MMWorld: Towards Multi-discipline Multi-faceted World Model Evaluation in Videos

Xuehai He[1] Weixi Feng[*2] Kaizhi Zheng[*1] Yujie Lu[*2] Wanrong Zhu[*2] Jiachen Li[*2]
Yue Fan[*1] Jianfeng Wang[3] Linjie Li[3] Zhengyuan Yang[3] Kevin Lin[3]
William Yang Wang[2] Lijuan Wang[3] Xin Eric Wang[1]
[1]UC Santa Cruz [2]UC Santa Barbara [3]Microsoft
{xhe89,xwang366}@ucsc.edu

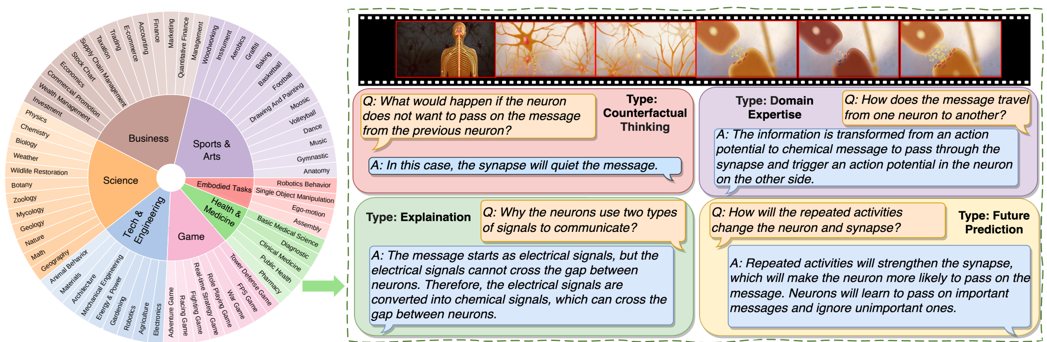


Figure 1: MMWorld covers seven broad disciplines and 69 subdisciplines, focusing on the evaluation of multi-faceted reasoning beyond perception (e.g., explanation, counterfactual thinking, future prediction, domain expertise). On the right is a video sample from the Health & Medicine discipline.

Abstract

Multimodal Language Language Models (MLLMs) demonstrate the emerging abilities of "world models"—interpreting and reasoning about complex real-world dynamics. To assess these abilities, we posit videos are the ideal medium, as they encapsulate rich representations of real-world dynamics and causalities. To this end, we introduce MMWorld, a new benchmark for multi-discipline, multi-faceted multimodal video understanding. MMWorld distinguishes itself from previous video understanding benchmarks with two unique advantages: (1) **multi-discipline**, covering various disciplines that often require domain expertise for comprehensive understanding; (2) **multi-faceted reasoning**, including explanation, counterfactual thinking, future prediction, etc. MMWorld consists of a human-annotated dataset to evaluate MLLMs with questions about the whole videos and a synthetic dataset to analyze MLLMs within a single modality of perception. Together, MMWorld encompasses 1,910 videos across seven broad disciplines and 69 subdisciplines, complete with 6,627 question-answer pairs and associated captions. The evaluation includes 2 proprietary and 10 open-source MLLMs, which struggle on MMWorld (e.g., GPT-4V performs the best with only 52.3% accuracy), showing large room for improvement. Further ablation studies reveal other interesting findings such as models' different skill sets from humans. We hope MMWorld can serve as an essential step towards world model evaluation in videos.

[*]Equal Contribution

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

1 Introduction

Foundation models, such as Large Language Models (LLMs) [49; 59; 26; 2] and Multimodal LLMs (MLLMs) [51; 58; 36; 33; 45; 10], have demonstrated remarkable abilities in text and image domains, igniting debates about their potential pathways to Artificial General Intelligence (AGI). This raises a critical question: how well do these models understand the dynamics of the real world? Are they equipped with an inherent World Model [28; 11; 21; 65] that can understand and reason about the underlying principles and causalities of the dynamic, multimodal world?

Videos, with their rich, dynamic portrayal of the real world, are ideally suited for evaluating the "world modeling" capabilities of MLLMs. Existing video understanding benchmarks [34; 47; 53; 34], however, fall short in two key perspectives for such evaluations. First, as LeCun et al. [28] discussed, the world model should be able to *(1) estimate missing information about the state of the world not provided by perception, and (2) predict plausible future states of the world*. Evaluation of such capabilities requires **multi-faceted reasoning** beyond perception level, including explaining the video dynamics, counterfactual thinking of alternative consequences, and predicting future activities within videos. Moreover, the **multi-discipline** nature of the multimodal world necessitates a grasp of diverse fundamental principles—ranging from physics and chemistry to engineering and business. Hence, domain expertise across a variety of disciplines is imperative for a thorough evaluation of a model's world understanding towards AGI [46; 73].

Therefore, we introduce MMWorld, a multi-discipline multi-faceted multimodal video understanding benchmark for a comprehensive evaluation of MLLMs[2]. MMWorld encompasses a wide range of disciplines and presents multi-faceted reasoning challenges that demand a combination of visual, auditory, and temporal understanding. It consists of 1,910 videos that span seven common disciplines, including *Art & Sports*, *Business*, *Science*, *Health & Medicine*, *Embodied Tasks*, *Tech & Engineering*, and *Games*, and 69 subdisciplines (see Figure 1) such as Robotics, Chemistry, Trading, and Agriculture, thereby fulfilling the objective of breadth in discipline coverage. The dataset includes a total of 1,559 question-answer pairs and captions annotated and reviewed by humans. Meanwhile, for multi-faceted reasoning, MMWorld mainly contains seven kinds of questions focusing on *explanation* (explaining the phenomenon in videos), *counterfactual thinking* (answering what-if questions), *future prediction* (predicting future events), *domain expertise* (answering domain-specific inquiries), *temporal understanding* (reasoning about temporal information), and etc. A video example with these four questions from the Health & Medicine discipline is depicted in Figure 1. MMWorld comprises two datasets: a human-annotated dataset for evaluating MLLMs on the whole video and a synthetic dataset designed to analyze MLLMs' perception within single visual or audio modalities. We evaluate 12 MLLMs that can handle videos or image sequences on MMWorld, including both open-source (e.g., Video-LLaVA-7B [36]) and proprietary models (GPT-4V [51] and Gemini [58]).

