

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

Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Kevin Lin, William Yang Wang, Lijuan Wang, Xin Eric Wang



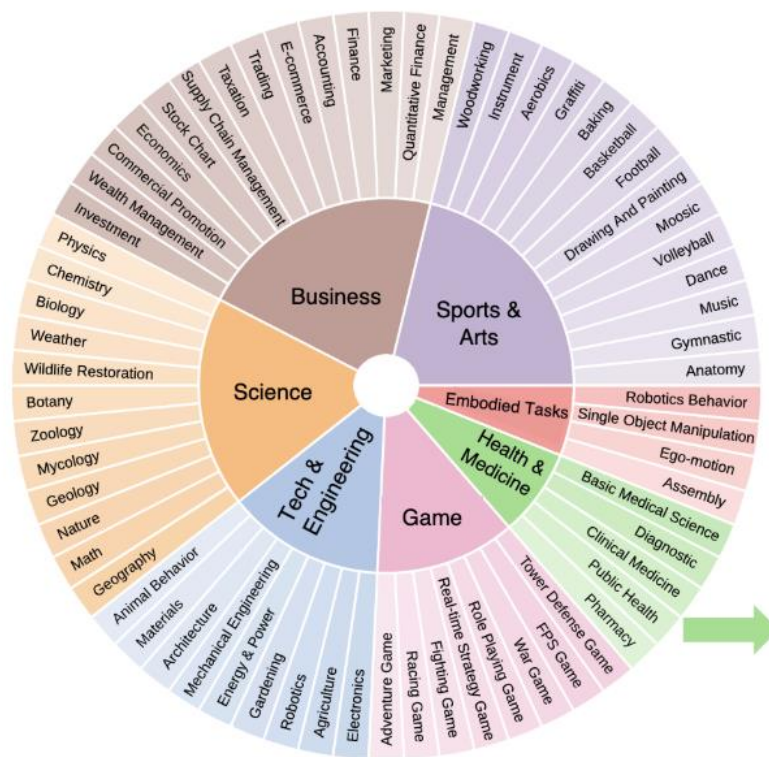


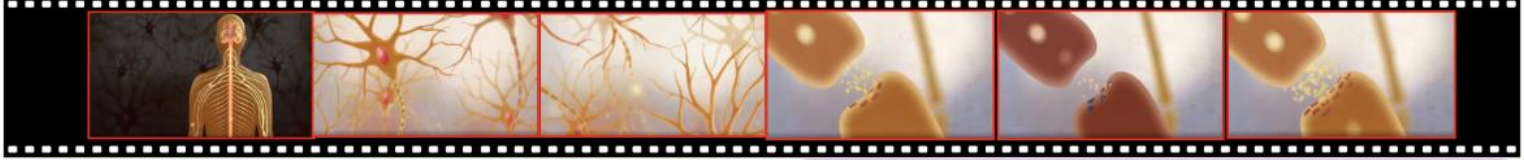
Motivation

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.



Multi-discipline Multi-faceted Video Understanding Benchmark





Type: Counterfactual Thinking

Q: What would happen if the neuron does not want to pass on the message from the previous neuron?

A: In this case, the synapse will quiet the message.

Type: Domain Expertise

Q: How does the message travel from one neuron to another?

A: The information is transformed from an action potential to chemical message to pass through the synapse and trigger an action potential in the neuron on the other side.

Type: Explanation

Q: Why the neurons use two types of signals to communicate?

A: The message starts as electrical signals, but the electrical signals cannot cross the gap between neurons. Therefore, the electrical signals are converted into chemical signals, which can cross the gap between neurons.

Type: Future Prediction

Q: How will the repeated activities change the neuron and synapse?

A: Repeated activities will strengthen the synapse, which will make the neuron more likely to pass on the message. Neurons will learn to pass on important messages and ignore unimportant ones.

