MLLMs for video modality

Vlad Shakhuro

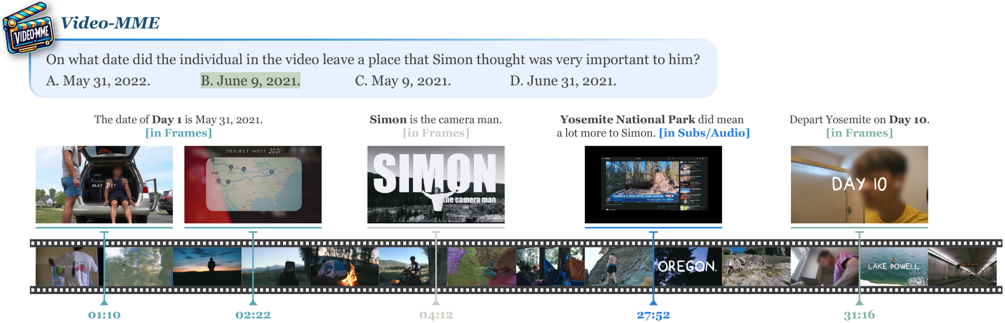
27 March 2025

Outline

1. Tasks and benchmarks

2. Models

Video-MME



Video-MME

On what date did the individual in the video leave a place that Simon thought was very important to him?

A. May 31, 2022.
B. June 9, 2021.
C. May 9, 2021.
D. June 31, 2021.

The date of **Day 1** is May 31, 2021. [in Frames]

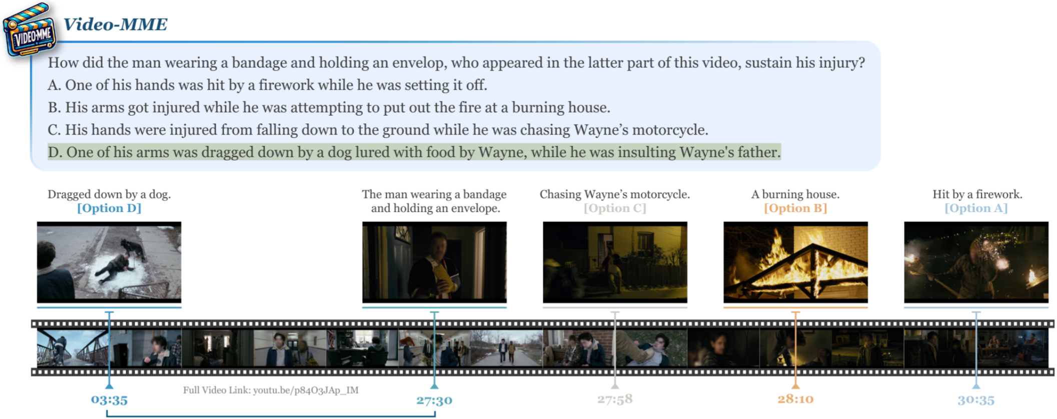
Simon is the camera man. [in Frames]

Yosemite National Park did mean a lot more to Simon. [in Subs/Audio]

Depart Yosemite on **Day 10**. [in Frames]

Fu et al. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. arXiv:2405.21075

Video-MME



Video-MME

How did the man wearing a bandage and holding an envelop, who appeared in the latter part of this video, sustain his injury?

