Zhejiang University

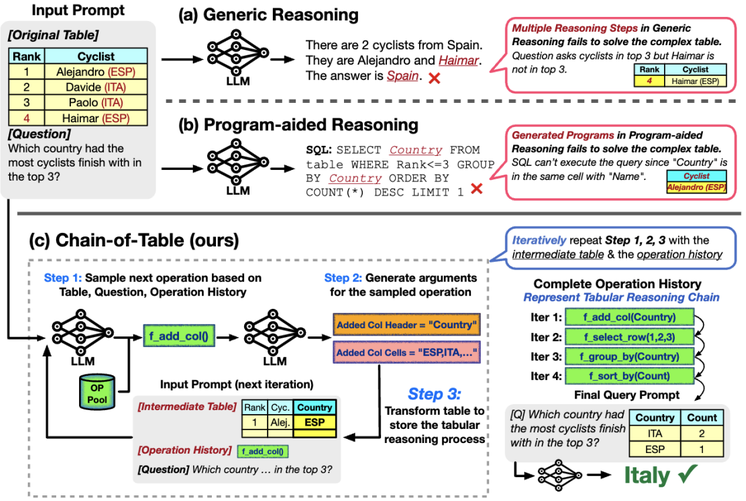
Better and Longer Video Understanding

ENXIN SONG

6 Sep, 2025

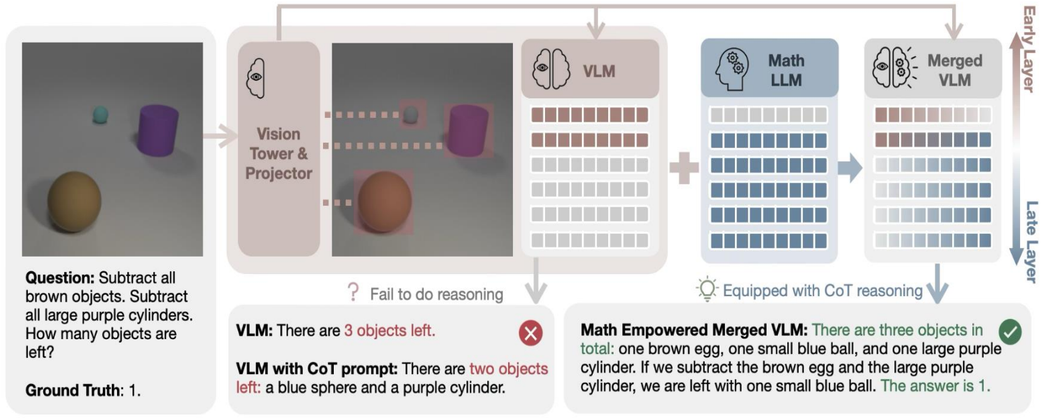
Reasoning Models

Text Reasoning



Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding, ICLR 2024

Vision Reasoning

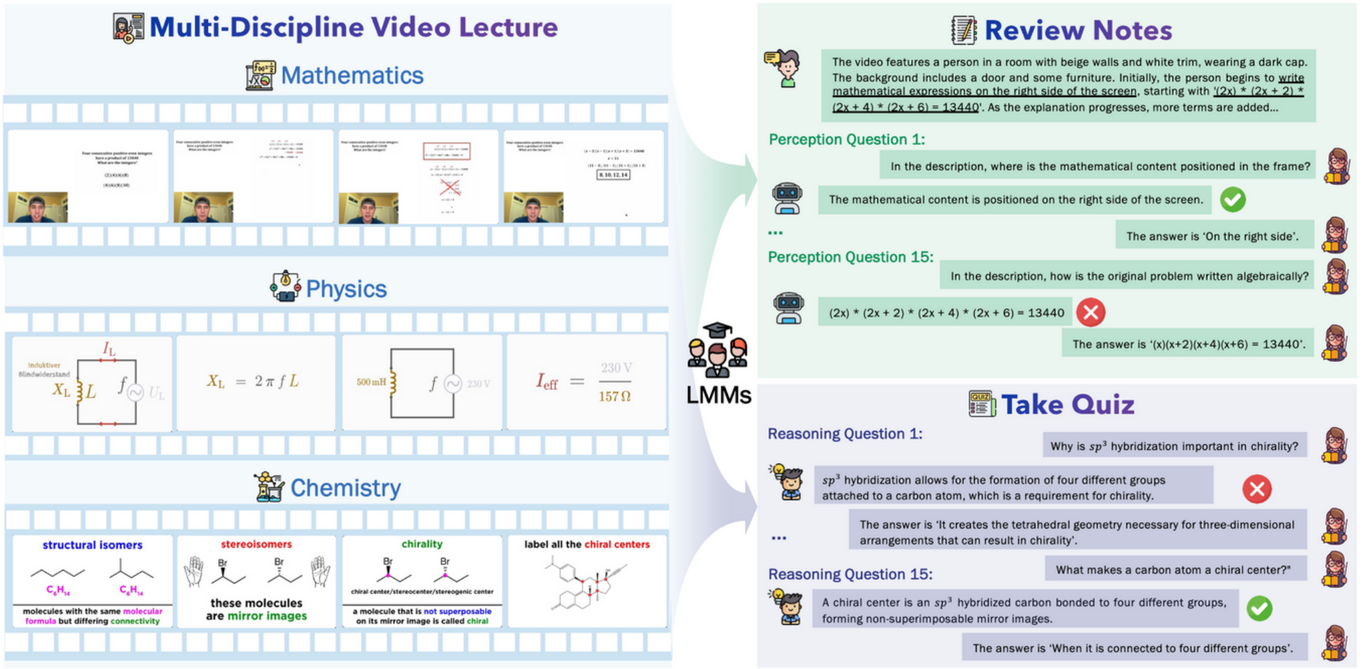


Bring Reason to Vision: Understanding Perception and Reasoning through Model Merging, ICML 2025

Lecture Video Understanding

Can video LLMs really understand real-world lectures?

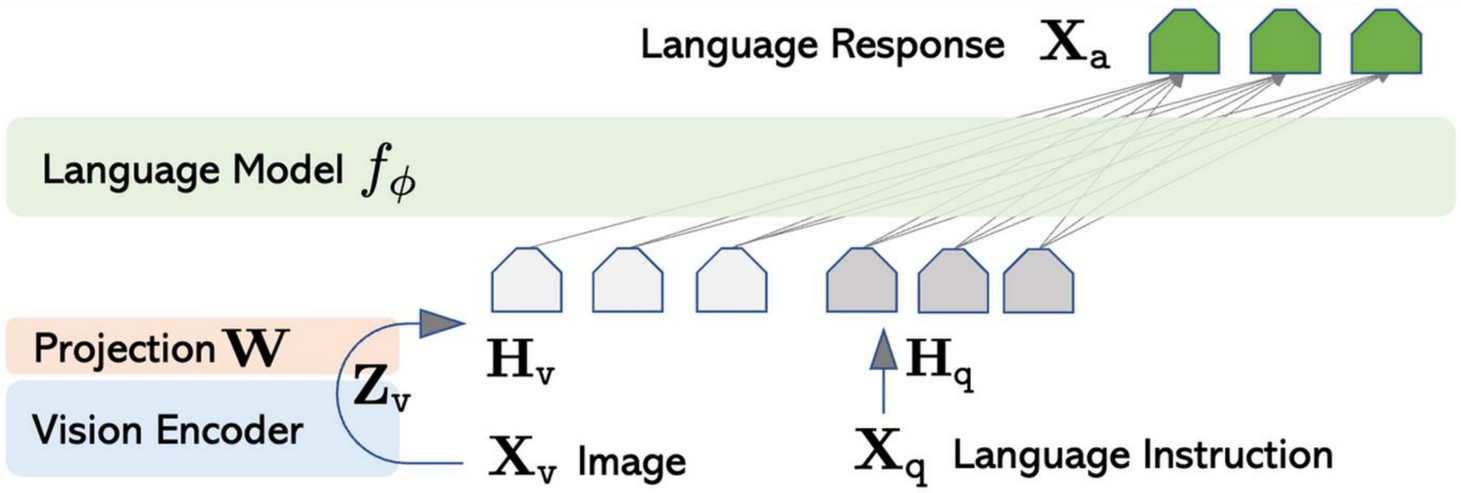
NOT YET.



