.sg
{zhangsongyang, lindahua, chenkai}@pjlab.org.cn

Abstract

With the rise of large language models, many efforts have been made to extend their capabilities to multimodal understanding, including video understanding. Traditional benchmarks like NeXT-QA [29], while offering quantitative measurements, often fall short in capturing the diverse range of video content. Moreover, their QA formats do not comprehensively evaluate the temporal understanding capabilities of models. We present a pioneering quantitative benchmark, MMBench-Video, to evaluate the performance of large vision-language models (LVLMs) in video understanding. MMBench-Video features long videos sourced from YouTube and free-form QA, reflecting real-world application scenarios. Focusing on video, we formulate questions that assess the model's temporal understanding capabilities. All questions are annotated by humans within a meticulously designed ability taxonomy. We leverage ChatGPT to achieve automated evaluation. MMBench-Video will serve as a valuable resource for the research community, facilitating improved evaluation of models and fostering future advancements in video understanding.

1. Introduction

Video, a ubiquitous multimedia format in our daily lives, serves an indispensable function. The significant proliferation of online videos caters to people's constant demands for information and entertainment. This contributes to an extensive and diverse network of visual experiences. The widespread accessibility and consumption of video content have reshaped how we communicate, learn, and connect in the digital age.

Video is more than just a source of visual enjoyment; it serves as a tool for transmitting knowledge and information. In the field of artificial intelligence, understanding video content is crucial. Videos contain contextual, emotional, and linguistic details, enabling artificial intelligence to explore and comprehend various facets of human culture, behavior, and social communication. This versatile nature of videos not only enhances the AI's comprehension but also presents opportunities for a nuanced examination of the complexities associated with human experiences and interactions.

The significant progress in computer vision and natural language processing has resulted in the emergence of various video models. Can video models understand videos like humans do? Can these models acquire knowledge from videos? Can they identify harmful content in video streams? This extends beyond mere image perception, encompassing nuanced comprehension and reasoning of context, emotion, and abstract concepts. Therefore, the evaluation and measurement of video model capabilities is an urgent task.

To address this problem, we have developed a novel quantitative benchmark, **MMBench-Video**, to evaluate the effectiveness of large vision-language models (LVLMs) in video understanding. MMBench-Video differs from conventional benchmarks in several key aspects:

- **Longer Videos and Free-Form QAs:** Unlike common benchmarks that focus primarily on visually classifying short videos, MMBench-Video shifts its focus to longer videos and evaluates the model's ability to answer free-form questions. In addition, it is a multimodal benchmark that considers visuals, speech, and text as important signals embedded in the video.
- **Dynamic Temporal Understanding:** MMBench-Video is carefully designed to emphasize the temporal understanding capabilities of models. Test cases evaluate only dynamic understanding, distinguishing it as a more authentically "video-centric" benchmark.
- **Comprehensive Evaluation through Diverse Abilities:** MMBench-Video explicitly includes videos from differ-

