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

Xinyu Fang*, Kangrui Mao*, Haodong Duan†, Xiangyu Zhao,
Yining Li, Dahua Lin, Kai Chen†

* Equal Contribution † Corresponding Author

Presenter: Xinyu Fang
Nov, 2024

The Existing VideoQA benchmarks have following limitations:

1. **Short Videos**: Existing VideoQA datasets primarily consist of **short videos** (less than a minute), that deviate from the real application scenario.

2. **Limited Capabilities**: Current VideoQA benchmarks are **limited to several basic video tasks**.

3. **Biased Evaluation**: Our preliminary study indicates that **GPT-3.5-based evaluation is less accurate** and exhibits **significant discrepancy relative to human preferences**, diminishing the credibility of the evaluation results.

Low duration and shot numbers

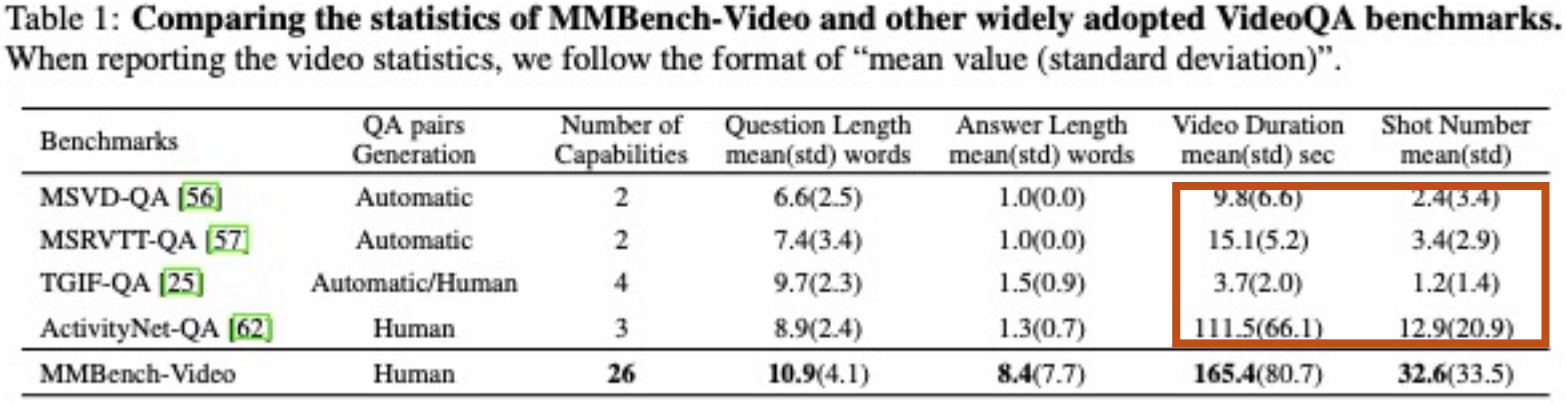
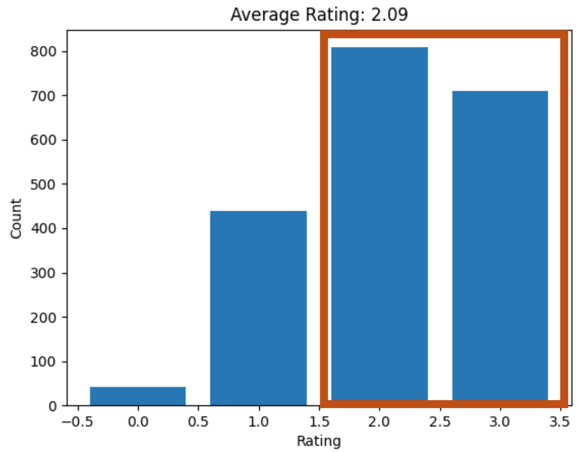


Table 1: **Comparing the statistics of MMBench-Video and other widely adopted VideoQA benchmarks.** When reporting the video statistics, we follow the format of "mean value (standard deviation)".

Benchmarks	QA pairs Generation	Number of Capabilities	Question Length mean(std) words	Answer Length mean(std) words	Video Duration mean(std) sec	Shot Number mean(std)
MSVD-QA [56]	Automatic	2	6.6(2.5)	1.0(0.0)	9.8(6.6)	2.4(3.4)
MSRVTT-QA [57]	Automatic	2	7.4(3.4)	1.0(0.0)	15.1(5.2)	3.4(2.9)
TGIF-QA [25]	Automatic/Human	4	9.7(2.3)	1.5(0.9)	3.7(2.0)	1.2(1.4)
ActivityNet-QA [62]	Human	3	8.9(2.4)	1.3(0.7)	111.5(66.1)	12.9(20.9)
MMBench-Video	Human	**26**	**10.9**(4.1)	**8.4**(7.7)	**165.4**(80.7)	**32.6**(33.5)