A. One of his hands was hit by a firework while he was setting it off.

B. His arms got injured while he was attempting to put out the fire at a burning house.

C. His hands were injured from falling down to the ground while he was chasing Wayne’s motorcycle.

D. One of his arms was dragged down by a dog lured with food by Wayne, while he was insulting Wayne's father.

Video-MME

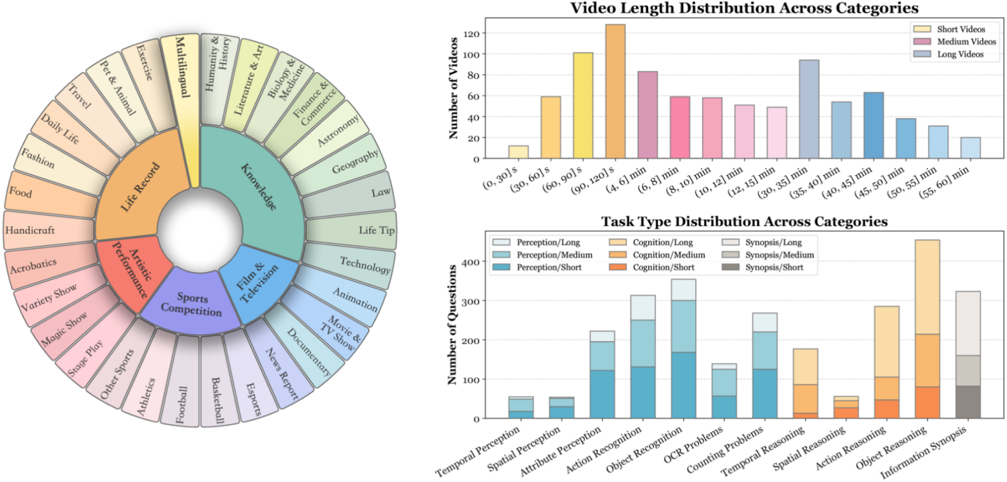
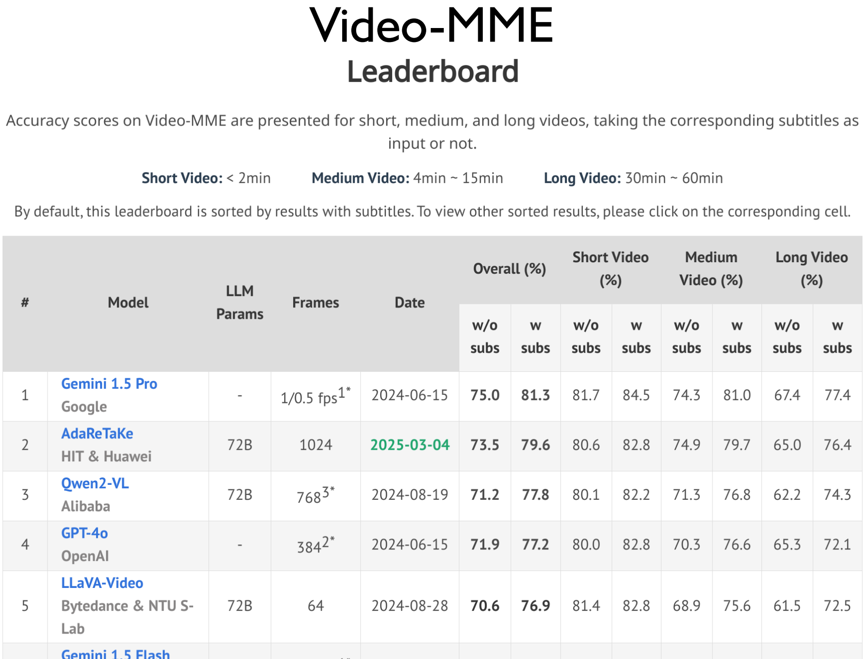


Figure 2: (Left) Video categories. Our benchmark covers 6 key domains and 30 sub-class video types. (Right) Video duration length and question type distributions. Video-MME has a full spectrum of video length and covers different core abilities of MLLMs.



Video-MME

Leaderboard

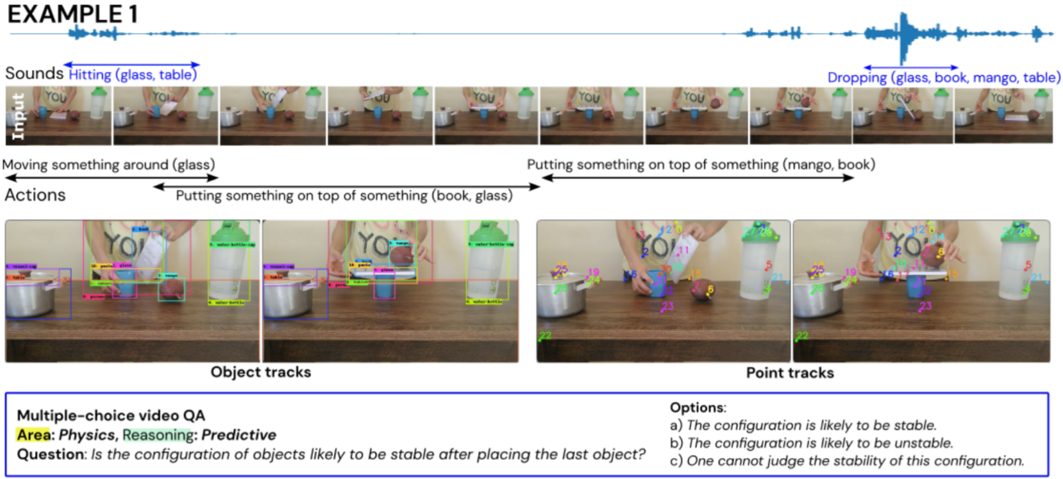
Accuracy scores on Video-MME are presented for short, medium, and long videos, taking the corresponding subtitles as input or not.