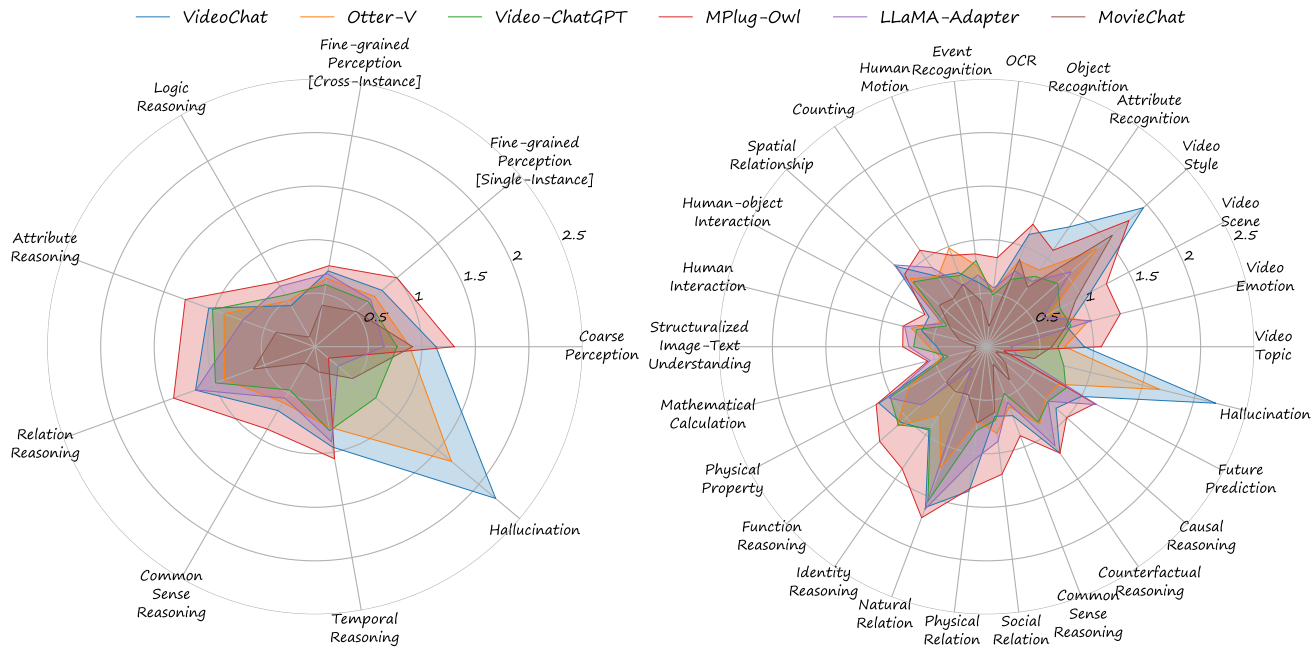


Figure 1. **Comparison of six mainstream video models on MMBench-Video.** The left figure illustrates performance in terms of broad (L2) abilities, while the right figure breaks down performance across detailed (L3) ability dimensions. Refer to Tab. 2 for comprehensive score details.

ent subject areas, with questions that assess numerous skills. The benchmark covers 26 specific abilities, allowing for a thorough evaluation of an LVLM's comprehension abilities.

To address the challenge of scoring free-form questions, we propose to use ChatGPT powered by GPT-4 [23] for automated scoring. Using a 3-point scale, the method focuses on semantic similarity, ignoring grammatical differences. Experimental results show agreement with human judgment, highlighting the reliability and consistency of the scoring system. This approach improves scoring by accounting for semantic nuance and provides a detailed analysis of the model's performance in generating responses to open-ended questions.

We conduct a comprehensive evaluation of six mainstream video models on MMBench-Video, reporting their performance across diverse capability dimensions (refer to Fig. 1). Performance rankings facilitate direct model comparisons, revealing insights into current video model shortcomings. With the help of MMBench-Video and other multimodal benchmarks, we thoroughly evaluated different approaches to understanding videos. This holistic assessment aims to contribute valuable information for advancing video model capabilities. It turns out that purely visual approaches to video understanding exhibit weak basic capabilities. In contrast, multimodal approaches that leverage visual, speech, and text modalities demonstrate enhanced performance on MMBench-Video, underscoring the efficacy of a holistic, multimodal understanding paradigm. To summarize our contribution:

1. **New Video Understanding Dataset:** MMBench-Video addresses the limitations of earlier datasets with regards to video duration and variety of questions. The videos within this dataset are sourced directly from popular online platforms, providing a more genuine representation of real-life scenarios.
2. **Comprehensive Evaluation:** We conducted a thorough evaluation of current video models on the MMBench-Video dataset, providing scores on various fine-grained abilities. This assessment helps the research community grasp the current state of video model development.
3. **Observation and Discovery:** Through observation of the current model's performance, we point out its limited understanding of videos, especially long videos, indicating the need for further development in this area. Additionally, we carry out several experiments to elucidate the model's shortcomings, offering valuable perspectives for future optimization.

2. Related Work

2.1. Video Question Answering Benchmarks

Video Question Answering (VideoQA) stands as a pivotal benchmark for evaluating the advancement of AI models

in comprehending video content. The research community has introduced a diverse spectrum of VideoQA benchmarks spanning various visual domains. These domains include human-centric videos, such as movies [26] and TV shows [11, 15], social media content [7, 12, 13, 28, 29, 31, 34], object-centric videos [22], synthetic scenarios [33], and egocentric videos [9]. These datasets traditionally target short videos, demanding concise answers. In contrast, our dataset deviates by focusing on longer videos, challenging models to produce detailed, free-form responses to intricate questions. This distinction underscores the unique contribution of our dataset in evaluating models' proficiency in handling extended video content and generating nuanced answers to complex queries.

2.2. Video-Language Model

Recent advances in large language models (LLMs) such as the GPT series [5, 24], LLaMA [27], and Vicuna [8], have significantly improved video-language models. Pioneering examples such as Flamingo [2, 3] and FrozenBiLM [30], demonstrate notable zero-shot learning capabilities. These models integrate a vision encoder with a language decoder for enhanced video understanding. With growing demand for human-like interaction, recent video-language models emphasize advanced, multi-modal human-model interactions. This progress is crucial for developing advanced chatbot technologies, which aim to create instruction-tuned models such as Otter-V [16], VideoChat [17], and LLaMA-Adapter [10], among others [21, 32, 35].