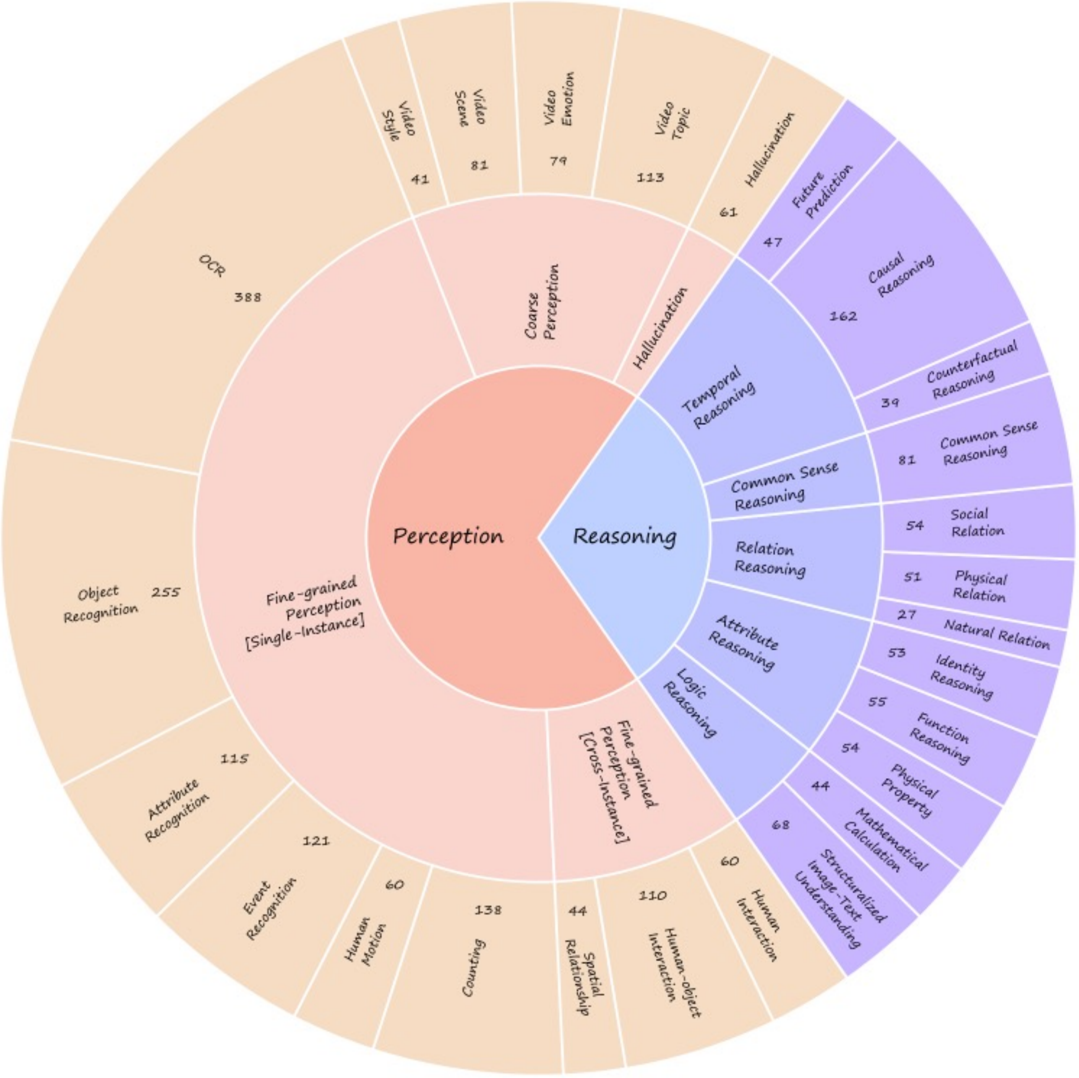
Great Bias In Judge

The project aims at designing a new VideoQA benchmark featuring the following characteristics:

1. The benchmark needs to cover videos of **multiple lengths and shots**, **mirroring practical use cases**.
2. This benchmark needs to cover a **wide range of capabilities** related to video comprehension, with **sufficient consideration of temporal**.
3. The benchmark should be **evaluated based on more advanced LLMs** (like GPT-4 or Qwen).

NEURAL INFORMATION PROCESSING SYSTEMS

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory



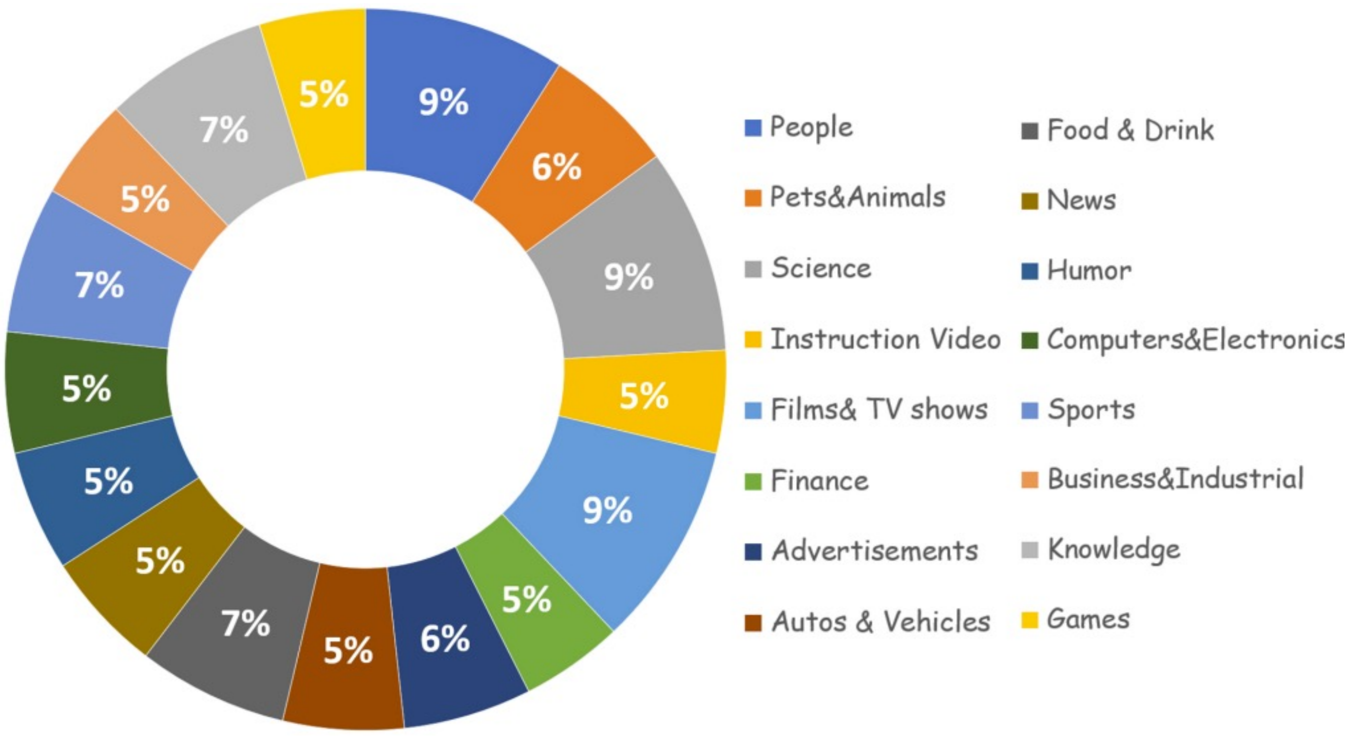
Follow the MMBench, We design a taxonomy of multi-modal video understanding capabilities:

1. The taxonomy features **3 capability levels** and **26 fine-grained capabilities**.
2. The two most fundamental L-1 capabilities are **perception** & **reasoning**.
3. Three additional L-2 capabilities: **Hallucination**, **Commonsense Reasoning**, **Temporal Reasoning**

NEURAL INFORMATION PROCESSING SYSTEMS

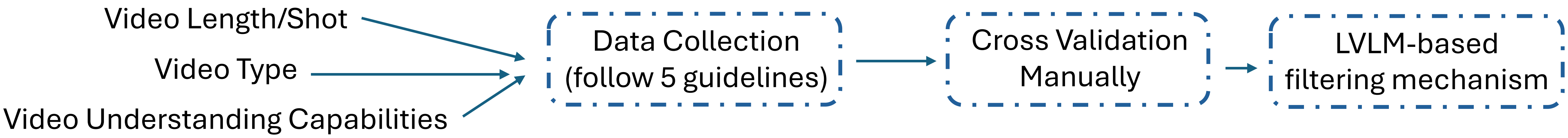
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Dataset collection and Quality Control:



Five Guidelines:

1. Each question should evaluate **one or multiple leaf capabilities**.
2. You are encouraged to formulate **temporal indispensable questions**…
3. **Avoid including specific timestamps** in the questions
4. The questions should **be free-form and exhibit linguistic diversified** ..
5. Please **provide informative and detailed answers** for each question



Video Length/Shot

Video Type

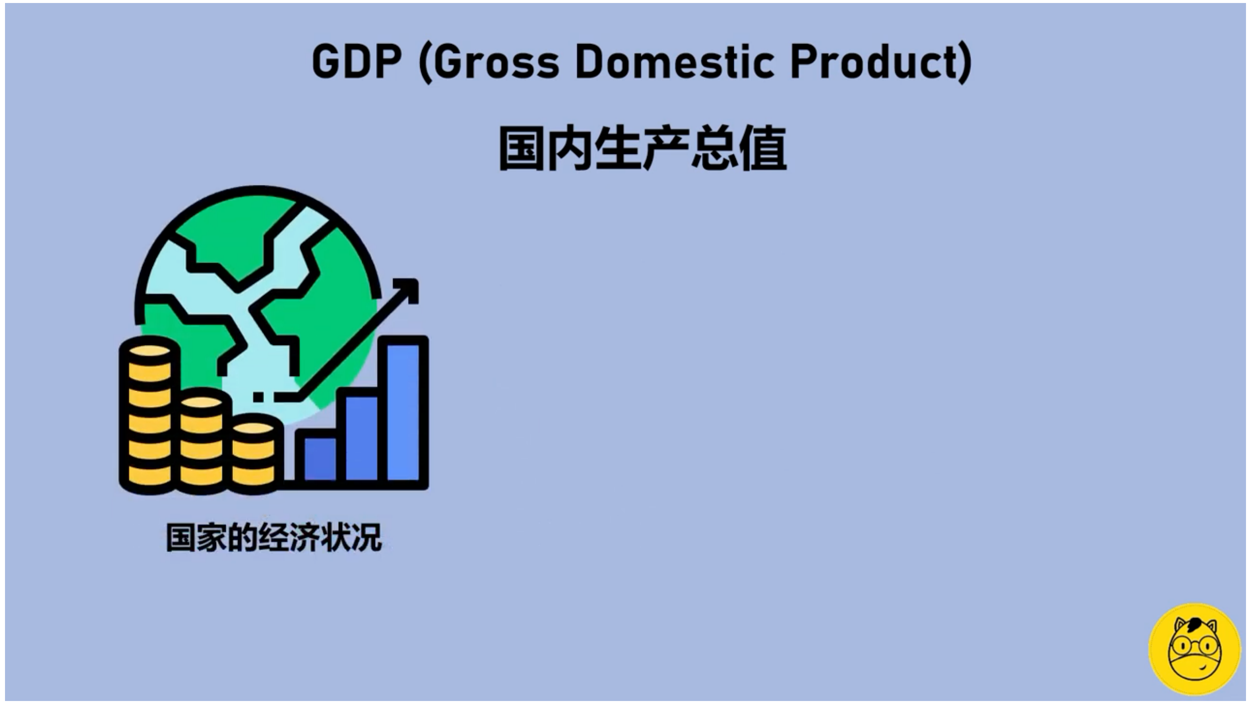
Video Understanding Capabilities

Data Collection (follow 5 guidelines)

Cross Validation Manually

LVLM-based filtering mechanism

MMBench-Video Example:



GDP (Gross Domestic Product)

国内生产总值

国家的经济状况

How to Calculate GDP according to this video?

GDP = C + I + G + (X − M). GDP equals the sum of consumption, private investment, government spending, and net exports.

How many financial terms does this video explain?

A total of 5 terms are discussed in the video. Respectively GDP, GNP, GNI, CPI, and PPI.

What is this video mainly about?

This video explains some financial terms.

MMBench-Video highlight features:

1. Long-form, multi-shot video benchmarks

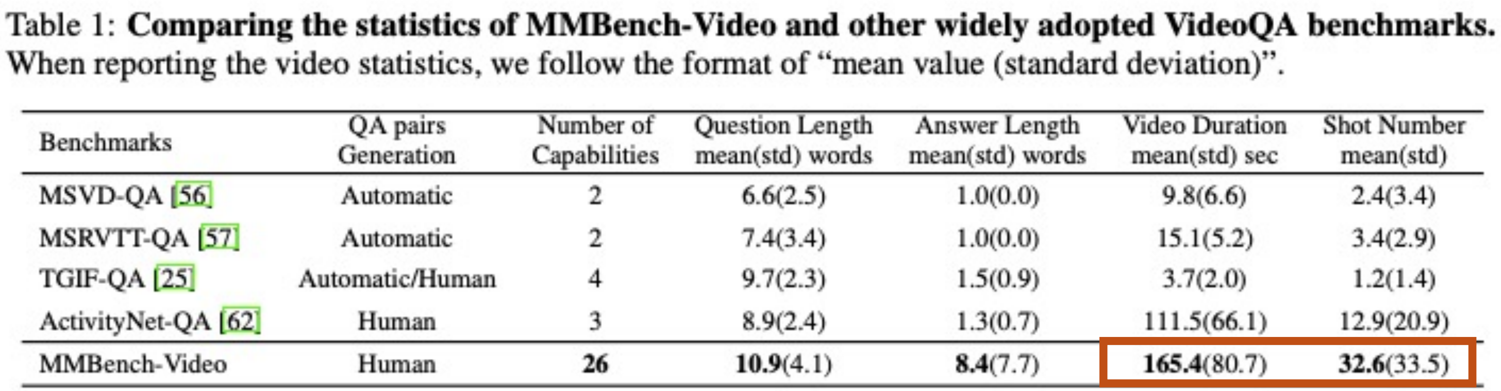
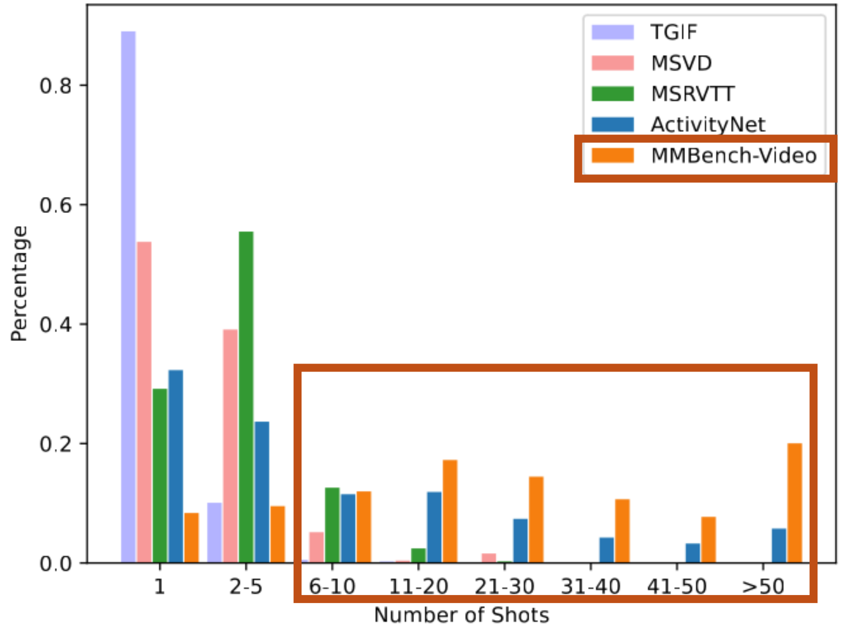