Video LLMs

How We Connect?

- Connect ViT and LLM
- Adapt from Image LLMs
- Handle longer sequences
- May need more compute
- But less data



Video LLMs

Short videos, short captions — can they tell the whole story?

Figure: Video example of MSR-VTT, which is a widely used video question answering and captioning benchmark.
Labeled caption: *Teams are playing soccer.*

Video-MMLU

Video-MMLU pushes LMMs to the limits

Dataset	Theme	# Video	# Ave. Duration (s)	Caption				Question-answering	
				Number	# Word	# Vocab.	Ave. Length	Number	Type
MovieChat-1K [112]	Movie	1,000	564	1,000	121,077	102,988	121	13,000	Open-ended
MMWorld [61]	Professional	1,910	107	1,910	-	-	66	6,627	Multiple-choice
MLVU [176]	Open	1,730	930	247	-	-	-	3,102	Multiple-choice
MVBench [2]	Open	4,000	16			×		4,000	Multiple-choice
LongVideoBench [133]	Open	3,763	473			×		6,678	Multiple-choice
TempCompass [93]	Open	410	< 30			×		7,540	Multiple-choice
Video-MMMU [65]	Professional	300	506			×		900	Multiple-choice
VATEX [127]	Open	41,250	10	41,250	4,994,768	44,103	15		×
VDC [21]	Open	1,027	28	1,027	515,441	20,419	501		×
LongCaptioning [131]	Open	10,000	93	10,000	-	-	1,198		×
Video-MMLU (ours)	Professional	1,065	109	1,065	520,679	27,613	489	15,746	Open-ended

Table: Benchmark comparison for video understanding tasks.

Video-MMLU

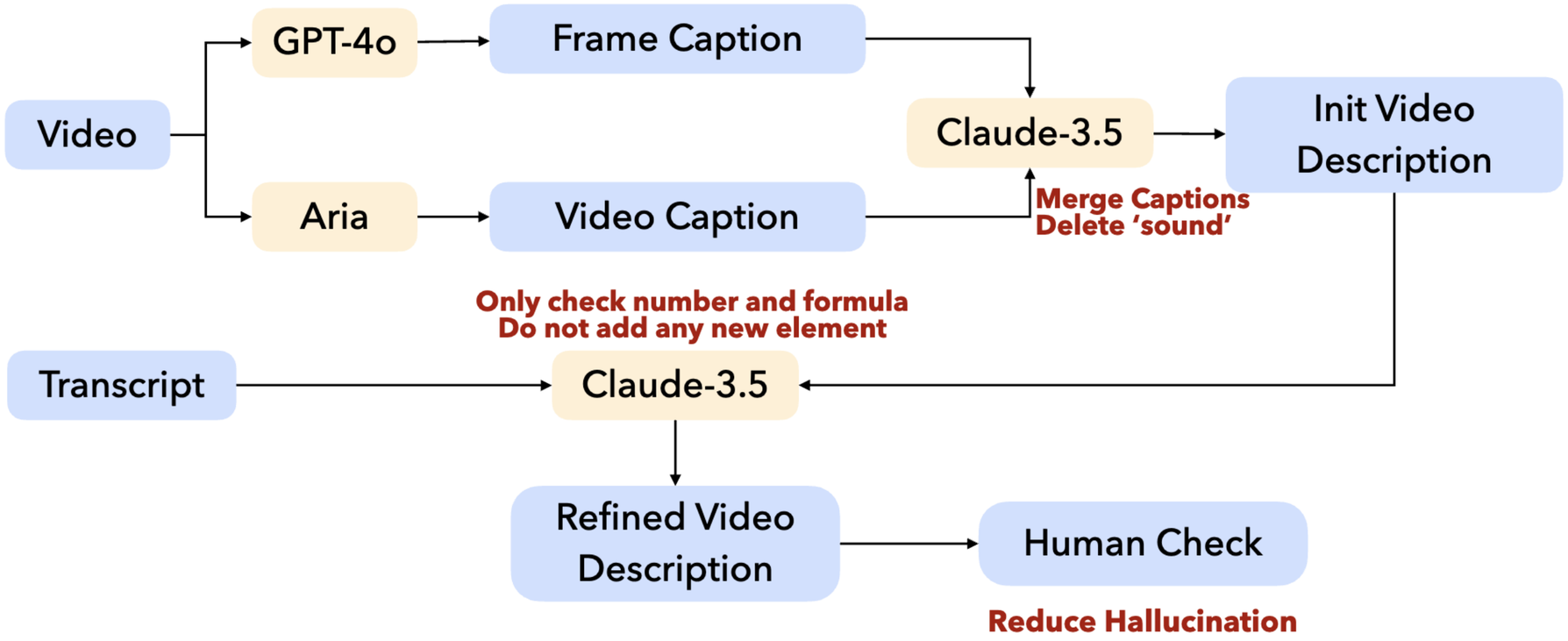
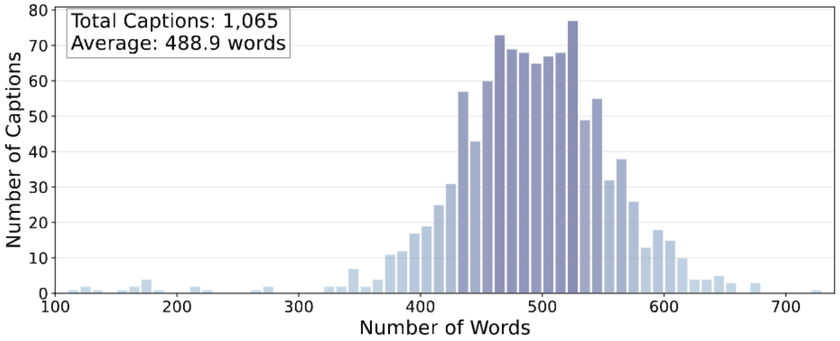


Fig. B1: **Video-MMLU** construction pipeline.

Video-MMLU



(a) Video detailed captions length distribution.

< 30s	30s~60s	60s~120s	120s~180s	> 180s
7.6%	30.6%	17.2%	26.8%	17.8%

(b) Video length duration.

< 10frm	10frm~20frm	20frm~30frm	30frm~40frm	> 40frm
8.0%	41.5%	19.1%	19.1%	12.4%

(c) Keyframes number distribution.

Fig. 2: Visualization of datasets statistics.