We summarized the contributions and key findings as follows:

- We introduce MMWorld, a new benchmark designed to rigorously evaluate the capabilities of Multimodal Large Language Models (MLLMs) in world modeling through the realm of video understanding. MMWorld spans a broad spectrum of disciplines, featuring a rich array of question types for multi-faceted reasoning.
- In addition to the human-annotated dataset, we develop an automatic data collection pipeline, streamlining video content selection and question-answer generation, and construct a well-controlled synthetic dataset to analyze MLLMs within single visual or audio modalities.
- We observe that existing MLLMs still face substantial challenges posed by MMWorld. Even the best performer, GPT-4V, can only achieve a 52.30% overall accuracy, and four MLLMs particularly trained on videos perform worse than random chance.
- Although there is stll a clear gap between open-source and proprietary models, the best open-source model Video-LLaVA-7B outperforms GPT-4V and Gemini on Embodied Tasks by a large margin

[2]Note that MMWorld is not a sufficient testbed for world model evaluation, but we believe overcoming the unique challenges presented in MMWorld is essential and necessary towards comprehensive world modeling.

Table 1: Comparison between MMWorld and previous benchmarks for real-world video understanding on a variety of criteria. Multi-faced include Explanation (`Explain.`), Counterfactual Thinking (`Count.`), Future Prediction (`Future.`) and Domain Expertise (`Domain.`) MMWorld is the first multi-discipline and multitask video understanding benchmark that covers wider reasoning questions, and also included first-party data annotations.

Benchmarks	**Multi-Discipline**	**Multi-Task**	**Multi-Faceted Reasoning**				**First-Party Annotation**
			`Explain.`	`Count.`	`Future.`	`Domain.`	
MovieQA [57], TVQA [29]			✓				✓
ActivityNet-QA [71]							✓
MSVD-QA [66], MSRVTT-QA [67]							✓
Sports-QA [31]				✓		✓	✓
VaTeX [61]		✓					✓
VALUE [35]		✓					
Video-Bench [48]		✓			✓	✓	
MVBench [34]		✓		✓	✓		
Perception Test [53]		✓	✓	✓	✓		
MMWorld (Ours)	✓	✓	✓	✓	✓	✓	✓

and performs similarly on Art & Sports, where spatiotemporal dynamics play a more crucial role in video understanding. This is further validated with its leading results on the Temporal Understanding question type.

- In our study comparing MLLMs with average humans (non-experts), we notice some correlation between question difficulties as perceived by humans and MLLMs. However, MLLMs present different skill sets than humans in that they can answer reasonable amount of difficult questions that humans completely fail but also struggle at easy questions that humans excel at. This indicates different perception, cognition, and reasoning abilities between MLLMs and humans.

2 Related Work

2.1 Multimodal Large Language Models (MLLMs)

Emerging MLLMs With recent breakthroughs [50; 18; 59; 12; 60; 4] in Large Language Models (LLMs), several counterparts in the vision-and-language domain have been proposed [14; 41; 40; 30; 78; 77; 5], and recently released GPT-4V [51], followed by Gemini Vision family [58]. Many MLLMs have expanded their capabilities beyond handling only text and image inputs. VideoChat [33] leverages the QFormer [32] to map visual representations to LLM [12], and performs a multi-stage training pipeline. Otter [30] proposes to conduct instruction finetuning based on Openflamingo [3]. PandaGPT [56] employs the ImageBind [23] as the backbone and finetunes it. mPLUG-Owl [68] introduces an abstractor module to perform visual and language alignment. VideoLLaMA [75] introduces a frame embedding layer and also leverages ImageBind to inject temporal and audio information into the LLM backend. Chat-UniVi [27] uses clustering to do feature fusion. Observing their emerging abilities in multimodal video understanding, we propose MMWorld to evaluate these models' skills in understanding the dynamics of the real world.

Benchmarking MLLMs To evaluate MLLMs, there is a flourishing of analysis [38; 76; 43; 15; 13; 20; 70; 16] and the establishment of innovative benchmarks such as VisIB-Bench [8] which evaluates models with real-world instruction-following ability given image inputs, MMMU [73] designed to access models on college-level image-question pairs that span among different disciplines, and VIM [44] which challenges the model's visual instruction following capability. However, these recent analyses and benchmarks only cover the image input, which hinders the evaluation of MLLM's performance as a world model. Recently, video benchmarks such as Perception Test [53] is proposed to focus on perception and skills like memory and abstraction. However, it uses scenarios with a few objects manipulated by a person, which limits the variety of contexts. MVBench [34] centers on temporal understanding, while MMWorld not only includes temporal reasoning but also evaluates other multi-faceted reasoning abilities.