3. MMBench-Video

This section presents an overview of MMBench-Video, including video and question selection and the development of the ability taxonomy. We introduce the production pipeline of the dataset and give various statistics. Importantly, we explain how to leverage ChatGPT for fully automated evaluation. The effectiveness and reasoning behind this approach are detailed. This section aims to offer insights into the benchmark's design philosophy, shedding light on its principles and methodologies.

3.1. Video Content

Previous video QA datasets, such as NeXT-QA [29], have shown limited variation in video content, mainly featuring single-shot videos. For verification, we process the videos in NeXT-QA test set with PySceneDetect [6]. Among 1000 videos, 742 videos only have one single shot. In contrast, our MMBench-Video dataset overcomes this limitation by incorporating videos with significantly higher complexity, reflecting the intricately edited and processed nature of videos watched daily. During the MMBench-Video collection, we noticed a difference in the number of shots per video, with an average of 32.2 shots per video, a significant boost compared to NeXT-QA. For a more in-depth analysis and comparison, please refer to Fig. 3b. These results highlight the increased diversity and realism incorporated in our dataset.

Games	Food & Drink
Autos & Vehicles	Business & Industrial
Computers & Electronics	Sports
People	Instruction Video
Films & TV Shows	Pets & Animals
Advertisements	Science
Humor	Knowledge
News	Finance

Table 1. **Video Categories in MMBench-Video.** The 16 categories range from engaging topics like Entertainment and Sports to informative subjects such as Science and Knowledge, ensuring a comprehensive evaluation of diverse content genres.

MMBench-Video sources its content from YouTube, affording notable advantages. Firstly, leveraging YouTube facilitates access to extensive metadata, encompassing video titles, click metrics, and subtitles, enriching the contextual understanding of the videos. Secondly, as the foremost global streaming platform, YouTube ensures dataset diversity through its broad global user base. Our categorization approach draws inspiration from YouTube-8M [1] and is adapted with minor modifications, culminating in the video category in MMBench-Video (Tab. 1).

3.2. Ability Taxonomy

Inherited from MMBench [19], we formulated a 3-level hierarchical ability taxonomy (refer to Fig. 2. At the top level, we delineate two general abilities: Perception and Reasoning. Within the Perception domain, our taxonomy encompasses Coarse Perception, Fine-grained Single-Instance Perception, and Fine-grained Cross-Instance Perception, representing varying degrees of perceptual acuity. In the Reasoning domain, we further refine our taxonomy to include Logic Reasoning, Attribute Reasoning, Relation Reasoning, Common Sense Reasoning, and Temporal Reasoning, capturing diverse facets of cognitive processing.

Additionally, we introduce Hallucination in the benchmark. This distinct dimension serves to gauge the extent to which a model generates content that is nonsensical or entirely fictional. By introducing this dimension, we aim to provide a comprehensive evaluation framework that not only scrutinizes the model's adherence to reality but also tests its susceptibility to generating misleading or inaccurate information. This comprehensive taxonomy provides a framework for evaluating and categorizing the capabilities of models within our benchmark.