Short Video: < 2min **Medium Video:** 4min ~ 15min **Long Video:** 30min ~ 60min

By default, this leaderboard is sorted by results with subtitles. To view other sorted results, please click on the corresponding cell.

#	Model	LLM Params	Frames	Date	Overall (%)		Short Video (%)		Medium Video (%)		Long Video (%)	
					w/o subs	w subs	w/o subs	w subs	w/o subs	w subs	w/o subs	w subs
1	Gemini 1.5 Pro Google	-	1/0.5 fps^{1*}	2024-06-15	**75.0**	**81.3**	81.7	84.5	74.3	81.0	67.4	77.4
2	AdaReTaKe HIT & Huawei	72B	1024	2025-03-04	**73.5**	**79.6**	80.6	82.8	74.9	79.7	65.0	76.4
3	Qwen2-VL Alibaba	72B	768^{3*}	2024-08-19	**71.2**	**77.8**	80.1	82.2	71.3	76.8	62.2	74.3
4	GPT-4o OpenAI	-	384^{2*}	2024-06-15	**71.9**	**77.2**	80.0	82.8	70.3	76.6	65.3	72.1
5	LLaVA-Video Bytedance & NTU S-Lab	72B	64	2024-08-28	**70.6**	**76.9**	81.4	82.8	68.9	75.6	61.5	72.5
	Gemini 1.5 Flash											

PerceptionTest

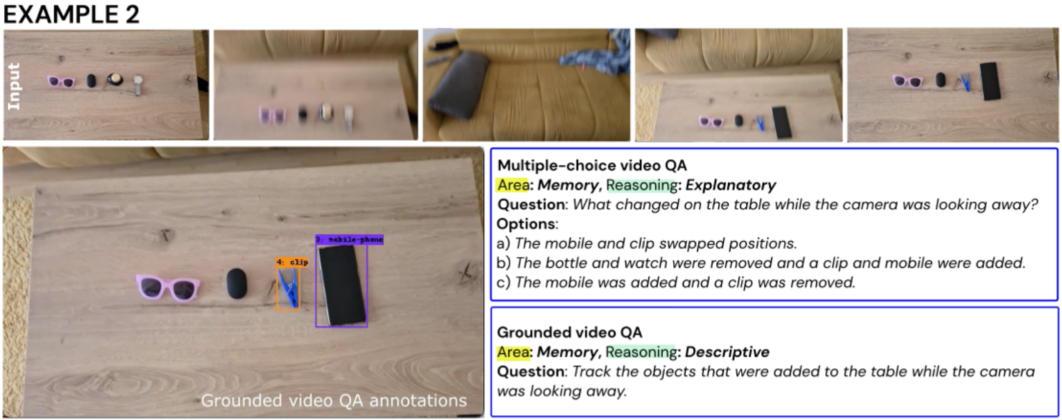

EXAMPLE 1
Sounds
Hitting (glass, table)
Dropping (glass, book, mango, table)
Input
Moving something around (glass)
Putting something on top of something (mango, book)
Actions
Putting something on top of something (book, glass)
Object tracks
Point tracks

Multiple-choice video QA
Area: ***Physics***, **Reasoning**: ***Predictive***
Question: *Is the configuration of objects likely to be stable after placing the last object?*

Options:
a) *The configuration is likely to be stable.*
b) *The configuration is likely to be unstable.*
c) *One cannot judge the stability of this configuration.*