2.2 Video Understanding Benchmarks

Previous video benchmarks, as shown in Table 1, focus on video understanding tasks, including activity-focused on web videos [72], description-based question answering [74], video completion [17], and video infilling [24]. Recently, Video-Bench [47] introduces a benchmark by collecting videos and annotations from multiple existing datasets. LWM [39] collects a large video and language dataset from public books and video datasets and trains a world model that is capable of processing more than millions of tokens. However, modeling millions of tokens is extremely difficult due to high memory cost, computational complexity, and lack of suitable datasets. Mementos [62] builds a benchmark for MLLM reasoning for input image sequences. STAR [64] builds a benchmark for situated reasoning in real-world videos. CLEVER [69] builds a benchmark containing videos focusing on objects with simple visual appearance. Our contribution, in contrast, presents a new video understanding benchmark designed to evaluate models on several pivotal components crucial for a comprehensive world model. These components encompass interdisciplinary coverage, task diversity, and multifaceted reasoning capabilities—including future prediction, counterfactual thinking, and more—underpinned by original human annotations and integrated domain knowledge.

3 The MMWorld Benchmark

The MMWorld benchmark is built on three key design principles: multi-discipline coverage and multi-faceted reasoning. It spans various disciplines that require domain expertise and incorporates diverse reasoning skills such as explanation, counterfactual thinking, and future prediction. The benchmark consists of two parts: a human-annotated dataset and a synthetic dataset. The human-annotated dataset serves as the main test bed to evaluate MLLMs from multiple perspectives. The synthetic dataset contains two subsets, focusing on evaluating MLLMs' perception behavior from both visual signals and audio inputs, respectively.

3.1 Manual Data Collection

We collect videos from YouTube with the Creative Licence in seven disciplines: Art & Sports (18.5%), Business (12.0%), Science (20.4%), Health & Medicine (12.0%), Embodied Tasks (12.0%%), Tech & Engineering (12.9%), and Game (12.2%). For Art & Sports, 29 videos are collected from the SportsQA dataset [31]. And for Embodied Tasks, 24 videos are sourced from IKEA Assembly [7], RT-1 [9], and Ego4D [19] datasets to increase video diversity.

Our manual benchmark collection takes two stages. In the first stage, we conduct a detailed examination of each of the seven primary disciplines to identify a comprehensive range of subdisciplines for inclusion in our benchmark. Our selection of videos is driven by two key principles:

The **first principle**, **multi-discipline** coverage, emphasizes the requirement for domain knowledge—selecting videos that inherently demand an understanding of specialized content across various disciplines. The **second principle**, **multi-faceted** annotation, involves collecting videos that enable the creation of question-answer pairs from multiple perspectives to evaluate world model properties comprehensively. The **third principle**, **temporal information**, prioritizes the inclusion of videos that provide meaningful content over time, as understanding temporal information is crucial for grasping world dynamics. This allows models to engage in temporal reasoning. Therefore, answering questions in our dataset requires implicit temporal reasoning, e.g., the model needs to understand temporal information to explain "why does the robot need to do the step shown in the video". We also design a "temporal understanding" question type to explicitly test models' ability to reason about temporal information (examples can be found in Section F in the Appendix).

During the second stage, our team embark on the task of question annotation. We craft questions that primarily test seven aspects of multimodal video understanding also from the perspective of **multi-faceted reasoning**: 1) Explanation: Questions ask the model to elucidate the underlying logic or purpose within the video; 2) Counterfactual Thinking: Tests the model's ability to hypothesize and consider alternative outcomes; 3) Future Prediction: Aims to predict future events based on the current

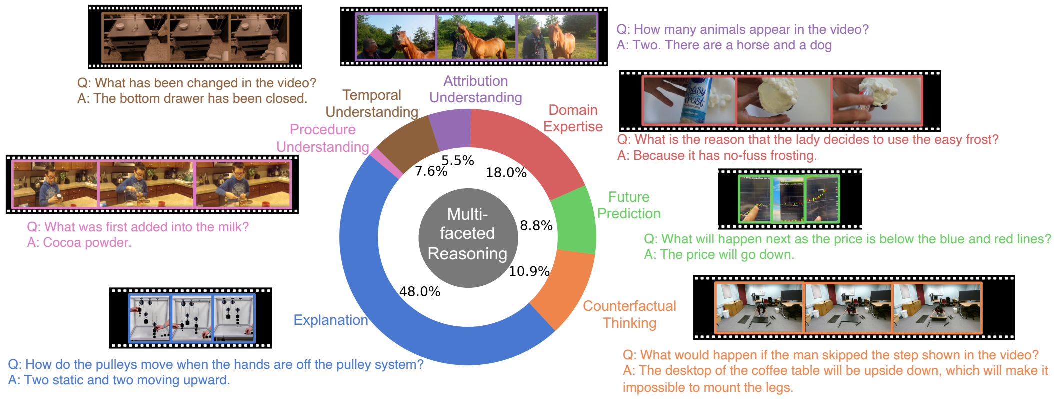


Figure 2: The questions in MMWorld primarily evaluate seven understanding and reasoning abilities of models to provide correct answers.

scenario, challenging the model's foresight; 4) Domain Expertise: Evaluates the model's depth of knowledge in specific fields, such as how to assemble a coffee table; 5) Temporal Understanding: Assesses the model's capability to reason about temporal sequences and dynamics; 6) Attribution Understanding: These questions focus on identifying cause-and-effect relationships within the video, including tasks like counting; 7) Procedure Understanding: Tests the model's ability to comprehend and explain procedural tasks shown in the video. The detailed distribution and examples are shown in Figure 2.

3.2 Automated Data Collection

Table 2: Key Statistics of the MMWorld Benchmark. The main subset is the human-annotated subset. Synthetic Subset I contains generated QA pairs focused exclusively on the audio content, while Synthetic Subset II contains QA pairs focused exclusively on the visual content of the video.