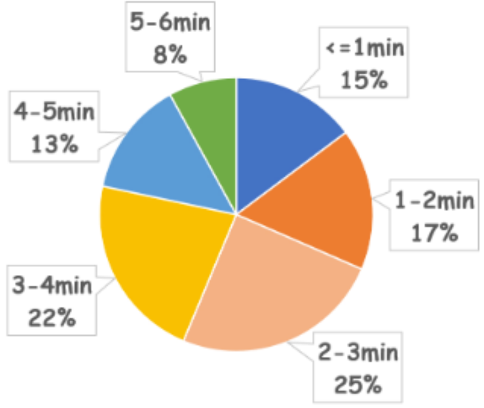
Table 1: **Comparing the statistics of MMBench-Video and other widely adopted VideoQA benchmarks.** When reporting the video statistics, we follow the format of "mean value (standard deviation)".

Benchmarks	QA pairs Generation	Number of Capabilities	Question Length mean(std) words	Answer Length mean(std) words	Video Duration mean(std) sec	Shot Number mean(std)
MSVD-QA [56]	Automatic	2	6.6(2.5)	1.0(0.0)	9.8(6.6)	2.4(3.4)
MSRVTT-QA [57]	Automatic	2	7.4(3.4)	1.0(0.0)	15.1(5.2)	3.4(2.9)
TGIF-QA [25]	Automatic/Human	4	9.7(2.3)	1.5(0.9)	3.7(2.0)	1.2(1.4)
ActivityNet-QA [62]	Human	3	8.9(2.4)	1.3(0.7)	111.5(66.1)	12.9(20.9)
MMBench-Video	Human	**26**	**10.9**(4.1)	**8.4**(7.7)	**165.4**(80.7)	**32.6**(33.5)

- boasts a substantially greater average duration than existing benchmarks.
- significantly surpasses all other benchmarks in average shot count.



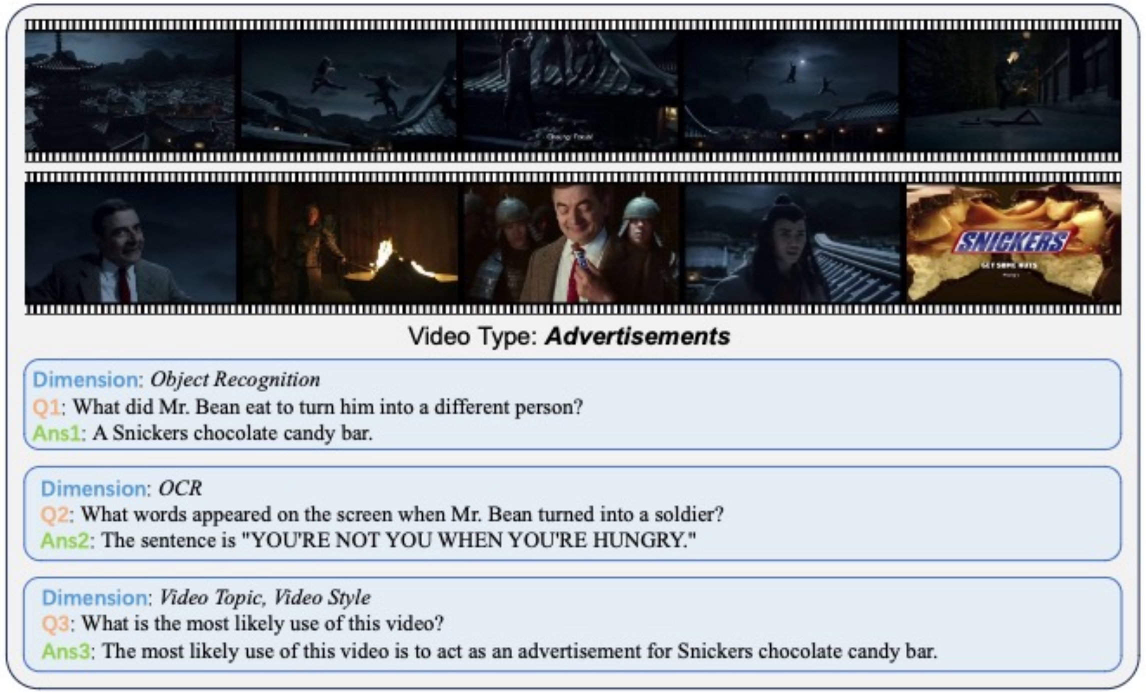
👆 Shot Number Distribution Comparison



👈 Duration Distribution of MMBench-Video

MMBench-Video highlight features:

2. Rich linguistic diversity
3. Comprehensive Capability Coverage in video understanding

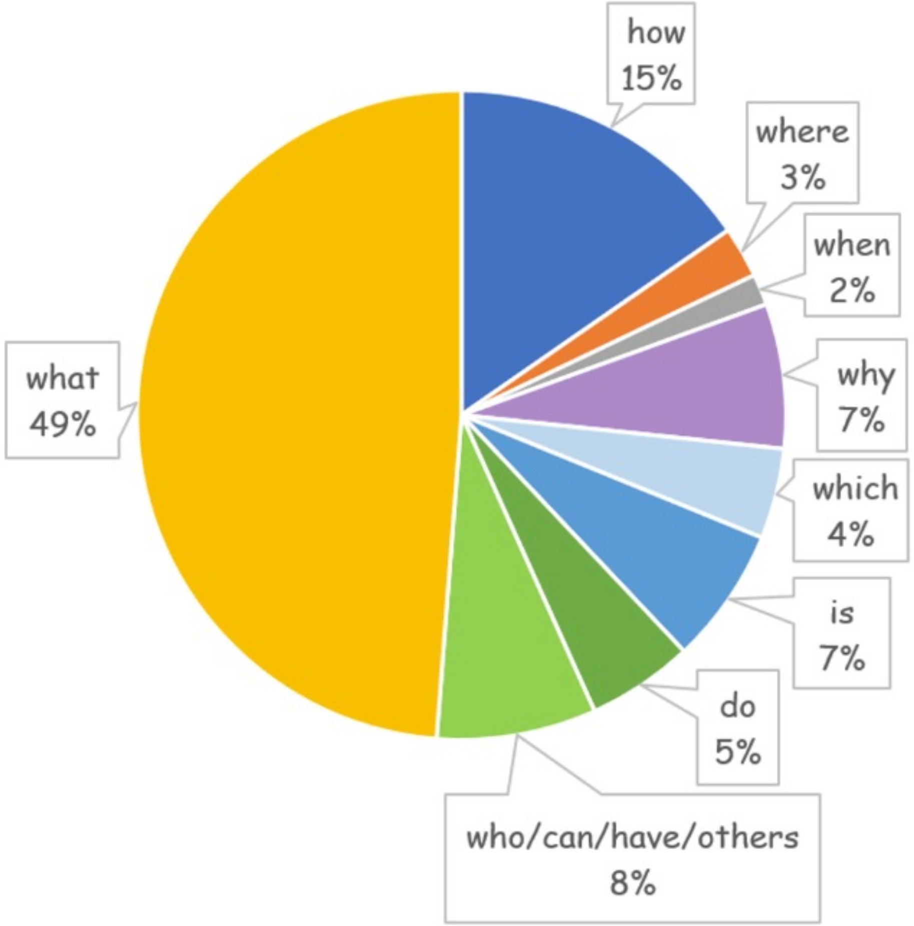


Video Type: ***Advertisements***

Dimension: *Object Recognition*
Q1: What did Mr. Bean eat to turn him into a different person?
Ans1: A Snickers chocolate candy bar.

Dimension: *OCR*
Q2: What words appeared on the screen when Mr. Bean turned into a soldier?
Ans2: The sentence is "YOU'RE NOT YOU WHEN YOU'RE HUNGRY."

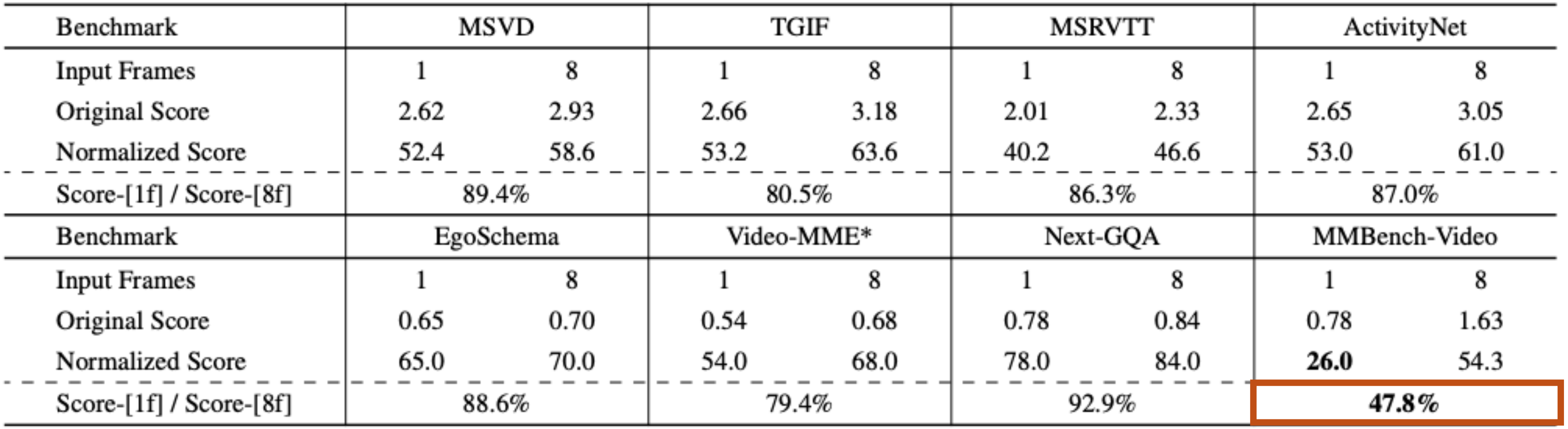
Dimension: *Video Topic, Video Style*
Q3: What is the most likely use of this video?
Ans3: The most likely use of this video is to act as an advertisement for Snickers chocolate candy bar.



MMBench-Video

MMBench-Video highlight features:

4. Adequate Temporal Indispensability

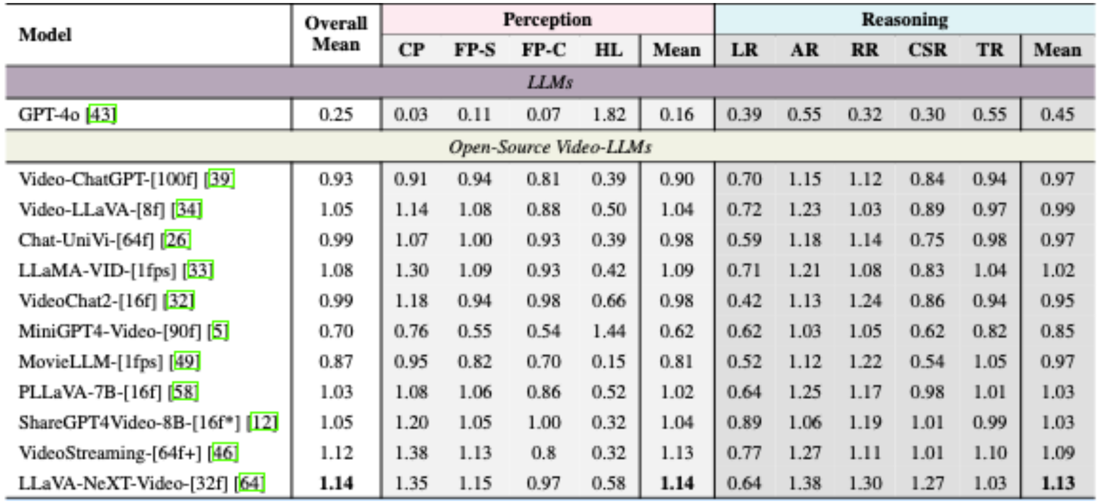


Benchmark	MSVD		TGIF		MSRVTT		ActivityNet	
Input Frames	1	8	1	8	1	8	1	8
Original Score	2.62	2.93	2.66	3.18	2.01	2.33	2.65	3.05
Normalized Score	52.4	58.6	53.2	63.6	40.2	46.6	53.0	61.0
Score-[1f] / Score-[8f]	89.4%		80.5%		86.3%		87.0%	
Benchmark	EgoSchema		Video-MME*		Next-GQA		MMBench-Video	
Input Frames	1	8	1	8	1	8	1	8
Original Score	0.65	0.70	0.54	0.68	0.78	0.84	0.78	1.63
Normalized Score	65.0	70.0	54.0	68.0	78.0	84.0	**26.0**	54.3
Score-[1f] / Score-[8f]	88.6%		79.4%		92.9%		**47.8%**	