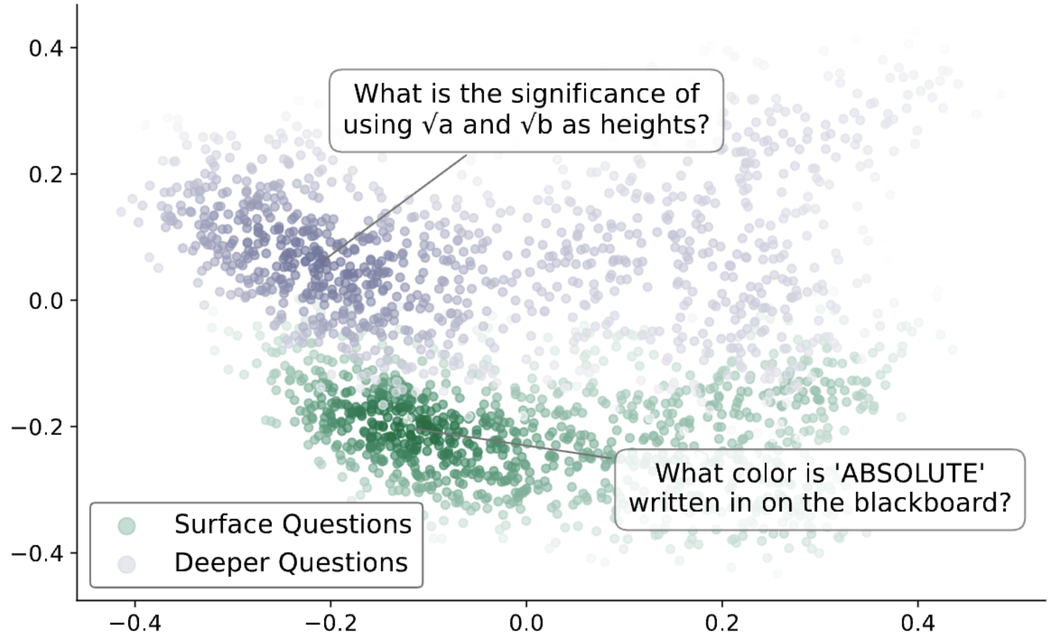


Fig. 3: The text embedding space distribution of surface perception questions in green and deeper reasoning questions in purple.

Video-MMLU

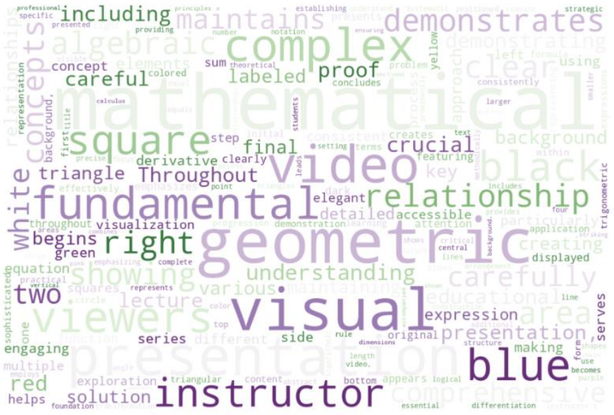
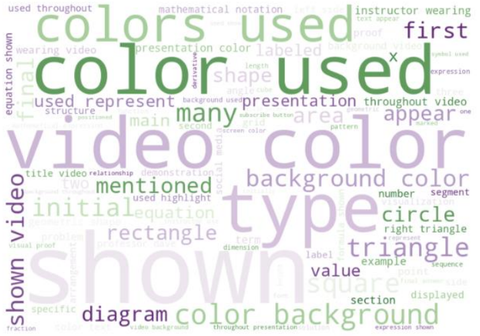
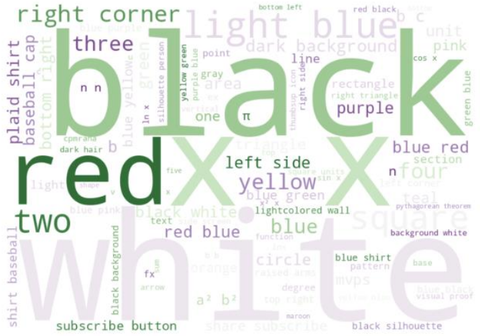


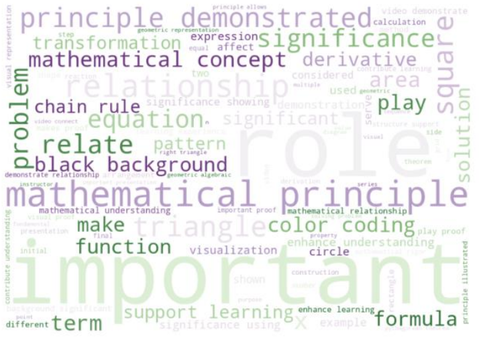
Fig. E2: Word cloud of lecture detailed captions in **Video-MMLU**, showing the diversit



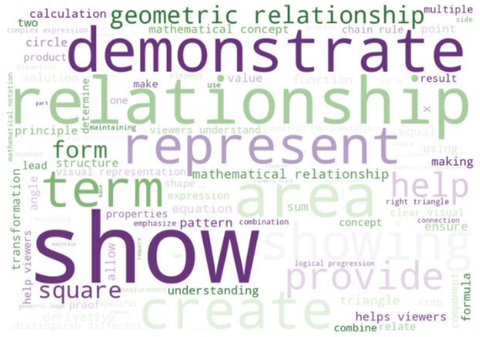
(a) Surface Question



(b) Surface Answer



(c) Deeper Question

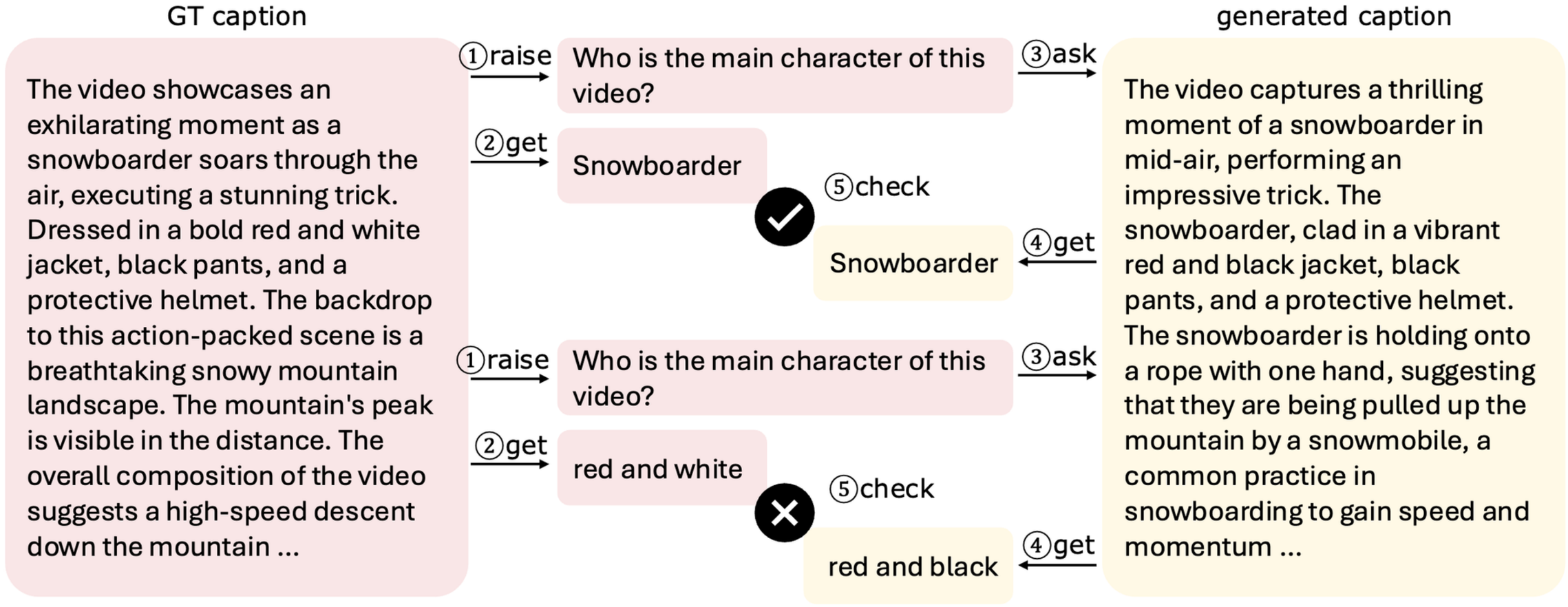


(d) Deeper Answer

Fig. E3: Word cloud of different question-answering pairs in **Video-MMLU**, showing the diversity.