Statistics	**Main Subset**	**Synthetic I**	**Synthetic II**
#Discipline/#Subdiscipline	7/61	7/51	7/54
#<Video-QA>	<417-1,559>	<746-2,969>	<747-2,099>
Avg Video Lengths (s)	102.3	103.4	115.8
Avg #Questions per Video	4.05	3.98	2.81
Avg #Options	3.90	4.00	4.00
Avg Question Length	11.39	15.12	17.56
Avg Option Length	7.27	6.01	5.19
Avg Answer Length	6.42	6.71	5.67
Avg Caption Length	27.00	71.87	82.33

Understanding real-world dynamics requires models to process both audio and visual modalities. To evaluate MLLMs' perception abilities in these modalities, we designed an automated data collection pipeline. This pipeline collects targeted videos and generates QA pairs based on either audio or visual information, ensuring the model's capabilities are assessed independently for each modality. By using a single modality's information for generation, our pipeline ensures that the synthetic data remains unbiased regarding input modality.

The synthetic data generation pipeline is illustrated in Figure 3. We employ a systematic approach to gather videos with Creative Commons licenses from YouTube and the extensive YouTube-8M dataset [1]. This method ensures a diverse and comprehensive collection of video data, which is important for the robust evaluation of multimodal video understanding models.

Video Collection and Processing We start with the video *Query Generator*. We start with the same seven disciplines as the manually collected dataset. For each discipline, a set of subdisciplines is defined to encapsulate a wide spectrum of topics, ensuring a diverse and comprehensive dataset. Once the queries are generated, the *Video Mapping and Filtering* step is initiated. We perform mapping of videos to YouTube-8M and online videos, constrained by a strict time limit of two minutes per query, keeping only the most pertinent videos that satisfy the predefined criteria. Simultaneously, the works in conjunction with the video transcripts to extract key terms and concepts. This iterative process refines the search parameters and enhances the semantic richness of the dataset by identifying and encoding the salient themes present in the videos. The *Video Summarization* module utilizes Query-

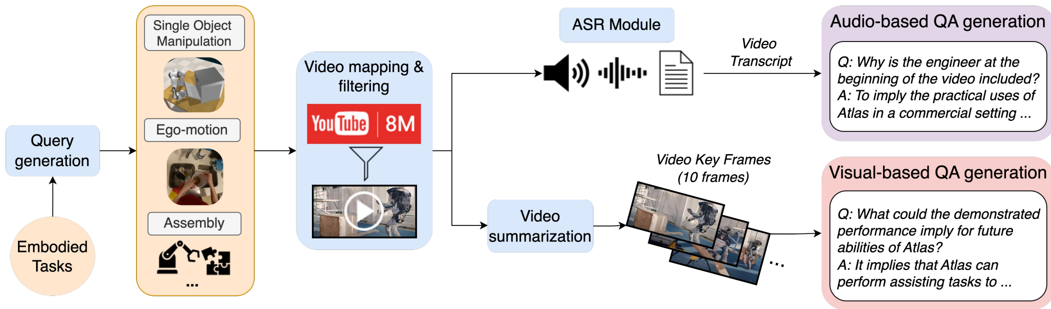


Figure 3: Schematic diagram of the synthetic data generation pipeline in MMWorld. It starts with generating subdiscipline-specific queries, followed by video retrieval from YouTube-8M [1] and YouTube. Keyframes are extracted for visual-based QA generation, and videos are transcribed using an ASR module for audio-based QA generation.

focused video summarization techniques based on Katna[3] and UniVTG [37]. This module selects ten representative frames from each video, distilling the essence of the content while preserving the narrative context. This summarization facilitates efficient storage and quicker processing times, which are crucial for large-scale analysis.

QA Generation The final stage in our pipeline is the *QA / Caption Generation* module, where we leverage the capabilities of GPT-4V to generate accurate and contextually relevant questions and answers, as well as captions, based on the video frames and transcripts. This step not only provides rich annotations for each video but also equips the dataset with a multimodal dimension that supports various downstream tasks such as video QA, captioning, and more.

Quality of the Synthetic Dataset Human evaluators were engaged to ascertain the reasonableness of automatically generated questions and answers, ensuring that the synthetic dataset maintains a high standard of quality and relevance. The findings from this human evaluation phase are detailed in Section 3 of the Appendix, offering insights into the dataset's efficacy and the realism of its constructed queries and responses.

Finally, the statistics of automated curated data, which is used for the ablation study, are shown in Table 2. The taxonomy of our dataset is shown in Figure 1. We note that only a portion of the subdisciplines are shown due to space concerns. Please refer to the Appendix for full information.

4 Experiments

4.1 Experimental Settings

In our study, we compare MLLM's performance on the MMWorld benchmark, including GPT-4V [51], Gemini Pro [58], Video-Chat [33], Video-LLaMA [75], ChatUnivi [27], mPLUG-Owl [68], Otter [30], ImageBind-LLM [23], PandaGPT [56], LWM [39], and X-Instruct-BLIP [52]. For both Gemini Pro and GPT-4V, we adhere to the default settings provided by their official APIs. They both take ten image frames extracted from the video content as the input. The Gemini Pro is set to process visual input and configured with safety settings to filter a range of harmful content. The configuration thresholds are set to 'BLOCK_NONE'. For PandaGPT, we set 'top_p' to 0.7 and 'temperature' to 0.5. For VideoChat, we set 'max_frames' to 100. For X-Instruct-BLIP, the model is implemented using four image frames. We use GPT-4-32K as the judge for judging whether the model answer is correct when it can not mapped to the option letter using the rule-based method. For others, we all use the default setting. All inferences are run on a NVIDIA A6000 workstation. The detailed implementation is given in the Appendix.

[3] `https://github.com/keplerlab/katna`

Table 3: MLLM accuracy across diverse disciplines (averaging over three runs). GPT-4V and Gemini Pro lead at most disciplines and achieve the best overall accuracy. The best open-source model Video-LLaVA-7B outperforms them on Embodied Tasks and perform similarly on Art & Sports.