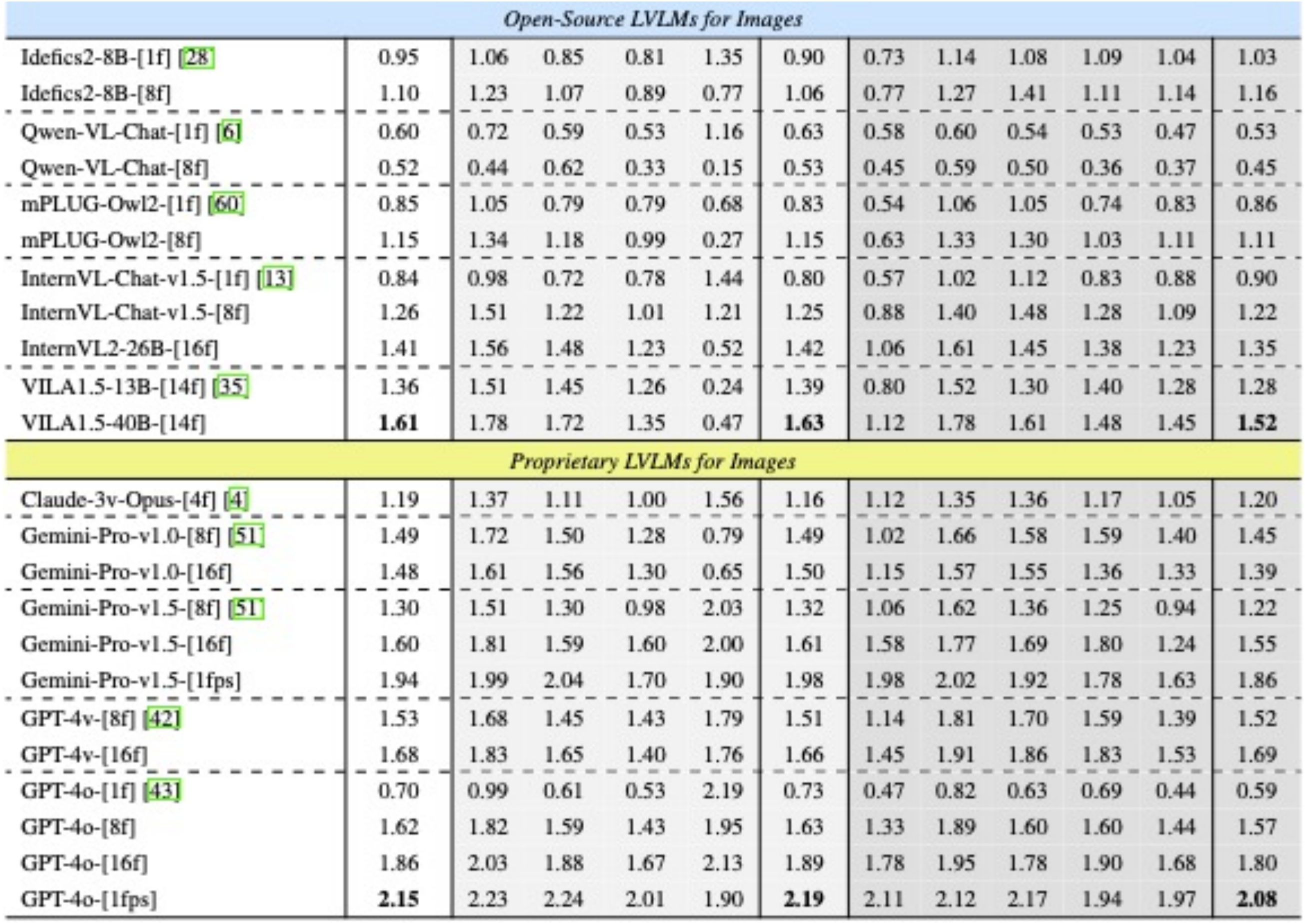
1. Allow most videos for its content to be adequately represented by a single frame.
2. Many of the QAs are too simplistic

Exhibit Great Temporal Importance of MMBench-Video

Main Results (Oct. 2024)