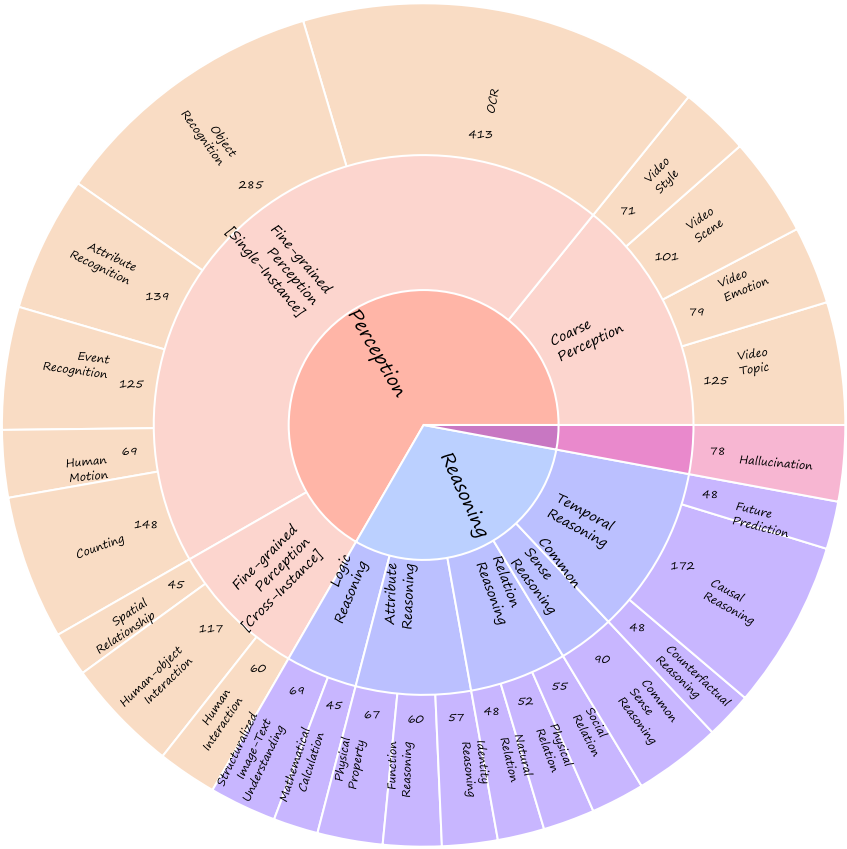


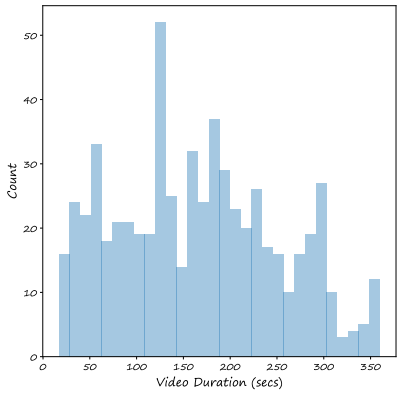
Figure 2. **Overview of ability dimensions in MMBench-Video.** Currently, MMBench-Video incorporates three levels of ability dimensions (from L-1 to L-3), which encompass 26 distinct leaf abilities.

3.3. Question-Answer Pair

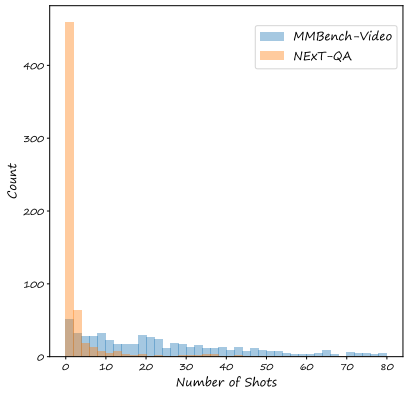
An issue with previous video QA datasets is the omission of temporal information in many questions. In essence, randomly selecting a frame from the video often suffices to answer the question, making the questions static. We define static questions as those in which more than 80% of the frames in the video contain the necessary information to answer it independently. In the case of NeXT-QA, a manual inspection of 100 randomly selected questions from the test set showed that 42% met our definition of static questions. These questions do not sufficiently test the model's temporal understanding capabilities. When developing MMBench-Video questions, we heavily emphasized incorporating temporal information and actively attempted to avoid static questions. However, a small number of static questions remained in our dataset. This issue results from including Coarse Perception questions in our defined ability taxonomy, which inherently evaluates the overall comprehension of the video.

In the MMBench-Video framework, each video is associated with various questions. The questions are independent of each other and have no progressive relationship. Each question corresponds with several distinct ability dimensions. Calculations are based on these ability dimensions when scoring and conducting statistical analyses. For example, a question involving multiple capability dimensions, like Object Recognition and Counting, is considered twice when calculating scores for each corresponding dimension.

In MMBench-Video, the questions and annotated answers have intentionally been designed to be free-form, departing from the fixed characteristics commonly observed in previous datasets, such as multiple-choice questions and standardized question formats such as What/Who/How? The questions have been formulated like conversations, reflecting linguistic diversity and aligning closely with real-world application scenarios. For instance, requests for the score of a football game may be expressed as "What is the score of the football match in the video?" or "Tell me the winning team and the final score." In annotating answers, efforts are made to capture as many details as possible. When answering Yes or No questions, the response includes not only the binary answer but also the reason-

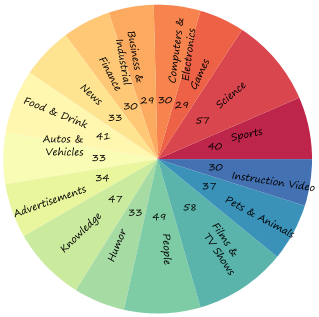


(a) Video duration distribution of MMBench-Video.