Model	Art& Sports	Business	Science	Health& Medicine	Embodied Tasks	Tech& Engineering	Game	Average
Random Choice	25.03	25.09	26.44	25.00	26.48	30.92	25.23	26.31
Proprietary MLLMs								
GPT-4V [51]	36.17 ±0.58	**81.59** ±1.74	**66.52** ±1.86	73.61 ±0.49	55.48 ±2.70	61.35 ±1.00	**73.49** ±1.97	**52.30** ±0.49
Gemini Pro [58]	37.12 ±2.68	76.69 ±2.16	62.81 ±1.83	**76.74** ±1.30	43.59 ±0.33	**69.86** ±2.01	66.27 ±2.60	51.02 ±1.35
Open-source MLLMs								
Video-LLaVA-7B [36]	35.91 ±0.96	51.28 ±0.87	56.30 ±0.76	32.64 ±0.49	**63.17** ±1.44	58.16 ±1.00	49.00 ±3.16	44.60 ±0.58
Video-Chat-7B [33]	**39.53** ±0.06	51.05 ±0.00	30.81 ±0.21	46.18 ±0.49	40.56 ±0.57	39.36 ±0.00	44.98 ±0.57	40.11 ±0.06
ChatUnivi-7B [27]	24.47 ±0.49	60.84 ±1.51	52.00 ±0.73	61.11 ±1.96	46.15 ±2.06	56.74 ±1.33	52.61 ±2.84	39.47 ±0.42
mPLUG-Owl-7B [68]	29.16 ±1.62	64.10 ±1.84	47.41 ±3.29	60.07 ±1.30	23.78 ±3.47	41.84 ±5.09	62.25 ±3.16	38.94 ±1.52
PandaGPT-7B [56]	25.33 ±0.54	42.66 ±3.02	39.41 ±2.67	38.54 ±3.07	35.43 ±0.87	41.84 ±2.79	40.16 ±4.65	32.48 ±0.45
ImageBind-LLM-7B [23]	24.82 ±0.16	42.66 ±0.99	32.15 ±1.11	30.21 ±1.47	46.85 ±1.14	41.49 ±1.50	41.37 ±0.57	31.75 ±0.14
X-Instruct-BLIP-7B [52]	21.08 ±0.27	15.85 ±0.87	22.52 ±1.11	28.47 ±0.49	18.41 ±1.44	22.34 ±0.87	26.10 ±0.57	21.36 ±0.18
LWM-1M-JAX [39]	12.04 ±0.53	17.48 ±0.57	15.41 ±0.91	20.49 ±0.98	25.87 ±1.98	21.99 ±2.19	11.65 ±3.01	15.39 ±0.32
Otter-7B [30]	17.12 ±1.17	18.65 ±0.87	9.33 ±0.36	6.94 ±0.98	13.29 ±1.51	15.96 ±1.74	15.26 ±0.57	14.99 ±0.77
Video-LLaMA-2-13B [75]	6.15 ±0.44	21.21 ±0.66	22.22 ±1.45	31.25 ±1.70	15.38 ±1.14	19.15 ±1.74	24.90 ±5.93	14.03 ±0.29

4.2 Evaluation

Our dataset includes multiple-choice questions and captions corresponding to each video, enabling tasks such as video question answering and video captioning. We focus on video question answering by evaluating a model's performance based on its accuracy in selecting the correct answer from the provided options. One challenge lies in reliably parsing the model's response to map it to one of the predefined choices. To address this, we employ two mapping strategies. We employ two mapping strategies. The first method employs automated scripts to parse the models' predictions and compare the parsed results with the ground truth, similar to the approach used in [73]. The second method involves models freely generating answers, which are then evaluated by GPT-4. Given the question, correct answer, and model's prediction, GPT-4 returns a True or False judgment. This approach is based on recent works in model evaluation [45; 25; 22; 42]. We validated this method with human evaluators, showing an error rate of 4.76% across 189 examples, confirming the effectiveness of GPT-4 as an evaluator. Detailed results for human evaluation and for these two different strategies are provided in Appendix B. In the main paper, all results are evaluated using the second approach.

4.3 Main Evaluation Results

We show in Table 3 the main evaluation results of different MLLMs. Among these, GPT-4V emerges as the top performer, closely followed by Gemini Pro. Video-LLaVA also demonstrates strong results, primarily due to the extensive training data which consists of 558K LAION-CCSBU image-text pairs and 702K video-text pairs from WebVid [6]. For instruction tuning, datasets were gathered from two sources: a 665K image-text instruction dataset from LLaVA v1.5 and a 100K video-text instruction dataset from Video-ChatGPT [45]. This superior performance may also be attributed to Video-LLaVA's adoption of CLIP ViT-L/14 trained in LanguageBind [36] as its vision model and the inclusion of a large volume of image-video-text pairings within the training data. On the other hand, models like Otter and LWM perform poorly across most disciplines, possibly due to their weaker backbone and architecture used. Otter uses the LLaMA-7B language encoder and a CLIP ViT-L/14 vision encoder, both of which are frozen, with only the Perceiver resampler module fine-tuned, which may contribute to its lower performance. Additionally, some MLLMs perform even worse than random, highlighting the challenging nature of MMWorld.

4.4 Study on Multi-faceted Reasoning on MMWorld

Figure 4 illustrates the multi-faceted reasoning performance for each MLLM. GPT-4V emerges as the strongest model across Future Prediction, Domain Expertise, and Attribution Understanding.