Pătrăucean et al. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. NeurIPS 2023

PerceptionTest



EXAMPLE 2

Multiple-choice video QA

Area: ***Memory***, Reasoning: ***Explanatory***

Question: *What changed on the table while the camera was looking away?*

Options:

a) *The mobile and clip swapped positions.*

b) *The bottle and watch were removed and a clip and mobile were added.*

c) *The mobile was added and a clip was removed.*

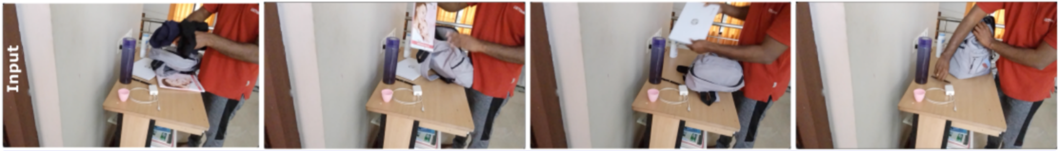
Grounded video QA

Area: ***Memory***, Reasoning: ***Descriptive***

Question: *Track the objects that were added to the table while the camera was looking away.*

PerceptionTest

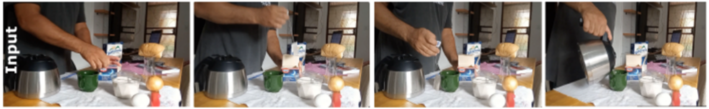
EXAMPLE 3



Multiple-choice video QA
Area: ***Memory***, **Reasoning**: ***Counterfactual***
Question: *If the person had put the objects in the backpack in reverse order, which object or objects would have been put in second?*

Options: a) shirt b) pen c) laptop

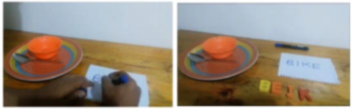
EXAMPLE 4



Multiple-choice video QA
Area: ***Semantics***, **Reasoning**: ***Explanatory***
Question: *What action or actions did the person fail to complete and why?*

Options:
a) *The person put the teabag next to the cup instead of inside the cup.*
b) *The person tried to pour water, but failed because they didn't tilt the container enough.*
c) *The person tried to pour water, but failed because the water container seems empty.*

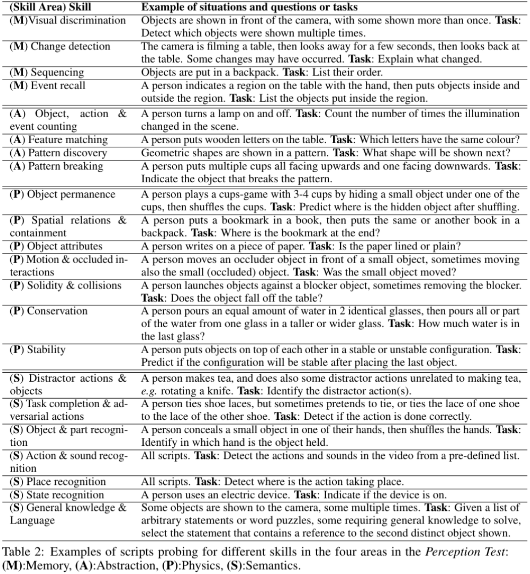
EXAMPLE 5



Multiple-choice video QA
Area: ***Abstraction***, **Reasoning**: ***Descriptive***
Question: *Which letters from the ones the person puts on the table have the same colour?*