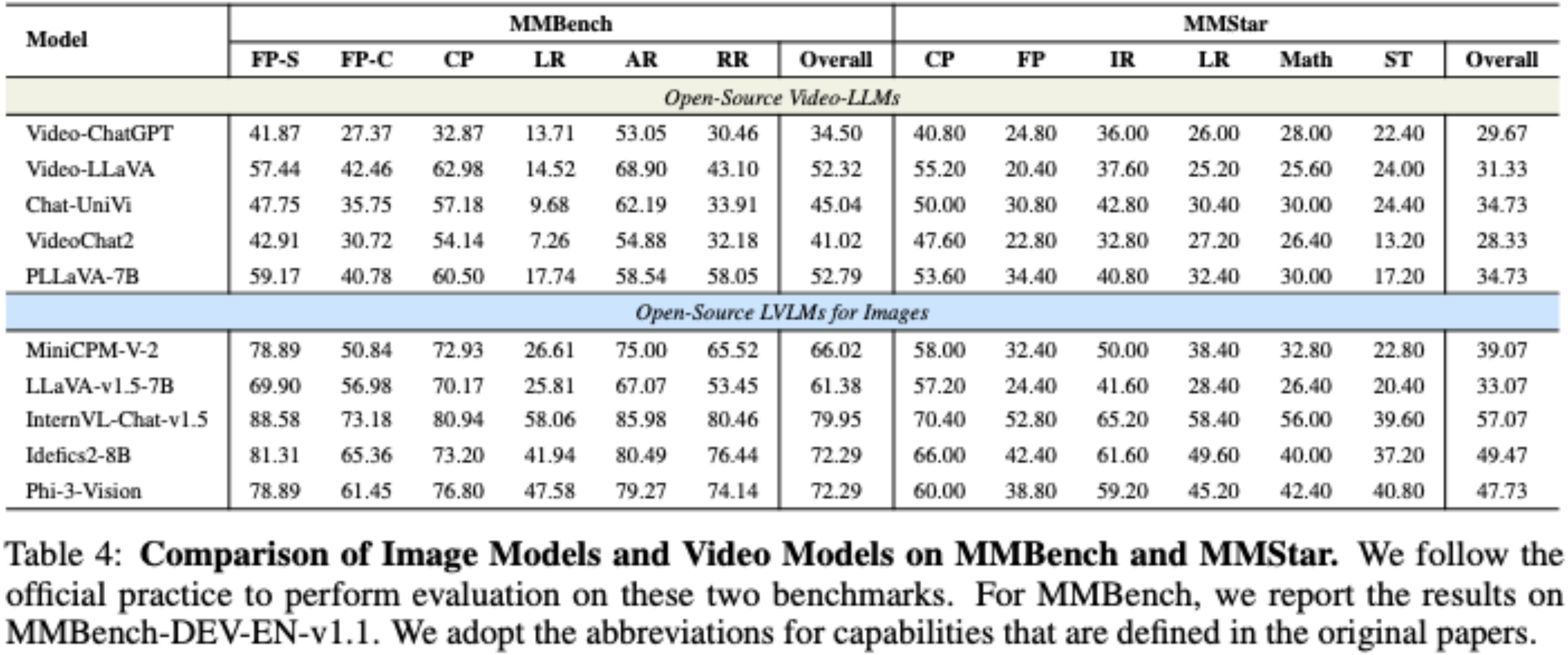
Model	Overall Mean	Perception					Reasoning					
		CP	FP-S	FP-C	HL	Mean	LR	AR	RR	CSR	TR	Mean
LLMs												
GPT-4o [43]	0.25	0.03	0.11	0.07	1.82	0.16	0.39	0.55	0.32	0.30	0.55	0.45
Open-Source Video-LLMs												
Video-ChatGPT-[100f] [39]	0.93	0.91	0.94	0.81	0.39	0.90	0.70	1.15	1.12	0.84	0.94	0.97
Video-LLaVA-[8f] [34]	1.05	1.14	1.08	0.88	0.50	1.04	0.72	1.23	1.03	0.89	0.97	0.99
Chat-UniVi-[64f] [26]	0.99	1.07	1.00	0.93	0.39	0.98	0.59	1.18	1.14	0.75	0.98	0.97
LLaMA-VID-[1fps] [33]	1.08	1.30	1.09	0.93	0.42	1.09	0.71	1.21	1.08	0.83	1.04	1.02
VideoChat2-[16f] [32]	0.99	1.18	0.94	0.98	0.66	0.98	0.42	1.13	1.24	0.86	0.94	0.95
MiniGPT4-Video-[90f] [5]	0.70	0.76	0.55	0.54	1.44	0.62	0.62	1.03	1.05	0.62	0.82	0.85
MovieLLM-[1fps] [49]	0.87	0.95	0.82	0.70	0.15	0.81	0.52	1.12	1.22	0.54	1.05	0.97
PLLaVA-7B-[16f] [58]	1.03	1.08	1.06	0.86	0.52	1.02	0.64	1.25	1.17	0.98	1.01	1.03
ShareGPT4Video-8B-[16f*] [12]	1.05	1.20	1.05	1.00	0.32	1.04	0.89	1.06	1.19	1.01	0.99	1.03
VideoStreaming-[64f+] [46]	1.12	1.38	1.13	0.8	0.32	1.13	0.77	1.27	1.11	1.01	1.10	1.09
LLaVA-NeXT-Video-[32f] [64]	**1.14**	1.35	1.15	0.97	0.58	**1.14**	0.64	1.38	1.30	1.27	1.03	**1.13**



Open-Source LVLMs for Images												
Idefics2-8B-[1f] [28]	0.95	1.06	0.85	0.81	1.35	0.90	0.73	1.14	1.08	1.09	1.04	1.03
Idefics2-8B-[8f]	1.10	1.23	1.07	0.89	0.77	1.06	0.77	1.27	1.41	1.11	1.14	1.16
Qwen-VL-Chat-[1f] [6]	0.60	0.72	0.59	0.53	1.16	0.63	0.58	0.60	0.54	0.53	0.47	0.53
Qwen-VL-Chat-[8f]	0.52	0.44	0.62	0.33	0.15	0.53	0.45	0.59	0.50	0.36	0.37	0.45
mPLUG-Owl2-[1f] [60]	0.85	1.05	0.79	0.79	0.68	0.83	0.54	1.06	1.05	0.74	0.83	0.86
mPLUG-Owl2-[8f]	1.15	1.34	1.18	0.99	0.27	1.15	0.63	1.33	1.30	1.03	1.11	1.11
InternVL-Chat-v1.5-[1f] [13]	0.84	0.98	0.72	0.78	1.44	0.80	0.57	1.02	1.12	0.83	0.88	0.90
InternVL-Chat-v1.5-[8f]	1.26	1.51	1.22	1.01	1.21	1.25	0.88	1.40	1.48	1.28	1.09	1.22
InternVL2-26B-[16f]	1.41	1.56	1.48	1.23	0.52	1.42	1.06	1.61	1.45	1.38	1.23	1.35
VILA1.5-13B-[14f] [35]	1.36	1.51	1.45	1.26	0.24	1.39	0.80	1.52	1.30	1.40	1.28	1.28
VILA1.5-40B-[14f]	**1.61**	1.78	1.72	1.35	0.47	**1.63**	1.12	1.78	1.61	1.48	1.45	**1.52**
Proprietary LVLMs for Images												
Claude-3v-Opus-[4f] [4]	1.19	1.37	1.11	1.00	1.56	1.16	1.12	1.35	1.36	1.17	1.05	1.20
Gemini-Pro-v1.0-[8f] [51]	1.49	1.72	1.50	1.28	0.79	1.49	1.02	1.66	1.58	1.59	1.40	1.45
Gemini-Pro-v1.0-[16f]	1.48	1.61	1.56	1.30	0.65	1.50	1.15	1.57	1.55	1.36	1.33	1.39
Gemini-Pro-v1.5-[8f] [51]	1.30	1.51	1.30	0.98	2.03	1.32	1.06	1.62	1.36	1.25	0.94	1.22
Gemini-Pro-v1.5-[16f]	1.60	1.81	1.59	1.60	2.00	1.61	1.58	1.77	1.69	1.80	1.24	1.55
Gemini-Pro-v1.5-[1fps]	1.94	1.99	2.04	1.70	1.90	1.98	1.98	2.02	1.92	1.78	1.63	1.86
GPT-4v-[8f] [42]	1.53	1.68	1.45	1.43	1.79	1.51	1.14	1.81	1.70	1.59	1.39	1.52
GPT-4v-[16f]	1.68	1.83	1.65	1.40	1.76	1.66	1.45	1.91	1.86	1.83	1.53	1.69
GPT-4o-[1f] [43]	0.70	0.99	0.61	0.53	2.19	0.73	0.47	0.82	0.63	0.69	0.44	0.59
GPT-4o-[8f]	1.62	1.82	1.59	1.43	1.95	1.63	1.33	1.89	1.60	1.60	1.44	1.57
GPT-4o-[16f]	1.86	2.03	1.88	1.67	2.13	1.89	1.78	1.95	1.78	1.90	1.68	1.80
GPT-4o-[1fps]	**2.15**	2.23	2.24	2.01	1.90	**2.19**	2.11	2.12	2.17	1.94	1.97	**2.08**

Full results shows on the OpenVLM Video Leaderboard.