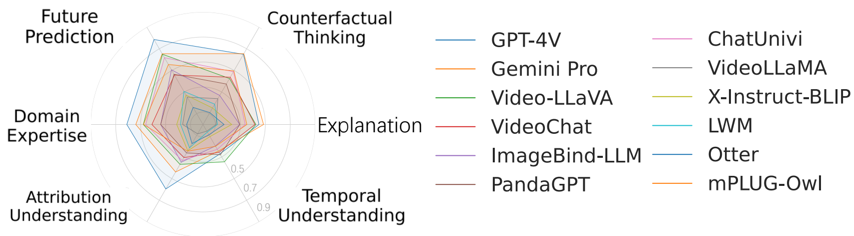
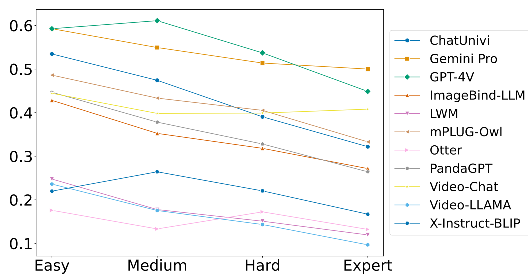
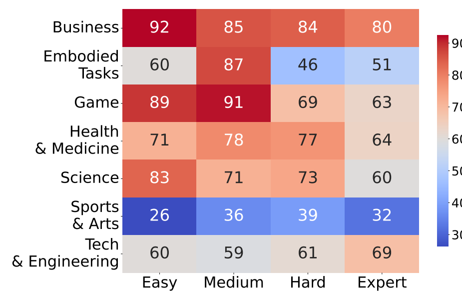


Figure 4: Results of different MLLMs on multi-faceted reasoning. The detailed performance numbers can be found in the Appendix.



(a) Accuracy of MLLMs at difficulty levels.



(b) GPT-4V results by disciplines at difficulty levels.

Figure 5: Model performance at different difficulty levels for average humans. Average human difficulty levels are defined by 3 turkers' performance per question: Easy (3/3 correct answers), medium (2/3 correct), hard (1/3 correct), and expert (0/3 correct).

Closed-source models like GPT-4V and Gemini Pro perform similarly on counterfactual thinking and outperform all others. However, for temporal understanding, Video-LLaVA performs the best. This may be due to its extensive training on large amounts of video-language data, which enhances its spatio-temporal reasoning abilities. This can be also observed in its high scores on the Art & Sports and Embodied Tasks, which involve dense spatio-temporal information, as shown in Table 3. Video-LLaVA's performance is comparable to GPT-4V and Gemini on explanation tasks, likely because of its two-stage training process and exposure to a large amount of instruction-tuning data in the second stage, which includes similar instructions.

4.5 Study on MLLM Performance at Different Difficulty Levels for Average Humans

Figure 5a indicate some correlation between the difficulty levels as perceived by humans and the performance of MLLMs. MLLMs generally follow a trend where accuracy decreases as the difficulty level increases, which aligns with human performance patterns. However, the correlation is not perfect, suggesting that while models and humans share some common ground in understanding question difficulty, there are also notable differences in their capabilities. The data reveals that MLLMs exhibit different skill sets compared to humans. As highlighted in Figure 5b, models like GPT-4V can correctly answer expert-level questions that humans often get wrong, particularly in disciplines such as Business and Health & Medicine, where humans often struggle, yet they sometimes falter on easier questions, likely due to the lack of contextual understanding. Notably, discrepancies in disciplines like Art & Sports and Tech & Engineering highlight areas where MLLMs' performance does not align with human results, suggesting different perception, cognition, and reasoning abilities in handling abstract concepts. These differences suggest that MLLMs can complement human capabilities, offering potential for enhanced task performance by combining the data-driven insights of models with human intuition and contextual knowledge.

Table 4: Performance on Synthetic Subsets I (Audio) and II (Visual). Synthetic Subset I contains QAs based solely on the audio content, while Synthetic Subset II focuses exclusively on the visual content of the video. We evaluated four MLLMs that can process both audio and visual inputs along with Gemini Pro (for the audio setting, only providing the question).

Model	Art&Sports		Business		Science		Health&Medicine		Embodied Tasks		Tech&Engineering		Game		Average	
	Audio	Visual	Audio	Visual	Audio	Visual	Audio	Visual	Audio	Visual	Audio	Visual	Audio	Visual	Audio	Visual
Random Choice	31.59	30.14	31.18	26.58	36.98	32.89	38.74	32.64	32.81	31.25	27.23	32.60	32.01	30.78	32.44	30.91
Video-Chat [33]	**33.98**	32.48	**46.47**	41.46	**41.86**	39.15	**45.95**	36.81	32.81	46.88	**37.48**	35.91	**32.98**	46.70	**38.82**	39.07
ChatUnivi [27]	30.03	43.22	30.19	52.85	38.75	54.59	34.76	50.69	20.14	40.63	24.17	46.41	29.98	45.44	31.82	48.44
Video-LLaMA [75]	30.15	30.23	36.18	33.17	31.33	31.34	30.90	32.78	**33.13**	30.05	31.18	30.55	20.49	27.20	29.08	30.47
Otter [30]	14.22	16.82	16.77	14.24	16.12	17.00	19.82	13.19	10.94	12.50	15.63	12.43	6.65	10.44	12.83	13.41
Gemini Pro [58]	20.88	**61.38**	29.43	**77.35**	30.62	**74.26**	30.14	**81.53**	22.57	**70.31**	18.83	**66.22**	29.96	**65.01**	24.45	**69.97**

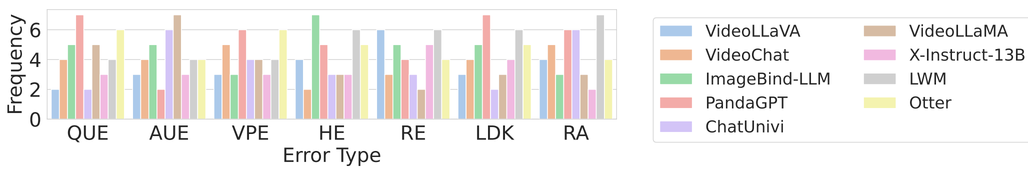


Figure 6: The frequency of different error types across various MLLMs. For each error type, 10 examples were evaluated. Error types are abbreviated as follows: QUE (Question Understanding Error), AUE (Audio Understanding Error), VPE (Visual Perception Error), HE (Hallucination Error), RE (Reasoning Error), LDK (Lack of Domain Knowledge), and RA (Reject to Answer).