Evaluation Metric

VDC



Video-MMLU

- 3 Proprietary Models
- 78 Open-Source LMMs
- 6 Vision-Blind Baselines
- 9 Vision Token Compression Models
- 4 Omni Models

How the base LLMs influence performance？

How the vision token influence performance？

Video-MMLU

Models	LLM	Size	Overall	Notebook Avg.	Notebook Math	Notebook Physics	Notebook Chemistry	Quiz Avg.	Quiz Math	Quiz Physics	Quiz Chemistry
Vision-Blind Baselines											
-	-	0.5B	4.29	2.25	2.31	4.44	0.01	6.33	4.92	8.88	5.19
-	-	1.5B	16.49	7.66	5.79	8.88	8.33	25.33	20.39	26.92	28.67
Qwen2.5 [109]	-	3B	21.64	11.31	11.88	14.24	7.81	31.97	27.58	32.12	36.23
	-	7B	22.34	10.02	6.08	12.34	11.66	34.66	33.17	34.65	33.16
	-	32B	24.76	13.65	9.85	17.45	13.66	35.87	34.14	36.15	37.33
	-	72B	24.99	9.44	8.88	12.77	6.69	40.54	37.55	39.76	43.31
Proprietary Models											
Gemini-1.5-Flash	-	-	43.63	39.46	27.69	53.36	37.33	47.77	44.36	67.51	31.43
GPT-4o	-	-	49.41	53.89	55.23	56.12	50.33	44.93	33.08	75.91	25.79
Claude-3.5-sonnet	-	-	69.34	67.43	63.74	65.91	72.66	71.24	68.29	77.64	67.80
Open-Source LMMs (~5B)											
SmolVLM-256M [38]	SmolLM2 [7]	135M	8.95	15.41	11.87	16.14	18.24	2.50	1.62	2.50	3.40
VILA1.5-3B [68]	Qwen2 [93]	1.5B	9.61	18.71	19.65	15.83	20.66	0.51	1.36	0.13	0.05
SmolVLM-500M [38]	SmolLM2 [7]	360M	11.05	17.24	11.29	21.75	18.68	4.86	3.65	7.14	3.81
SmolVLM [38]	SmolLM2 [7]	1.7B	14.14	17.25	14.91	20.00	16.86	11.03	6.09	15.71	11.29
DeepSeek-VL-1.3B [74]	DeepSeek-LLM [14]	1.3B	15.28	20.59	18.30	20.51	22.98	9.98	10.35	8.57	11.02
InternVL2-2B [26]	InternLM2 [16]	1.8B	15.60	24.61	22.44	26.31	25.09	6.59	5.68	7.85	6.25
Mini-InternVL-Chat-2B-V1.5 [26]	InternLM2 [16]	1.8B	16.12	21.19	22.32	20.00	21.25	11.05	10.15	10.35	12.65
XinYuan-VL-2B [32]	Qwen2 [93]	1.5B	17.58	29.65	25.08	32.63	31.26	5.52	4.26	6.07	6.25
InternVL2-1B [26]	Qwen2 [93]	0.5B	18.59	26.59	22.77	26.66	30.34	10.59	10.35	10.00	11.43
Qwen2-VL-2B [93]	Qwen2 [93]	1.5B	19.33	30.19	28.67	32.98	28.92	8.47	7.10	10.71	7.62
LLaVA-OneVision-OV [61]	Qwen2 [93]	0.5B	19.60	23.77	22.42	20.89	28.01	15.43	14.82	13.92	17.55
XComposer2-1.8B [36]	InternLM2 [16]	1.8B	19.77	20.79	14.21	23.85	24.32	18.76	13.60	19.28	23.40
InternVL2-4B [26]	Phi3-mini [1]	3.8B	20.44	27.44	26.28	30.87	25.19	13.45	11.16	17.50	11.70
Qwen2.5-VL-3B [13]	Qwen2.5 [109]	3B	22.40	31.06	31.20	32.63	29.36	13.74	10.05	17.85	13.33
Aquila-VL-2B [45]	Qwen2.5 [109]	1.5B	23.94	13.78	13.14	15.08	13.14	34.10	30.45	36.07	35.78
Apollo-1.5B [130]	Qwen2.5 [109]	1.5B	25.89	26.43	26.32	21.66	31.33	25.35	26.02	20.01	30.03
Apollo-3B [130]	Qwen2.5 [109]	3B	27.27	33.26	32.30	30.83	36.66	21.28	17.12	26.66	20.07
InternVL2.5-1B [25]	Qwen2.5 [109]	0.5B	27.57	31.71	26.97	34.38	33.79	23.43	22.84	23.92	23.53
InternVL2.5-2B [25]	InternLM2.5 [16]	1.8B	28.62	33.26	27.94	34.02	37.83	23.99	22.43	23.57	25.98
Phi-3-Vision [1]	Phi3-mini [1]	3.8B	28.69	21.85	21.88	23.85	19.84	35.54	25.98	41.07	39.59
SAIL-VL-2B [35]	Qwen2.5 [109]	1.5B	28.86	25.65	23.27	25.96	27.74	32.08	27.71	33.57	34.96
Phi-3.5-Vision [1]	Phi3.5-mini [1]	3.8B	34.39	29.55	23.20	32.38	33.09	39.23	35.32	40.35	42.04
Mini-InternVL-Chat-4B-V1.5 [26]	Phi3-mini [1]	3.8B	39.98	25.76	23.71	30.17	23.42	54.20	45.27	61.42	55.91
InternVL2.5-4B [25]	Qwen2.5 [109]	3B	40.74	36.75	31.35	36.30	42.61	44.74	41.82	46.42	45.98
Aria [62]	66 Experts MoEs	3.9B	42.87	45.09	41.45	45.83	48.0	40.65	39.17	42.66	40.12
Open-Source LMMs (~8B)											
XComposer [123]	InternLM [36]	7B	10.91	20.52	12.96	23.50	25.10	1.29	1.42	0.71	1.76
InstructBLIP-7B [33]	Vicuna [29]	7B	11.06	19.26	14.54	21.75	21.49	2.86	1.11	4.64	2.85
Mantis-8B-Fuyu [56]	Fuyu [4]	8B	12.31	16.50	12.41	17.19	19.91	8.12	6.09	8.21	10.06
Cambrian-8B [89]	LLaMA3 [6]	8B	12.68	20.17	20.38	21.75	18.38	5.19	3.35	6.78	5.44
Mantis-8B-siglip-llama3 [56]	LLaMA3 [6]	8B	13.74	23.95	14.12	28.42	29.33	3.54	2.33	1.78	6.53
Mantis-8B-Idefics2 [56]	Mistral [5]	7B	14.19	21.41	16.81	23.50	23.94	6.98	6.70	6.78	7.48
LLaVA-1.5 [70]	Vicuna [29]	7B	15.71	22.31	15.30	22.81	28.84	9.11	7.81	8.92	10.61
Video-LlaVA-7B [67]	Vicuna [29]	7B	15.89	15.32	11.15	16.14	18.67	16.47	13.29	17.5	18.63
VideoChat2-HD [64]	Mistral [5]	7B	16.74	18.07	15.63	19.65	18.94	15.40	12.79	13.15	20.26
Mantis-8B-clip-llama3 [56]	LLaMA3 [6]	8B	18.78	21.62	14.37	24.56	25.95	15.95	13.29	13.21	21.36
Video-ChatGPT [77]	Vicuna [29]	7B	19.37	16.30	10.88	15.78	22.25	22.45	19.08	20.00	28.29
LLaVA-NeXT [63]	Vicuna [29]	7B	21.46	18.06	9.79	21.40	22.99	24.87	21.31	22.85	30.47
Qwen-VL [12]	Qwen [11]	7B	21.98	24.35	19.56	25.61	27.88	19.62	16.95	16.07	25.85
mPLUG-Owl3 [113]	Qwen2 [93]	7B	22.59	22.55	18.61	24.91	24.13	22.64	18.57	25.00	24.35
ShareGPT4V-7B [20]	Vicuna [29]	7B	22.81	23.48	18.99	25.16	26.30	22.15	17.83	22.24	26.40
LLaVA-NeXT [63]	LLaMA3 [6]	8B	23.29	16.53	8.97	20.00	20.64	30.05	24.06	32.50	33.60
PLLaVA [107]	Vicuna [29]	7B	23.85	16.08	12.75	18.18	17.33	31.63	21.55	41.44	31.91
InternVL2-8B [26]	InternLM2.5 [16]	7B	24.06	31.43	26.12	33.33	34.85	16.69	13.19	21.78	15.10
DeepSeek-VL-7B [74]	DeepSeek-LLM [14]	7B	24.12	26.20	25.62	25.65	27.33	22.04	20.50	18.57	27.07
VILA1.5-8B [68]	LLaMA3 [6]	8B	24.20	27.95	25.38	25.83	32.66	20.45	14.38	26.73	20.24
XComposer2 [36]	InternLM2 [16]	7B	25.62	16.24	12.68	17.89	18.16	35.00	26.90	38.92	39.18
LLaVA-NeXT [63]	Mistral [5]	7B	25.83	20.31	18.48	21.05	21.42	31.45	26.09	33.57	34.69
Qwen2-VL-7B [93]	Qwen2 [93]	7B	28.83	34.22	27.58	35.37	39.72	23.44	19.59	24.07	26.66
LLaVA-NeXT-Video-7B [125]	Qwen2 [93]	7B	31.55	35.75	30.03	36.69	40.54	27.35	29.32	24.07	28.66
LLaVA-OneVision-OV [61]	Qwen2 [93]	7B	33.99	34.55	29.53	35.66	38.46	33.44	30.35	35.71	34.28
Apollo-7B [130]	Qwen2.5 [109]	7B	36.78	38.22	33.50	39.16	42.00	35.33	29.45	26.56	49.98
Qwen2.5-VL-7B[13]	Qwen2.5 [109]	7B	37.47	42.02	39.43	44.91	41.73	32.93	24.36	41.78	32.65
Valley-Eagle [103]	Qwen2.5 [109]	7B	37.96	39.22	33.21	41.40	43.07	36.71	28.40	37.14	44.58
MiniCPM-V 2.6 [112]	Qwen2 [93]	7B	39.31	43.57	36.13	47.01	47.59	35.06	28.79	37.66	38.73
InternVL2.5-8B [25]	InternLM2.5 [16]	7B	39.51	34.51	30.48	34.53	38.52	44.51	39.91	46.29	47.33
MiniCPM-o 2.6 [112]	Qwen2.5 [109]	7B	44.89	54.83	47.86	57.54	59.10	34.95	39.28	22.85	42.72

