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

Anonymous Author(s)
Affiliation
Address
email

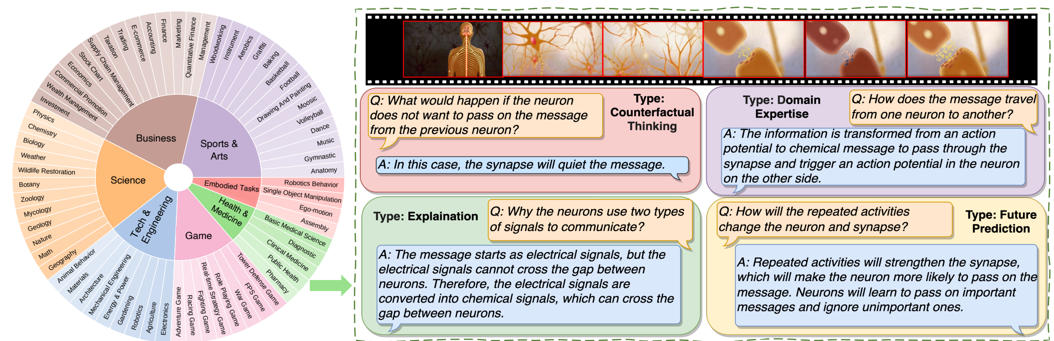

Figure 1: MMWorld covers seven broad disciplines and 69 subdisciplines, focusing on the evaluation of multi-faceted reasoning beyond perception (e.g., explanation, counterfactual thinking, future prediction, domain expertise). On the right is a video sample from the Health & Medicine discipline.

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.

1 Introduction

Foundation models, such as Large Language Models (LLMs) [OpenAI, 2023b; Touvron et al., 2023; Jiang et al., 2023; Anil et al., 2023] and Multimodal LLMs (MLLMs) [OpenAI, 2023a; Team et al., 2023; Lin et al., 2023; Li et al., 2023b; Maaz et al., 2024; Chen et al., 2023], 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 [LeCun, 2022; Chen et al., 2024; Ha and Schmidhuber, 2018; Xiang et al., 2024] 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 [Li et al., 2023c; Ning et al., 2023; Pătrăucean et al., 2023; Li et al., 2023c], however, fall short in two key perspectives for such evaluations. First, as LeCun et al. [LeCun, 2022] 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 [Morris et al., 2023; Yue et al., 2023].

Therefore, we introduce MMWorld, a multi-discipline multi-faceted multimodal video understanding benchmark to comprehensively evaluate MLLMs' abilities in reasoning and interpreting real-world dynamics [1]. 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 video 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.

2 Experiments

2.1 Main Evaluation Results

We show in Table 1 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 [Bain et al., 2021]. 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 [Maaz et al., 2024]. This superior performance may also be attributed to Video-LLaVA's adoption of CLIP ViT-L/14 trained in LanguageBind [Lin et al., 2023] 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.

[1] 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: 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-4o [OpenAI, 2024]	47.87 ±1.47	**91.14** ±0.87	**73.78** ±2.88	**83.33** ±1.47	62.94 ±3.47	**75.53** ±2.61	**80.32** ±2.05	**62.54** ±0.79
Claude-3.5-Sonnet [Anthropic, 2024]	**54.58** ±0.45	63.87 ±0.40	59.85 ±1.28	54.51 ±1.28	30.99 ±0.40	58.87 ±0.61	59.44 ±0.68	54.54 ±0.29
GPT-4V [OpenAI, 2023a]	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 [Team et al., 2023]	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 [Lin et al., 2023]	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 [Li et al., 2023b]	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 [Jin et al., 2023]	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 [Ye et al., 2023]	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
Video-ChatGPT-7B [Maaz et al., 2024]	26.84 ±0.69	39.16 ±3.02	36.45 ±1.31	53.12 ±0.00	36.60 ±3.25	41.49 ±1.74	36.55 ±2.27	33.27 ±0.97
PandaGPT-7B [Su et al., 2023]	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 [Han et al., 2023]	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 [Panagopoulou et al., 2023]	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 [Liu et al., 2024]	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 [Li et al., 2023a]	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 [Zhang et al., 2023]	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

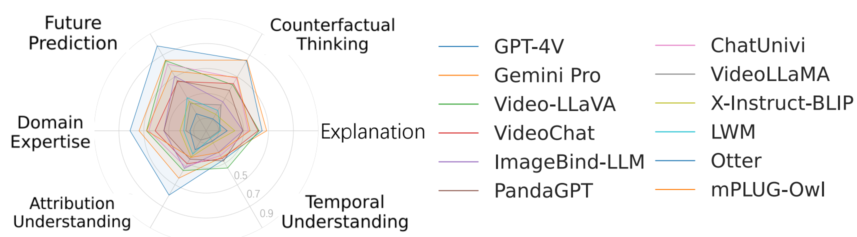

Figure 2: Results of different MLLMs on multi-faceted reasoning. The detailed performance numbers can be found in the Appendix.

2.2 Study on Multi-faceted Reasoning on MMWorld

Figure 2 illustrates the multi-faceted reasoning performance for each MLLM. GPT-4V emerges as the strongest model across Future Prediction, Domain Expertise, and Attribution Understanding. 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 1. 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.

3 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 [Wang et al., 2024; Shen et al., 2023].

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023.

Anthropic. Introducing the next generation of Claude. `https://www.anthropic.com/news/claude-3-family`, 2024. Accessed: 2024-07-29.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023.

William Chen, Oier Mees, Aviral Kumar, and Sergey Levine. Vision-language models provide promptable representations for reinforcement learning. *arXiv preprint arXiv:2402.02651*, 2024.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022.

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*, 2023a.

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*, 2023b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv: 2311.17005*, 2023c.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 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.

Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi: Operationalizing progress on the path to agi. *arXiv preprint arXiv:2311.02462*, 2023.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023.

OpenAI. Gpt-4v(ision) system card. *https://openai.com/research/gpt-4v-system-card*, 2023a.

OpenAI. Gpt-4 technical report, 2023b.

OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024. Accessed: 2024-07-29.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*, 2023.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=HYEGXFnPoq`.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv: 2308.03825*, 2023.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: Evaluating safeguards in LLMs. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.findings-eacl.61`.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Pandora: Towards general world model with natural language actions and video states. 2024.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

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.