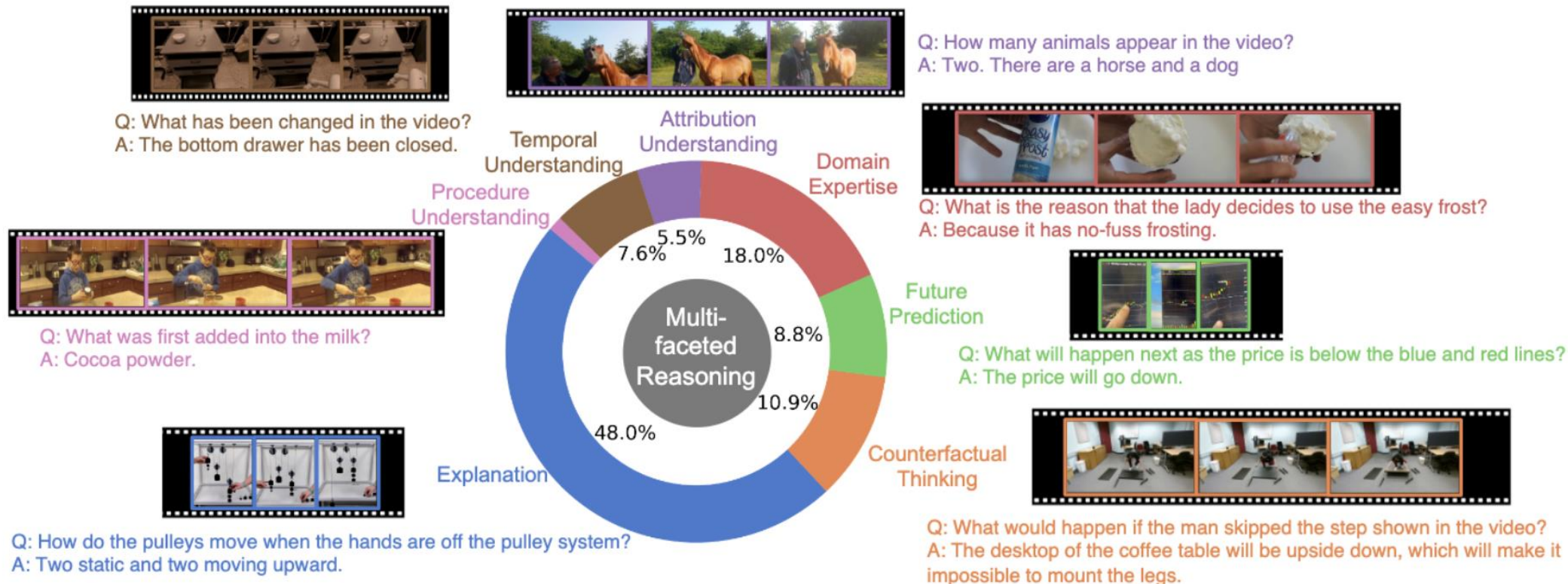


Dataset Characteristics

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

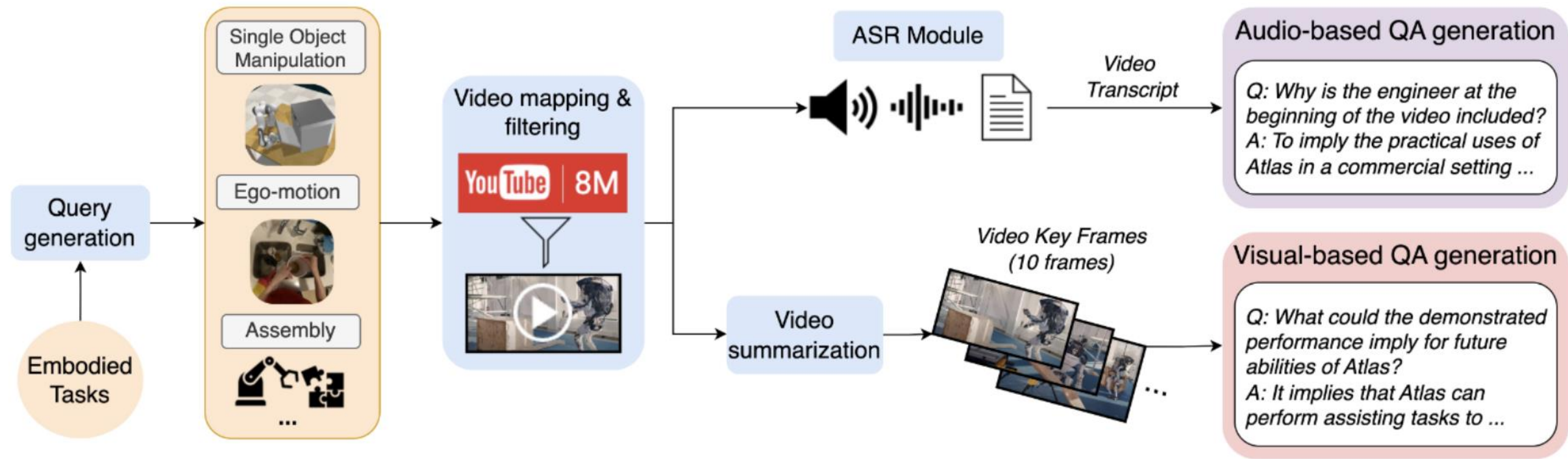


Question Types Distribution



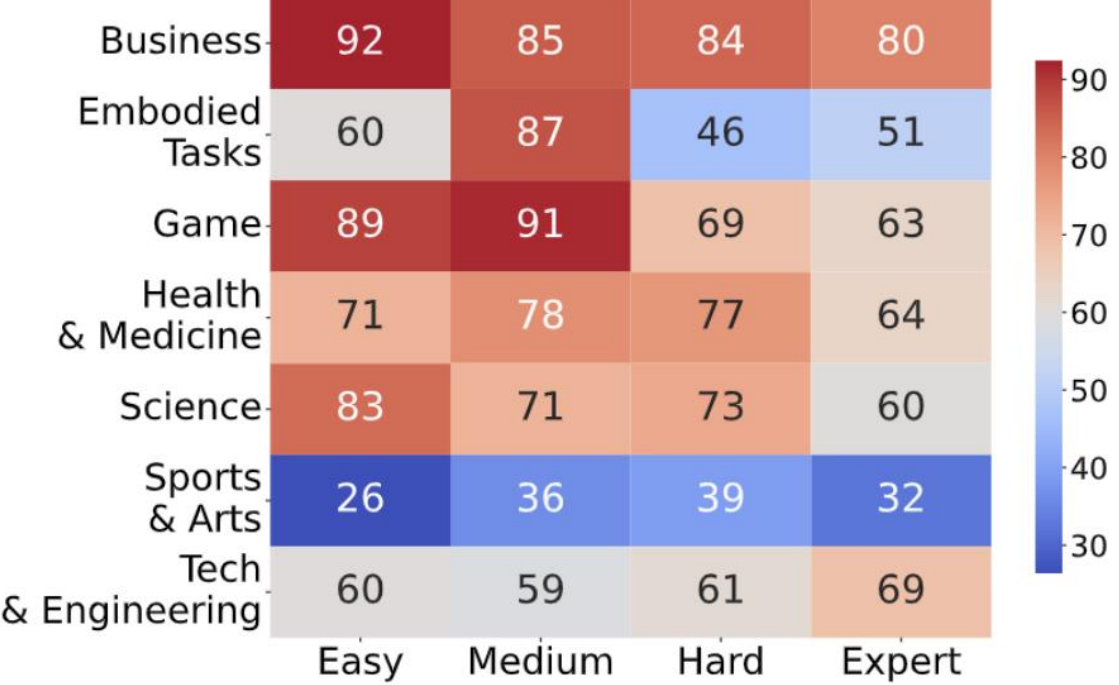
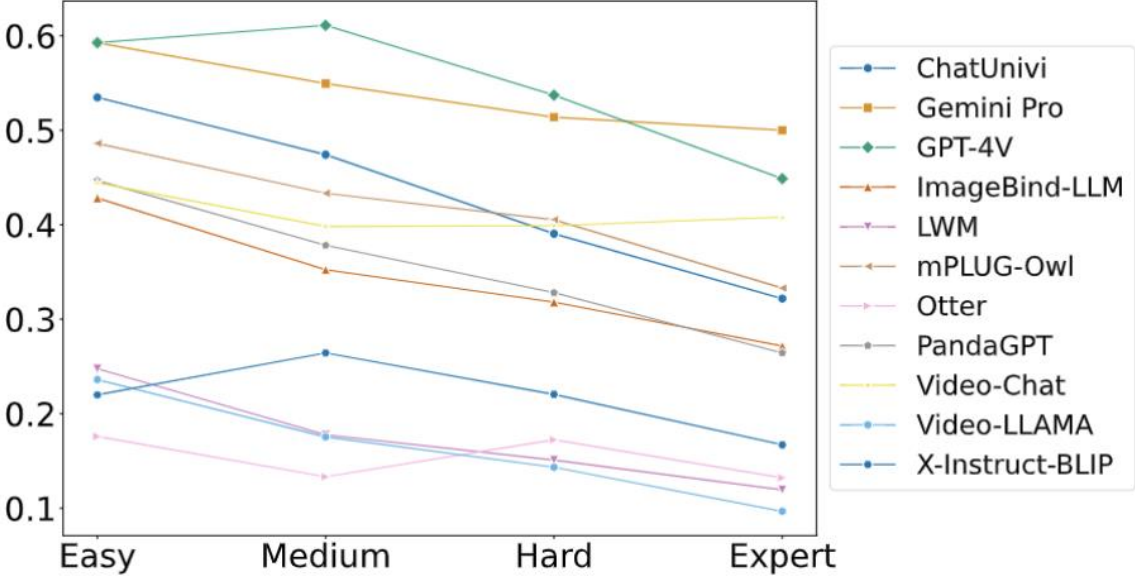


Synthetic Data Generation Pipeline





Study on MLLM Performance at Different Difficulty Levels for Average Humans





MLLM accuracy across diverse disciplines

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, 2023b)	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., 2023a)	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., 2023c)	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., 2024b)	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., 2023a)	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



Conclusions and Future Works

- 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.