Table 3. Results on **Video-MMLU** including proprietary models, and open-source LMMs (<40B), across overall performance, detailed captioning (Notebook), and reasoning QA (Quiz) in different disciplines. Darker shades indicate better performance.

Models	LLM	Size	Overall	Notebook Avg.	Notebook Math	Notebook Physics	Notebook Chemistry	Quiz Avg.	Quiz Math	Quiz Physics	Quiz Chemistry
Proprietary Models											
Gemini-1.5-Flash	-	-	43.63	39.46	27.69	53.36	37.33	47.77	44.36	67.51	31.43
GPT-4o	-	-	49.41	53.89	55.23	56.12	50.33	44.93	33.08	75.91	25.79
Claude-3.5-sonnet	-	-	69.34	67.43	63.74	65.91	72.66	71.24	68.29	77.64	67.80
Open-Source LMMs (~20B)											
LLaVA-NeXT [63]	Vicuna [29]	13B	8.13	16.27	9.95	22.55	16.32	0.0	0.0	0.0	0.0
ShareGPT4V-13B [20]	Vicuna [29]	13B	11.57	18.37	17.13	19.31	18.69	4.78	4.06	5.00	5.30
Cambrian-13B [89]	Vicuna [29]	13B	14.56	21.77	20.14	24.21	20.97	7.36	4.16	10.71	7.21
VILA1.5-13B [68]	Vicuna [29]	13B	15.71	24.95	24.21	22.66	28.00	6.48	2.73	6.68	10.05
InstructBLIP-13B [33]	Vicuna [29]	13B	15.89	22.32	14.96	26.51	25.49	9.47	5.17	15.35	7.89
InternVL-Chat-V1-1 [26]	LLaMA2 [90]	13B	21.53	24.83	22.30	26.31	25.88	18.22	13.29	21.78	19.59
LLaVA-1.5 [70]	Vicuna [29]	13B	21.58	16.74	12.75	21.75	15.72	26.42	23.14	26.07	30.06
OmChat-v2.0-13B [127]	Qwen2 [93]	7B	21.91	24.57	21.26	25.61	26.85	19.26	15.32	20.71	21.76
PLLaVA-13B [107]	Vicuna [29]	13B	26.25	21.08	17.75	21.42	24.07	31.43	25.27	28.21	40.81
InternVL-Chat-V1-5 [26]	InternLM [36]	20B	28.76	26.00	21.35	26.31	30.34	31.53	32.33	29.62	32.66
CogVLM2-LLaMA3-Chat-19B [52]	LLaMA3 [6]	8B	31.99	24.08	21.33	24.91	26.01	39.90	32.58	41.42	45.71
Open-Source LMMs (~40B)											
Cambrian-34B [89]	Nous-Hermes-2-Yi [80]	34B	12.73	19.90	20.52	22.10	17.08	5.56	4.06	6.78	5.85
InternVL2-26B [26]	InternLM2 [16]	20B	21.33	29.68	26.19	28.77	34.01	12.98	13.84	11.11	14.00
InternVL-Chat-V1-2-Plus [26]	Nous-Hermes-2-Yi [80]	34B	24.25	18.88	21.23	21.05	14.36	29.62	22.13	38.57	28.16
InternVL2-40B [26]	Nous-Hermes-2-Yi [80]	34B	27.44	32.74	28.67	33.58	35.99	22.15	19.79	25.71	20.95
InternVL-Chat-V1-2 [26]	Nous-Hermes-2-Yi [80]	34B	29.00	21.42	23.14	26.66	14.47	36.58	23.85	50.00	35.91
VILA1.5-40B [68]	Nous-Hermes-2-Yi [80]	34B	30.72	32.30	29.91	28.33	38.66	29.13	31.51	20.15	35.72
PLLaVA-34B [107]	Nous-Hermes-2-Yi [80]	34B	30.91	21.09	20.53	22.10	20.65	40.74	30.62	47.22	44.38
LLaVA-NeXT [63]	Nous-Hermes-2-Yi [80]	34B	34.16	25.07	23.58	25.72	25.93	43.25	21.42	58.33	50.00
LLaVA-NeXT [63]	Qwen1.5 [10]	32B	40.43	26.98	23.10	22.96	34.90	53.88	46.10	55.55	60.00
InternVL2.5-26B [25]	InternLM2.5 [16]	20B	44.39	39.71	32.90	47.91	38.33	49.07	47.21	50.00	50.00
InternVL2.5-38B [25]	Qwen2.5 [109]	32B	49.35	37.01	34.25	35.13	41.66	61.68	60.47	59.25	65.33

Video-MMLU

Proprietary models consistently outperform open-source models.