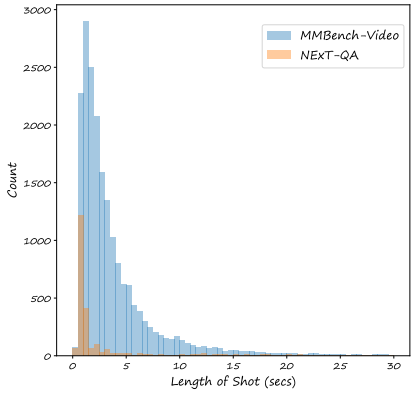


(b) Shot-per-video distribution of MMBench-Video and NExT-QA.

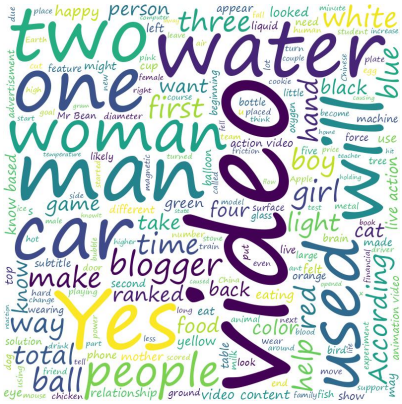
(c) The wordcloud of questions in MMBench-Video.



(d) Video category distribution of MMBench-Video.



(e) Shot duration distribution of shots in MMBench-Video and NExT-QA.



(f) The wordcloud of answers in MMBench-Video.

Figure 3. **Statistics of MMBench-Video.**

ing. We marked as many situations as possible if it is an enumeration question. Adverbs such as "maybe" or "most likely" are used when the answer is uncertain. Moreover, to evaluate the model's Hallucination, we intentionally included a few questions that cannot be answered based on the video content. These questions are annotated like "The answer cannot be known based on the video content." All questions and answers undergo human annotation, ensuring the dataset's quality.

3.4. Dataset Statistics

MMBench-Video contains **610** video clips from 16 major categories (Fig. 3d), ranging from 30 seconds to 6 minutes, as well as **2,256** question-answer pairs. Each QA evaluates one or multiple capabilities of a vision-language model. In Figs. 3c and 3f, we visualize the word clouds of questions and answers in MMBench-Video, respectively. The total length of those videos is **27.7h**, and the average length of video clips is **165 seconds**. We demonstrate the duration distribution of clips in MMBench-Video in Fig. 3a.

One major characteristic of MMBench-Video is that it contains videos with a large number of video shots. We sample the same number of videos from the text split of NExT-QA (a typical video-QA benchmark), split each video into multiple shots, and visualize the shot-per-video number distribution across two benchmarks (Fig. 3b). The average shot number of videos in MMBench-Video is significantly larger than videos in NExT-QA (32.2 *vs.* 4.5). In Fig. 3e, we further visualize the shot duration distribution of both datasets. Besides containing more shots in each video, video shots in MMBench-Video also exhibit longer duration, which makes MMBench-Video a more challenging benchmark.

3.5. Data Collection Pipeline

MMBench-Video is collected in a three-stage pipeline: Finding & Self-Annotation, Peer-Annotation, and Post-processing. We hired about 20 well-educated undergraduate students to do the annotation work.

Finding & Self-Annotation The annotators are given a guideline containing information about target videos and question types. They browse YouTube freely. Once they

come across a video that matches our guidelines, they record its YouTube ID, develop questions about the video, and annotate answers.

Peer-Annotation After Stage 1, annotators have collected enough videos. Now, they start to watch the videos found by others and develop new questions. Through this stage, videos undergo a rigorous cross-validation process, ensuring diverse perspectives and minimizing bias in the dataset. The iterative nature of this stage allows for continuous refinement, resulting in a more comprehensive and well-curated dataset for computer vision research.

Post-processing In this phase, videos are crawled and refined for optimal use. Firstly, if an annotator selects a clip from a larger video, we employ clipping to extract the relevant segment. Additionally, if the annotator thinks subtitles are necessary for answering the question, we embed them into the video.

3.6. Evaluation

The evaluation process involves tackling the challenge of assessing responses to open-ended questions and ensuring alignment with human annotations. This is particularly complex given the inherent subjectivity of human judgment. To automate this intricate process, we leverage ChatGPT powered by GPT-4 [23] for scoring, employing a prompt configured with a 3-point scale (0, 1, 2, 3). This method streamlines the evaluation, providing a quantitative measure of the model's performance compared to human values.