4.6 Study on Modality of Perception

We conduct ablations to evaluate MLLMs ability to perceiving the world on the synthetic dataset of MMWorld. With our synthetic dataset, we considered scenarios where only one modality—either audio or visual—is available. Table 4 shows the results which evaluates the model's ability to interpret spoken language, background noises, and other audio elements without the aid of visual context and the model's perception ability to operate without any audio input. For the visual perception test, Gemini Pro performed the best, demonstrating its strong ability to process visual information. Interestingly, Video-Chat exhibited better audio perception than ChatUnivi, despite its poorer visual perception. This may be attributed to its use of the Whisper [54] speech recognition model. It also explains that in Table 3, Video-Chat outperforms ChatUnivi in the Art & Sports discipline, which requires a greater understanding of music, voice, and background audio. However, in other disciplines such as Science and Health & Medicine, Video-Chat's performance is significantly poorer.

4.7 Error Analysis

To gain deeper insights into the limitations of MLLMs, we prompted the models to explain the reasoning behind their choices, particularly when errors occurred. Through this analysis, we identified common error patterns and summarized them into seven distinct categories. We conducted a simple test where the same questions that triggered errors in GPT-4V were also posed to other MLLMs. The frequencies of each type of error are presented in Figure 6, as annotated by human evaluators. Detailed qualitative examples of these errors and further analysis are provided in the Appendix.

5 Conclusion

Our MMWorld Benchmark represents a significant step forward in the quest for advanced multi-modal language models capable of understanding complex video content. By presenting a diverse array of videos across seven disciplines, accompanied by questions that challenge models to demonstrate explanation, counterfactual thinking, future prediction, and domain expertise, we have created a rigorous testing ground for the next generation of AI. While using LLMs for data generation can introduce hallucination issues, these challenges are manageable and are commonly addressed [63; 55]. Another potential risk is the misuse of MLLMs for surveillance or privacy invasion. The ability of models to understand video content and perform reasoning could be exploited to monitor individuals without their consent, leading to serious ethical and legal concerns regarding privacy.

References

[1] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)

[2] Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J.H., Shafey, L.E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G.H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C.A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A.C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D.R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., Wu, Y.: Palm 2 technical report (2023)

[3] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)

[4] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)

[5] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023)

[6] Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021)

[7] Ben-Shabat, Y., Yu, X., Saleh, F., Campbell, D., Rodriguez-Opazo, C., Li, H., Gould, S.: The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 847–859 (2021)

[8] Bitton, Y., Bansal, H., Hessel, J., Shao, R., Zhu, W., Awadalla, A., Gardner, J., Taori, R., Schimdt, L.: Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. arXiv preprint arXiv:2308.06595 (2023)

[9] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)

[10] Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning (2023)

[11] Chen, W., Mees, O., Kumar, A., Levine, S.: Vision-language models provide promptable representations for reinforcement learning. arXiv preprint arXiv:2402.02651 (2024)

[12] Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), `https://lmsys.org/blog/2023-03-30-vicuna/`

[13] Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., Yao, H.: Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. arXiv preprint arXiv:2311.03287 (2023)

[14] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)

[15] Fan, Y., Gu, J., Zhou, K., Yan, Q., Jiang, S., Kuo, C.C., Guan, X., Wang, X.E.: Muffin or chihuahua? challenging large vision-language models with multipanel vqa (2024)

[16] Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)

[17] Fu, T.J., Yu, L., Zhang, N., Fu, C.Y., Su, J.C., Wang, W.Y., Bell, S.: Tell Me What Happened: Unifying Text-guided Video Completion via Multimodal Masked Video Generation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

[18] Google: Bard - chat based ai tool from google, powered by palm 2. https://bard.google.com/?hl=en (2023)

[19] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)

[20] Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

[21] Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)

[22] Hackl, V., Müller, A.E., Granitzer, M., Sailer, M.: Is gpt-4 a reliable rater? evaluating consistency in gpt-4 text ratings. arXiv preprint arXiv:2308.02575 (2023)

[23] Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., et al.: Imagebind-llm: Multi-modality instruction tuning. arXiv preprint arXiv:2309.03905 (2023)

[24] Himakunthala, V., Ouyang, A., Rose, D., He, R., Mei, A., Lu, Y., Sonar, C., Saxon, M., Wang, W.Y.: Let's think frame by frame with vip: A video infilling and prediction dataset for evaluating video chain-of-thought (2023)

[25] Hsu, T.Y., Huang, C.Y., Rossi, R., Kim, S., Giles, C.L., Huang, T.H.K.: Gpt-4 as an effective zero-shot evaluator for scientific figure captions. arXiv preprint arXiv:2310.15405 (2023)

[26] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)

[27] Jin, P., Takanobu, R., Zhang, C., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. arXiv preprint arXiv:2311.08046 (2023)

[28] LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62**(1) (2022)

[29] Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696 (2018)

[30] Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023)