Lecture understanding in models relies more on textual content in frames than on animations.

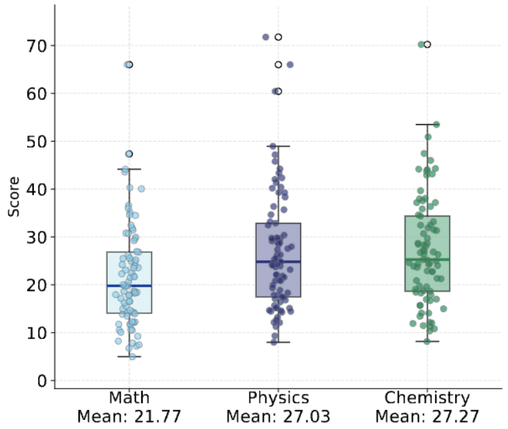
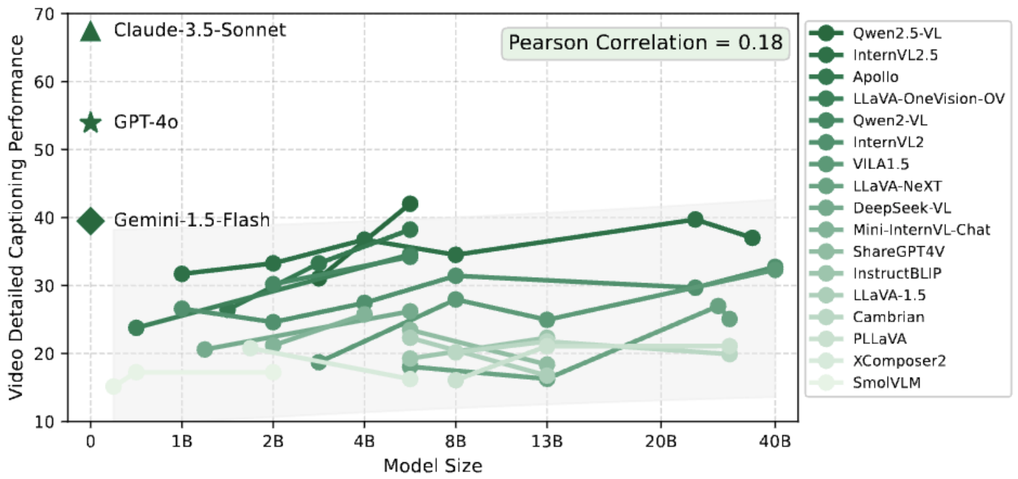


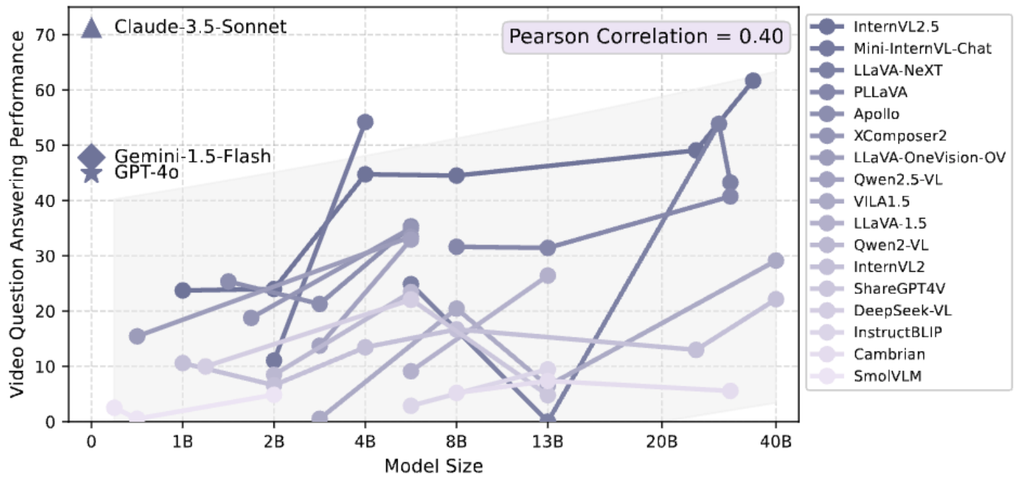
Fig. 6: Score distribution.

Video-MMLU

Open-source video LMMs may not exhibit clear advantages over image LMMs.

Large scale LMMs do not show clear advantages over smaller ones.





Video-MMLU

Larger LLMs enhance lecture understanding but with diminishing returns.