The prompt establishes a standardized approach for evaluating candidate answers. We ensure a robust assessment by explicitly instructing the model to consider semantic similarity and disregard grammatical differences. The 3-point scale employed in our evaluation process serves as a well-defined metric for assessing candidate answers. A rating of 0 signifies no similarity, indicating that the ca answer is entirely wrong. A rating of 1 suggests low similarity, denoting that the answer is largely incorrect. A rating of 2 implies high similarity, indicating that the answer is largely correct. Finally, a rating of 3 signifies semantic equivalence, conveying that the answer is entirely correct. This tiered structure provides a nuanced and granular evaluation, allowing for a more detailed analysis of the model's performance.

Experimental findings reinforce the effectiveness of our scoring system. In comparison with human judgment, the scores exhibit no significant differences, highlighting the reliability of the 3-point scale in capturing semantic similarity. Furthermore, the consistency of scores is evident as the experimental data reveals minimal variation. Specifically, the standard deviation of scores for a given question-answer pair is consistently less than 0.1, affirming the stability and reproducibility of our evaluation framework. This robustness reinforces the credibility of our approach in objectively assessing the model's responses.

This approach ensures that the evaluation process goes beyond simple correctness and considers the subtleties of semantic understanding, providing a more refined and insightful assessment of the model's performance. By instructing the AI assistant to focus on the essence of the response rather than superficial linguistic variations, we aim to enhance the model's capacity to generate accurate and contextually appropriate answers to free-form questions.

4. Experiment

In this section, we present the evaluation result of various technical solutions to multimodal understanding on MMBench-Video.

4.1. Video-based Models on MMBench-Video

We summarize the results of video-based models in Tab. 2. Most video models had an average score of less than 1 out of 3, with the best-performing model scoring only 1.05. These scores indicate that the current video model's proficiency in comprehending MMBench-Video is in the early stages of development. This observation highlights existing challenges and underscores the need for further improvements in video understanding models to increase their ability and effectiveness in understanding diverse video content.

There is an interesting observation from the comparison of mPLUG-Owl [32] and VideoChat [17]. Although mPLUG-Owl achieved the highest scores in each Perception and Reasoning dimension, it lagged in the Hallucination test, resulting in an overall lower score than VideoChat. mPLUG-Owl's strategy of trying to answer every question, even when the video content could not provide a solution, led to this result. In contrast, VideoChat excelled in the hallucination test, achieving a remarkable score of 2.21. It is noteworthy that honesty, where the model admits uncertainty by saying "I don't know", is crucial. This practice prevents the model from providing incorrect information, thereby mitigating the potential negative impact on human users.

4.2. Video-based Models on Image Multimodal Benchmark

A video, fundamentally, is a sequential compilation of images. Consequently, for a model to comprehend a video, it necessitates a foundational understanding of individual images. To evaluate the image understanding capabilities of existing video models, we performed tests on MMBench [19], a benchmark on image multimodal understanding. To evaluate the image dataset with the video model, we employed a technique known as still-inflation, where an image is expanded to a pseudo video clip by simply repeating. In this experiment, an image is transformed into a 1-second video with a frame rate of 30 frames per second (FPS) and a total of 30 frames.

Model	Mean	Mean (w/o HL)	CP	FP-S	FP-C	LR	AR	RR	CSR	TR	HL
VideoChat [17]	**1.05**	0.90	1.19	0.86	0.79	0.44	1.05	1.20	0.69	1.03	**2.21**
Video-ChatGPT [20]	0.75	0.75	0.79	0.71	0.67	0.53	1.01	1.01	0.47	0.82	0.74
Otter-V [16]	0.87	0.77	0.94	0.80	0.72	0.49	0.90	0.91	0.60	0.79	1.67
LLaMA-Adapter [37]	0.73	0.79	0.72	0.71	0.80	0.63	0.70	1.20	0.56	1.02	0.28
mPLUG-Owl [32]	0.97	**1.07**	**1.35**	**1.00**	**0.82**	**0.68**	**1.29**	**1.42**	**0.89**	**1.12**	0.17
MovieChat [25]	0.44	0.43	0.97	0.56	0.41	0.11	0.40	0.61	0.18	0.24	0.46

Table 2. **Evaluation Result of Video Models on MMBench-Video.** CP, FP[S], FP[C] stands for Perception tasks, LR, AR, RR, CSR, TR stand for Reasoning tasks. HL stands for the Hallucination test. All scores are on a 3-point scale. The higher HL score, the less hallucination the model has. Due to the fact that the Hallucination test does not directly assess the video understanding ability of the model (responding with "I don't know" yields full scores), we present two mean scores: one with HL scores and the other excluding it.