[31] Li, H., Deng, A., Ke, Q., Liu, J., Rahmani, H., Guo, Y., Schiele, B., Chen, C.: Sports-qa: A large-scale video question answering benchmark for complex and professional sports. arXiv preprint arXiv:2401.01505 (2024)

[32] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)

[33] Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)

[34] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., Qiao, Y.: Mvbench: A comprehensive multi-modal video understanding benchmark. arXiv preprint arXiv: 2311.17005 (2023)

[35] Li, L., Lei, J., Gan, Z., Yu, L., Chen, Y.C., Pillai, R., Cheng, Y., Zhou, L., Wang, X.E., Wang, W.Y., et al.: Value: A multi-task benchmark for video-and-language understanding evaluation. arXiv preprint arXiv:2106.04632 (2021)

[36] Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)

[37] Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2794–2804 (2023)

[38] Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. In: Proceedings of the International Conference on Learning Representations (2024)

[39] Liu, H., Yan, W., Zaharia, M., Abbeel, P.: World model on million-length video and language with ringattention. arXiv preprint arXiv:2402.08268 (2024)

[40] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)

[41] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)

[42] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: Gpteval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634 (2023)

[43] Lu, Y., Jiang, D., Chen, W., Wang, W., Choi, Y., Lin, Y.: Wildvision arena: Benchmarking multi-modal llms in the wild (February 2024), `https://huggingface.co/spaces/WildVision/vision-arena/`

[44] Lu, Y., Li, X., Wang, W.Y., Choi, Y.: Vim: Probing multimodal large language models for visual embedded instruction following (2023)

[45] Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024) (2024)

[46] Morris, M.R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., Legg, S.: Levels of agi: Operationalizing progress on the path to agi. arXiv preprint arXiv:2311.02462 (2023)

[47] Ning, M., Zhu, B., Xie, Y., Lin, B., Cui, J., Yuan, L., Chen, D., Yuan, L.: Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. arXiv preprint arXiv:2311.16103 (2023)

[48] Ning, M., Zhu, B., Xie, Y., Lin, B., Cui, J., Yuan, L., Chen, D., Yuan, L.: Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. arXiv preprint arXiv:2311.16103 (2023)

[49] OpenAI: Gpt-4 technical report (2023)

[50] OpenAI: Gpt-4: Technical report. arXiv preprint arXiv:2303.08774 (2023)

[51] OpenAI: Gpt-4v(ision) system card. https://openai.com/research/gpt-4v-system-card (2023)

[52] Panagopoulou, A., Xue, L., Yu, N., Li, J., Li, D., Joty, S., Xu, R., Savarese, S., Xiong, C., Niebles, J.C.: X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. arXiv preprint arXiv:2311.18799 (2023)

[53] Pătrăucean, V., Smaira, L., Gupta, A., Continente, A.R., Markeeva, L., Banarse, D., Koppula, S., Heyward, J., Malinowski, M., Yang, Y., Doersch, C., Matejovicova, T., Sulsky, Y., Miech, A., Frechette, A., Klimczak, H., Koster, R., Zhang, J., Winkler, S., Aytar, Y., Osindero, S., Damen, D., Zisserman, A., Carreira, J.: Perception test: A diagnostic benchmark for multimodal video models. In: Advances in Neural Information Processing Systems (2023), `https://openreview.net/forum?id=HYEGXFnPoq`

[54] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. International Conference on Machine Learning (2022). https://doi.org/10.48550/arXiv.2212.04356

[55] Shen, X., Chen, Z., Backes, M., Shen, Y., Zhang, Y.: "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. arXiv preprint arXiv: 2308.03825 (2023)

[56] Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355 (2023)

[57] Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4631–4640 (2016)

[58] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

[59] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[60] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

[61] Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4581–4591 (2019)

[62] Wang, X., Zhou, Y., Liu, X., Lu, H., Xu, Y., He, F., Yoon, J., Lu, T., Bertasius, G., Bansal, M., et al.: Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. arXiv preprint arXiv:2401.10529 (2024)

[63] Wang, Y., Li, H., Han, X., Nakov, P., Baldwin, T.: Do-not-answer: Evaluating safeguards in LLMs. In: Graham, Y., Purver, M. (eds.) Findings of the Association for Computational Linguistics: EACL 2024. pp. 896–911. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024), https://aclanthology.org/2024.findings-eacl.61

[64] Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: 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)

[65] Xiang, J., Liu, G., Gu, Y., Gao, Q., Ning, Y., Zha, Y., Feng, Z., Tao, T., Hao, S., Shi, Y., Liu, Z., Xing, E.P., Hu, Z.: Pandora: Towards general world model with natural language actions and video states (2024)

[66] Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1645–1653 (2017)

[67] Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016), https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/

[68] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)

[69] Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: CLEVRER: collision events for video representation and reasoning. In: ICLR (2020)

[70] Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)

[71] Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9127–9134 (2019)

[72] Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: AAAI. pp. 9127–9134 (2019)

[73] Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502 (2023)

[74] Zeng, K.H., Chen, T.H., Chuang, C.Y., Liao, Y.H., Niebles, J.C., Sun, M.: Leveraging video descriptions to learn video question answering. Proceedings of the AAAI Conference on Artificial Intelligence **31**(1) (Feb 2017). https://doi.org/10.1609/aaai.v31i1.11238, https://ojs.aaai.org/index.php/AAAI/article/view/11238

[75] Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)

[76] Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W.Y., Petzold, L.R.: Gpt-4v(ision) as a generalist evaluator for vision-language tasks (2023)

[77] Zheng, K., He, X., Wang, X.E.: Minigpt-5: Interleaved vision-and-language generation via generative vokens. arXiv preprint arXiv:2310.02239 (2023)

[78] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We included the code and data in the supplemental material and we also provided a URL link.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 4.3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]