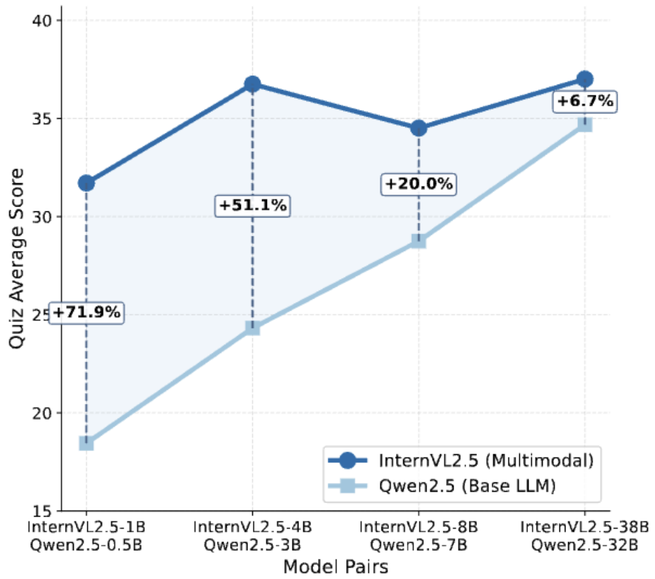
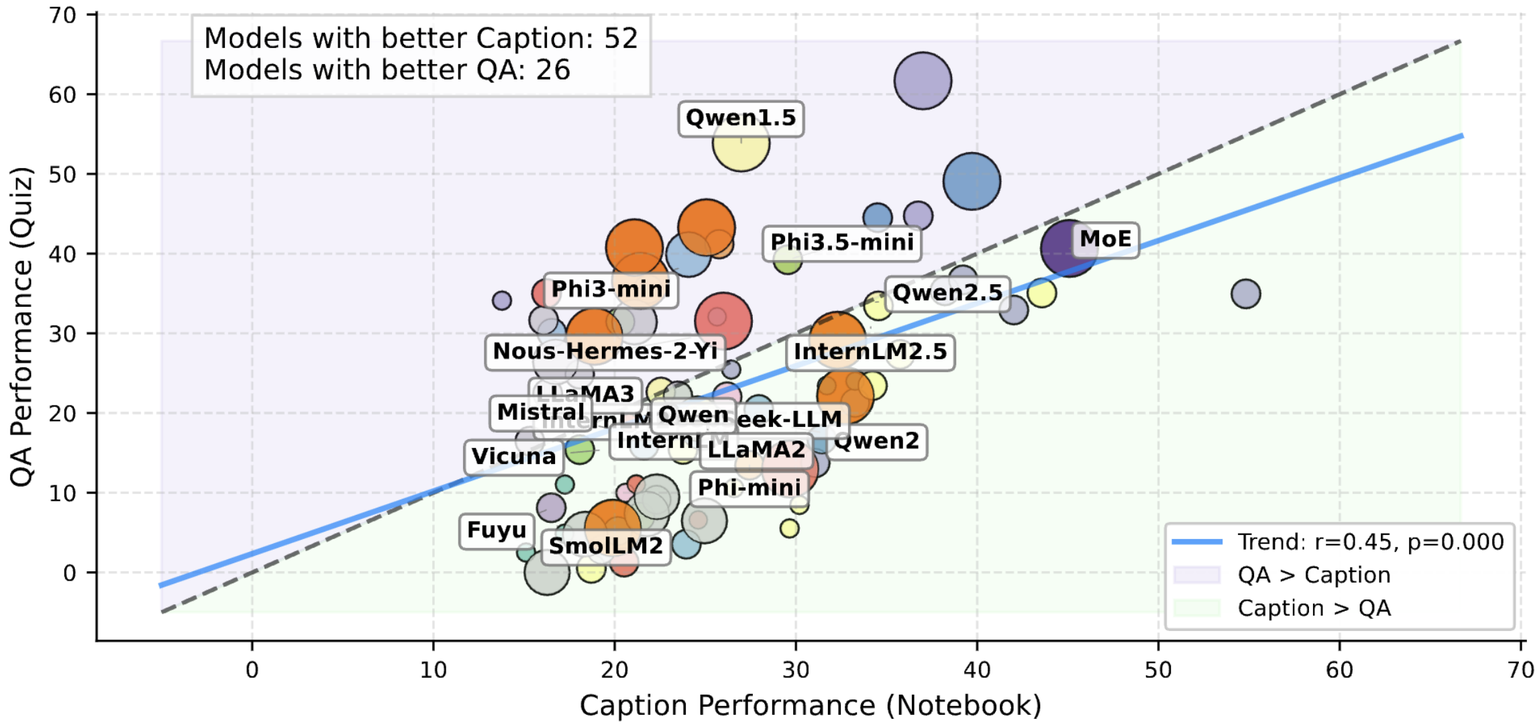


Fig. 8: Impact of LLM size.

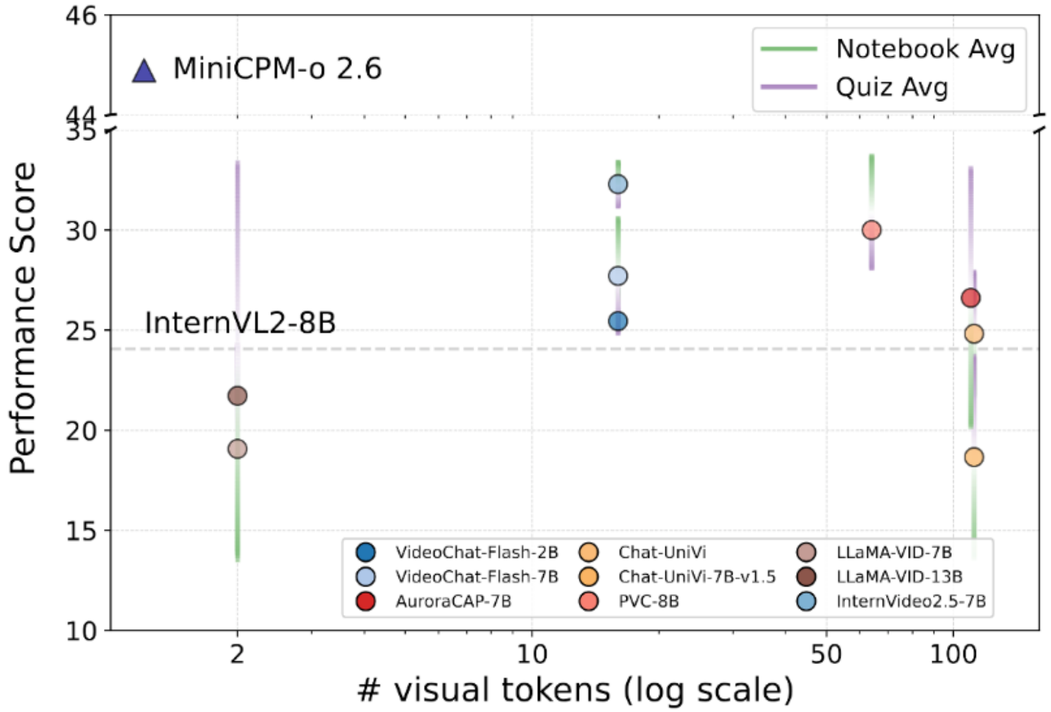
Video-MMLU

LLM Architecture shapes LMMs' balance between perception and reasoning.



Video-MMLU

Can LMMs with visual token compression sustain strong performance in complex, context-rich lecture understanding tasks like Video-MMLU?



Video-MMLU

Can LMMs with visual token compression sustain strong performance in complex, context-rich lecture understanding tasks like Video-MMLU?

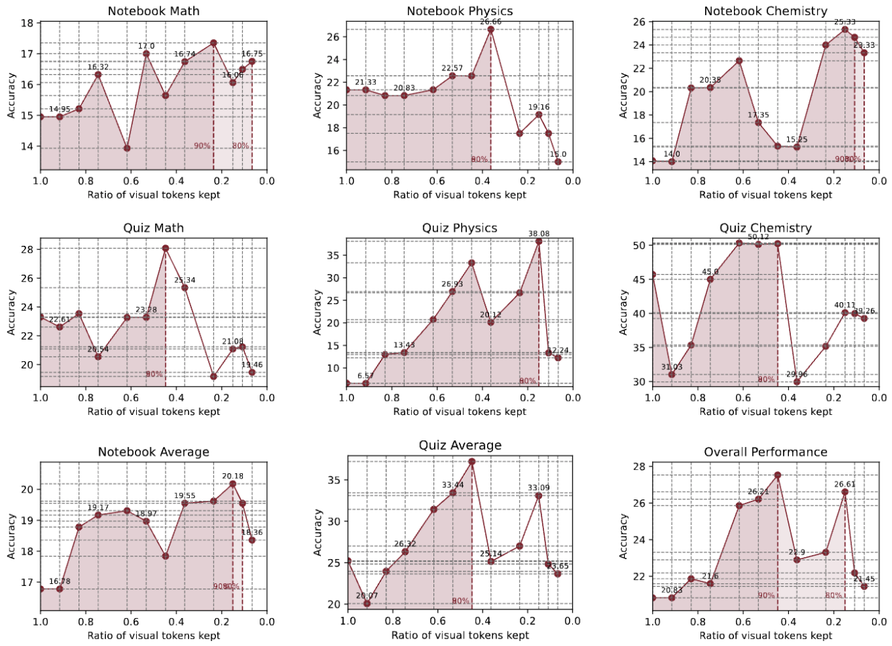


Fig. J5: Ablation study of token merging in AuroraCap [21] on **Video-MMLU**.