Type	Model	Overall	LR	AR	RR	FP-S	FP-C	CP
Image	InternLM-XComposer [36]	**74.4**	**50.6**	**82.0**	**76.1**	**79.3**	59.2	**81.7**
	LLaVA-v1.5-13B [18]	67.0	39.9	74.7	61.6	70.9	**59.9**	75.4
	Qwen-VL-Chat [4]	61.8	40.5	74.3	47.9	66.3	46.2	72.8
	IDEFICS-80B-Instruct [14]	54.6	29.0	67.8	46.5	56.0	48.0	61.9
Video	VideoChat [17]	**27.9**	7.0	**43.8**	**22.3**	**23.6**	12.6	**40.0**
	Video-ChatGPT [20]	16.8	**12.1**	28.5	14.2	14.1	5.7	20.8
	Otter-V [16]	17.9	4.0	33.0	9.5	14.8	5.3	27.0
	LLaMA-Adapter [37]	13.7	11.6	20.5	14.2	9.8	**15.0**	12.8
	mPLUG-Owl [32]	8.9	6.4	12.2	3.8	14.3	3.6	8.4

Table 3. **Comparison of Image Models and Video Models on MMBench.** The abbreviations LR, AR, RR, FP-S, FP-C, CP correspond to Logic Reasoning, Attribute Reasoning, Relation Reasoning, Fine-Grained Perception [Single-Instance], Fine-Grained Perception [Cross-Instance], and Coarse Perception, respectively, as defined in [19]. MMBench uses CircularEval method on multiple choice questions. The reported numbers represent the percentage of correct answers (out of 100) for each ability.

We present the evaluation results in Tab. 3. The findings demonstrate a significant image comprehension performance gap between video-based and image-based models. Notably, VideoChat is the highest-performing video-based model on MMBench, which is consistent with its superior performance on MMBench-Video. This substantiates our hypothesis that foundational competence in video understanding inherently stems from proficient image comprehension.

4.3. Image-based Models for Video Understanding

Recognizing the limitations of existing video models and the impressive performance of image-based models, we considered using image models directly for video understanding. To apply image-based models to video understanding, we investigate two settings, namely 1-image and 9-image.

Under the 1-image setting, we randomly pick a single frame from the video, depending solely on that image to respond to the questions. Although there is a natural loss of in-depth information, this approach is skillful at catching the primary global aspects of the video, which allows for sufficient information gathering to address the relevant questions.

The 9-image setting involves uniformly sampling nine frames from the video and arranging them in a 3×3 grid to create a composite image. The model is explicitly instructed on the temporal sequencing of these frames, thereby enhancing its ability to comprehend the video's temporal dynamics. This precise experimental design evaluates the effectiveness of utilizing image models for temporal modeling in video comprehension.

The results are presented in Tab. 4. The 9-image setting demonstrates superior performance on average relative to video models. As expected, scores on nine images exceed those on one image for most image models. Notably, some image models perform better than video models, even with only one image. These findings support our hypothesis that existing video approaches inadequately capture temporal information. The apparent differences in performance highlight the need for better techniques in modeling the time-sensitive details of video comprehension.

4.4. LLM-based Approaches on MMBench-Video

Remarkable advancements have been made in large language models, encouraging further investigation into their

Method	Model	Mean	Mean (w/o HL)	CP	FP-S	FP-C	LR	AR	RR	CSR	TR	HL
9-image	VideoChat [17]	**1.07**	**0.97**	1.16	0.93	**0.93**	0.45	**1.19**	1.43	0.73	**0.94**	**1.88**
	Otter-I [16]	0.85	0.82	1.00	0.82	0.75	0.38	0.96	0.99	**0.93**	0.71	1.10
	InternLM-XComposer [36]	0.76	0.81	1.13	0.78	0.67	**0.47**	1.05	1.04	0.62	0.70	0.40
	Qwen-VL-Chat [4]	1.00	**0.97**	**1.35**	**1.02**	0.83	0.61	1.14	1.13	0.84	0.85	1.26
	Idefics_9B_Instruct [14]	0.90	0.95	1.20	0.87	0.78	0.44	1.14	**1.45**	0.91	0.81	0.49
1-image	VideoChat [17]	**1.03**	0.90	1.14	0.85	0.77	0.50	1.01	1.15	0.70	**1.11**	**2.06**
	Otter-I [16]	0.77	0.77	0.98	0.79	0.66	0.46	0.89	0.91	0.71	0.76	0.76
	InternLM-XComposer [36]	0.72	0.76	1.05	0.82	0.55	0.47	0.86	0.99	0.59	0.71	0.45
	Qwen-VL-Chat [4]	0.91	0.84	1.12	0.83	0.53	0.52	1.01	0.97	0.83	0.90	1.47
	Idefics_9B_Instruct [14]	0.96	**1.02**	**1.17**	**0.99**	**0.88**	**0.68**	**1.09**	**1.46**	**0.86**	1.06	0.49