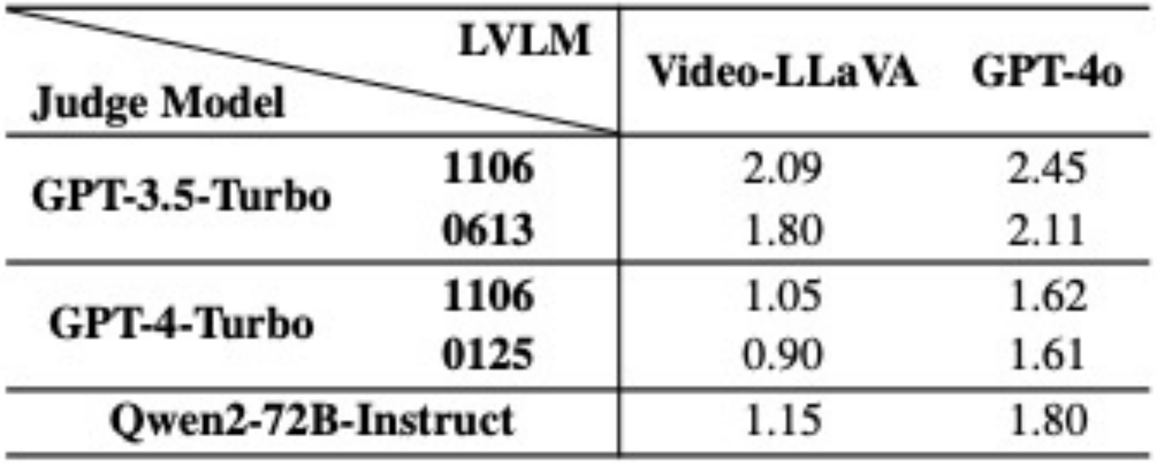
Performance of Video-LLMs on Image VQA Benchmarks



Model	MMBench							MMStar						
	FP-S	FP-C	CP	LR	AR	RR	Overall	CP	FP	IR	LR	Math	ST	Overall
Open-Source Video-LLMs														
Video-ChatGPT	41.87	27.37	32.87	13.71	53.05	30.46	34.50	40.80	24.80	36.00	26.00	28.00	22.40	29.67
Video-LLaVA	57.44	42.46	62.98	14.52	68.90	43.10	52.32	55.20	20.40	37.60	25.20	25.60	24.00	31.33
Chat-UniVi	47.75	35.75	57.18	9.68	62.19	33.91	45.04	50.00	30.80	42.80	30.40	30.00	24.40	34.73
VideoChat2	42.91	30.72	54.14	7.26	54.88	32.18	41.02	47.60	22.80	32.80	27.20	26.40	13.20	28.33
PLLaVA-7B	59.17	40.78	60.50	17.74	58.54	58.05	52.79	53.60	34.40	40.80	32.40	30.00	17.20	34.73
Open-Source LVLMs for Images														
MiniCPM-V-2	78.89	50.84	72.93	26.61	75.00	65.52	66.02	58.00	32.40	50.00	38.40	32.80	22.80	39.07
LLaVA-v1.5-7B	69.90	56.98	70.17	25.81	67.07	53.45	61.38	57.20	24.40	41.60	28.40	26.40	20.40	33.07
InternVL-Chat-v1.5	88.58	73.18	80.94	58.06	85.98	80.46	79.95	70.40	52.80	65.20	58.40	56.00	39.60	57.07
Idefics2-8B	81.31	65.36	73.20	41.94	80.49	76.44	72.29	66.00	42.40	61.60	49.60	40.00	37.20	49.47
Phi-3-Vision	78.89	61.45	76.80	47.58	79.27	74.14	72.29	60.00	38.80	59.20	45.20	42.40	40.80	47.73

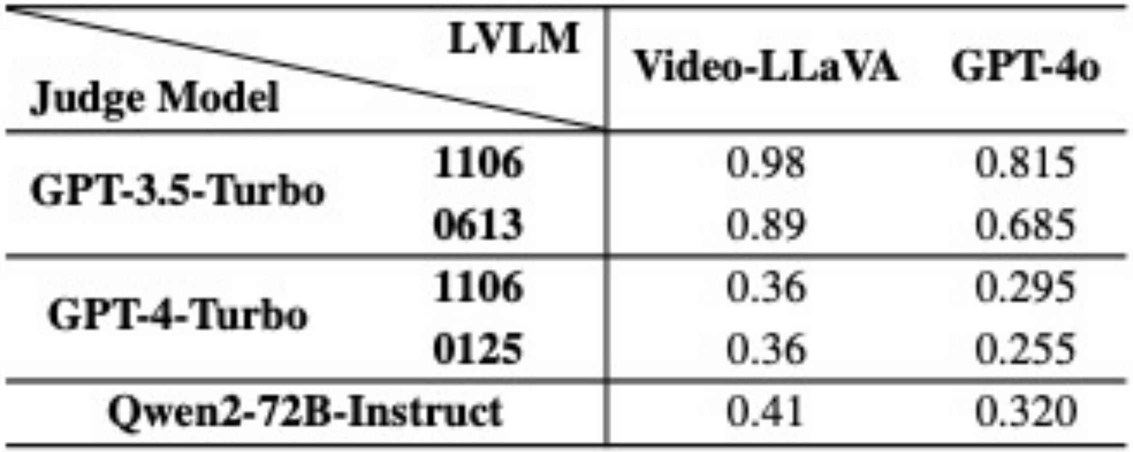
Table 4: **Comparison of Image Models and Video Models on MMBench and MMStar.** We follow the official practice to perform evaluation on these two benchmarks. For MMBench, we report the results on MMBench-DEV-EN-v1.1. We adopt the abbreviations for capabilities that are defined in the original papers.

The Superior Performance of GPT-4 as a Judge



Judge Model \ LVLM		Video-LLaVA	GPT-4o
GPT-3.5-Turbo	1106	2.09	2.45
	0613	1.80	2.11
GPT-4-Turbo	1106	1.05	1.62
	0125	0.90	1.61
Qwen2-72B-Instruct		1.15	1.80

Table 6: **Evaluation results obtained with different GPT judges on MMBench-Video.** The overall mean scores are reported.



Judge Model \ LVLM		Video-LLaVA	GPT-4o
GPT-3.5-Turbo	1106	0.98	0.815
	0613	0.89	0.685
GPT-4-Turbo	1106	0.36	0.295
	0125	0.36	0.255
Qwen2-72B-Instruct		0.41	0.320

Table 7: **The mean absolute error (MAE) of different GPT Judges with human preferences on a randomly selected subset.**

NEURAL INFORMATION PROCESSING SYSTEMS

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Thanks for your attention!

VLMEvalKit
MM' 24

OpenVLM Video Leaderboard

Prism
NeurIPS' 24

MMBench-Video
NeurIPS' 24
D&B Track