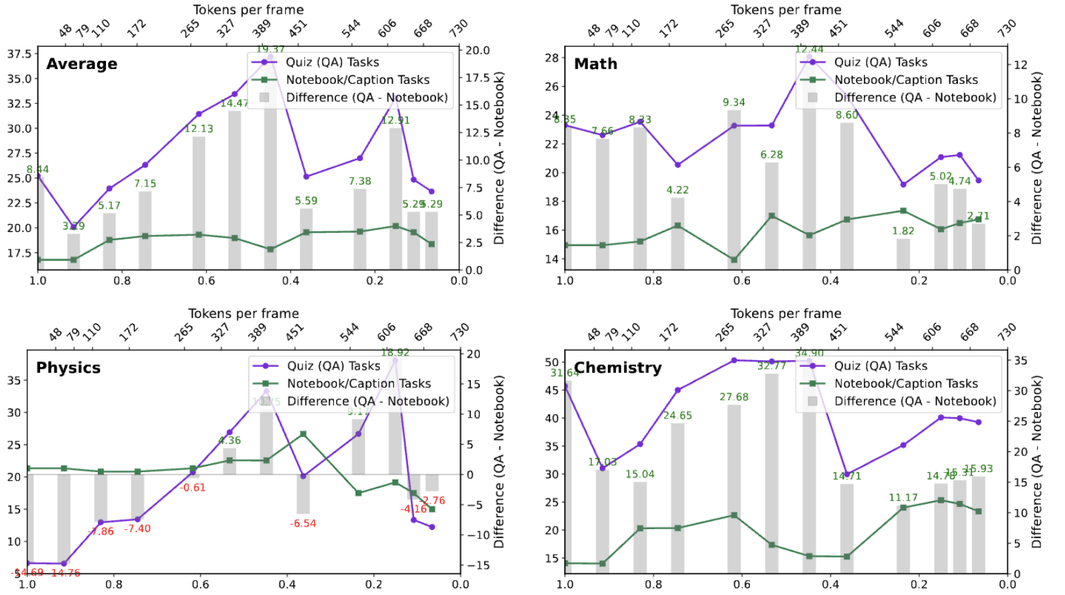


Fig. J6: Performance comparsion of token merging in AuroraCap [21] on **Video-MMLU**across different discipline.

Video-MMLU

- Video-MMLU: https://arxiv.org/abs/2504.14693
- Github: https://github.com/Espere-1119-Song/Video-MMLU
- Benchmark: https://huggingface.co/datasets/Enxin/Video-MMLU
- Eval Code: https://github.com/open-compass/VLMEvalKit
- Project Page: https://enxinsong.com/Video-MMLU-web/
- Workshop Page: https://sites.google.com/view/loveucvpr24/track1b

Video-MMLU

Multimodal Video Agent Workshop hosted in CVPR 2025.

Track 1B: Multi-Discipline Lecture Understanding Challenge

- This track evaluates models' ability to understand and reason over lecture videos in mathematics, physics, and chemistry.
- The task consists of **video detailed captioning**(Review Notes) and **question answering** (Take Quiz), both required for final evaluation.
- The benchmark is based on the Video-MMLU.
- The **final score** is the average of the captioning and QA scores.
- The **top winner** will be recognized at the workshop and awarded official certificates.
- **If you encounter computational resource limitations during the evaluation, we can assist with testing.**

Model	Size	#F	#T	Overall	Notebook	Quiz
SilVR	7B	32	18.75	80.41	78.8	82.0
Claude-3.5-sonnet-20241022	-	32	-	69.34	67.4	71.2
GPT-4o-2024-05-13	-	32	-	49.41	53.9	44.9
InternVL2.5-38B	38B	32	-	49.35	37.0	61.7
TW-GRPO	7B	32	-	48.9	42.1	55.7
FlexSelect	7B	32	18.75	48.75	32.4	65.1
MiniCPM-o 2.6	8B	32	-	44.89	54.8	35.0
InternVL2.5-26B	26B	32	-	44.39	39.7	49.1
Gemini-1.5-Flash-002	-	32	-	43.63	39.5	47.8
Aria	3.9B	32	-	42.87	45.1	40.6
InternVL2.5-4B	4B	32	-	40.74	36.8	44.7
LLaVA-NeXT-Qwen-32B	32B	16	-	40.43	27.0	53.9
Mini-InternVL-Chat-4B-V1.5	4B	32	-	39.98	25.8	54.2
InternVL2.5-8B	8B	32	-	39.51	34.5	44.5
MiniCPM-V 2.6	7B	32	-	39.31	43.6	35.1

Some of our work on LLM-based video understanding

Long

MovieChat: From Dense Token to Sparse Memory for Long Video Understanding
Computer Vision and Pattern Recognition (CVPR), 2024

MovieChat+: Question-aware Sparse Memory for Long Video Question Answering
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2025

Detail

AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark International Conference on Learning Representations (ICLR), 2025

Knowledge

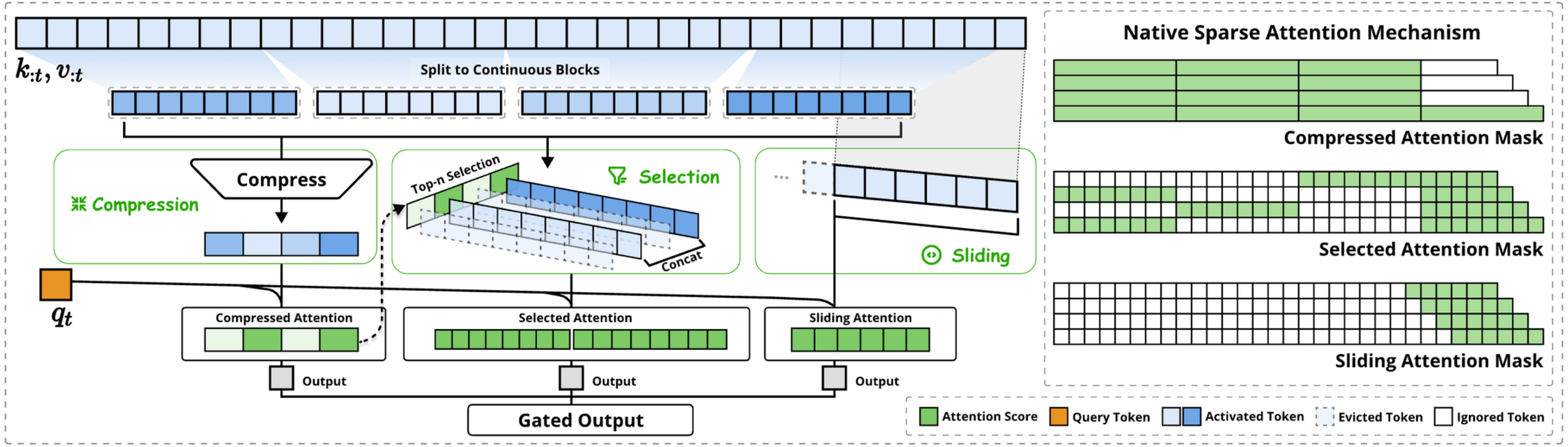
Video-MMLU: A Massive Multi-Discipline Lecture Understanding Benchmark
International Conference on Computer Vision (ICCV) Workshop @ Findings

Efficient

AuroraLong: Bringing RNNs Back to Efficient Open-Ended Video Understanding
International Conference on Computer Vision (ICCV), 2025

Something Ongoing

Can Sparse Attention performs well in Video Understanding? (Figure: DeepSeek NSA)



THANK YOU

02 May, 2024