Table 4. **Evaluation Result of Image Models on MMBench-Video.**

Method	Model	Mean	Mean (w/o HL)	CP	FP-S	FP-C	LR	AR	RR	CSR	TR	HL
Question	GPT-4* [23]	0.84	0.63	0.19	0.25	0.29	0.59	0.88	1.05	0.58	1.22	**2.50**
	GPT-4 [23]	0.88	0.69	0.22	0.23	0.37	0.67	0.86	1.17	0.55	1.43	2.42
Subtitle	GPT-4 [23]	1.34	1.25	1.38	0.98	1.01	0.60	**1.72**	**1.76**	0.86	**1.65**	2.10
Localization +Video	GPT-4 [23] + VideoChat [17]	**1.54**	**1.46**	**1.62**	**1.48**	**1.46**	1.20	1.65	1.49	**1.19**	1.60	2.19
	GPT-4 [23] + Otter [16]	1.39	1.38	1.45	1.31	1.19	**1.37**	1.53	1.60	1.04	1.57	1.47

Table 5. **Evaluation Result of Video Models on MMBench-Video.** *The first row of data was tested on all MMBench-Video data, while the other data was tested only on the subset of MMBench-Video with subtitles.

potential for describing video content. In theory, all video information can be encoded in text. Therefore, it is enough to answer all video-related questions from the text. However, the length of such text may exceed the context window of language models. Therefore, accurately summarizing videos via text presents a significant research challenge. Fortunately, subtitles provided by YouTube in MMBench-Video present a promising opportunity. Since the videos typically contain narrators delivering information, we employed GPT-4 [23] to analyze YouTube subtitles, examining the feasibility of comprehending videos through a text-based method. We attempted three methods. The initial approach was to input the question directly into the LLM without any context from the video. The second method was to input both the video subtitles and questions into the LLM, enabling it to answer according to the subtitles. The third method utilizes an LLM to identify the video interval through subtitles and then employs a video model to analyze this segment and answer the question.

The results are presented in Tab. 5. Inputting only questions results in a low score, which is reasonable since no information about the content of the video is given. However, the score is better than 0. There are two reasons for this. First, we score based on the partial correctness of the answer. The model may be able to guess part of the answer through the questions, such as Yes or No questions. Additionally, we found that it scores higher on reasoning questions. Some videos in MMBench-Video provide factual information, such as physical laws. These videos contain questions that a human expert could potentially answer without watching the video. With knowledge encoded in existing language models, it is possible for LLMs to achieve the same. The subtitle method scores significantly higher than video models. Localization + video model further improved the score. This shows that the current video model has inadequate temporal retrieval capabilities, resulting in challenges when processing long videos in MMBench-Video.

5. Future Work

In the future, we aim to expand our dataset by collecting more data. Additionally, we intend to explore more cost-effective annotation methods that minimize manual labor while upholding annotation quality. Presently, we have a test set, and our goal is to contribute a training set to the community, addressing the shortage of video training data. Ultimately, our goal is to provide detailed guidelines that can direct the enhancement of models for more advanced video reasoning abilities.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark, 2016. 3

[2] 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. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3

[3] 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 arXiv:2308.01390*, 2023. 3

[4] 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 arXiv:2308.12966*, 2023. 7, 8

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems (NeuIPS)*, 33:1877–1901, 2020. 3

[6] Brandon Castellano. Pyscenedetect. *Last accessed*, 2020. 3

[7] Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan C. Stroud, and Rada Mihalcea. Fiber: Fill-in-the-blanks as a challenging video understanding evaluation framework, 2022. 3

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 3

[9] Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1675–1685, 2021. 3

[10] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 3

[11] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020. 3

[12] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning, 2021. 3

[13] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video Question Answering with Spatio-Temporal Reasoning. *IJCV*, 2019. 3

[14] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. 7, 8

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 3

[16] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 3, 7, 8

[17] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3, 6, 7, 8

[18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 7

[19] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2023. 3, 6, 7

[20] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 7

[21] Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*, 2023. 3

[22] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875, 2017. 3

[23] OpenAI. Gpt-4 technical report, 2023. 2, 6, 8

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[25] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding, 2023. 7

[26] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 3

[27] 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 arXiv:2302.13971*, 2023. 3

[28] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3

[29] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 1, 3

[30] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, 2022. 3

[31] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3480–3491, 2022. 3

[32] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023. 3, 6, 7

[33] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020. 3

[34] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019. 3

[35] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3

[36] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023. 7, 8

[37] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 7