Options:
a) *EI* b) *BE* c) *IK*

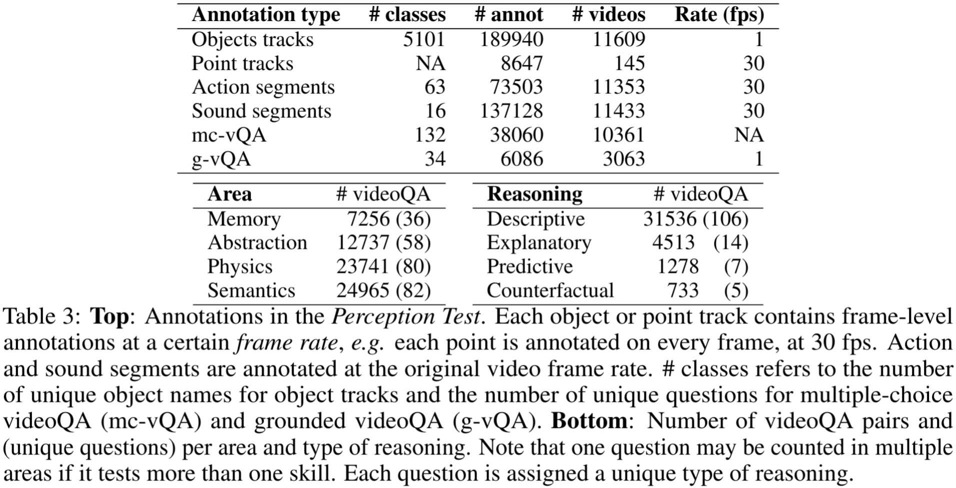
PerceptionTest



(Skill Area) Skill	Example of situations and questions or tasks
(**M**)Visual discrimination	Objects are shown in front of the camera, with some shown more than once. **Task**: Detect which objects were shown multiple times.
(**M**) Change detection	The camera is filming a table, then looks away for a few seconds, then looks back at the table. Some changes may have occurred. **Task**: Explain what changed.
(**M**) Sequencing	Objects are put in a backpack. **Task**: List their order.
(**M**) Event recall	A person indicates a region on the table with the hand, then puts objects inside and outside the region. **Task**: List the objects put inside the region.
(**A**) Object, action & event counting	A person turns a lamp on and off. **Task**: Count the number of times the illumination changed in the scene.
(**A**) Feature matching	A person puts wooden letters on the table. **Task**: Which letters have the same colour?
(**A**) Pattern discovery	Geometric shapes are shown in a pattern. **Task**: What shape will be shown next?
(**A**) Pattern breaking	A person puts multiple cups all facing upwards and one facing downwards. **Task**: Indicate the object that breaks the pattern.
(**P**) Object permanence	A person plays a cups-game with 3-4 cups by hiding a small object under one of the cups, then shuffles the cups. **Task**: Predict where is the hidden object after shuffling.
(**P**) Spatial relations & containment	A person puts a bookmark in a book, then puts the same or another book in a backpack. **Task**: Where is the bookmark at the end?
(**P**) Object attributes	A person writes on a piece of paper. **Task**: Is the paper lined or plain?
(**P**) Motion & occluded interactions	A person moves an occluder object in front of a small object, sometimes moving also the small (occluded) object. **Task**: Was the small object moved?
(**P**) Solidity & collisions	A person launches objects against a blocker object, sometimes removing the blocker. **Task**: Does the object fall off the table?
(**P**) Conservation	A person pours an equal amount of water in 2 identical glasses, then pours all or part of the water from one glass in a taller or wider glass. **Task**: How much water is in the last glass?
(**P**) Stability	A person puts objects on top of each other in a stable or unstable configuration. **Task**: Predict if the configuration will be stable after placing the last object.
(**S**) Distractor actions & objects	A person makes tea, and does also some distractor actions unrelated to making tea, *e.g.* rotating a knife. **Task**: Identify the distractor action(s).
(**S**) Task completion & adversarial actions	A person ties shoe laces, but sometimes pretends to tie, or ties the lace of one shoe to the lace of the other shoe. **Task**: Detect if the action is done correctly.
(**S**) Object & part recognition	A person conceals a small object in one of their hands, then shuffles the hands. **Task**: Identify in which hand is the object held.
(**S**) Action & sound recognition	All scripts. **Task**: Detect the actions and sounds in the video from a pre-defined list.
(**S**) Place recognition	All scripts. **Task**: Detect where is the action taking place.
(**S**) State recognition	A person uses an electric device. **Task**: Indicate if the device is on.
(**S**) General knowledge & Language	Some objects are shown to the camera, some multiple times. **Task**: Given a list of arbitrary statements or word puzzles, some requiring general knowledge to solve, select the statement that contains a reference to the second distinct object shown.

Table 2: Examples of scripts probing for different skills in the four areas in the *Perception Test*: (**M**):Memory, (**A**):Abstraction, (**P**):Physics, (**S**):Semantics.

PerceptionTest



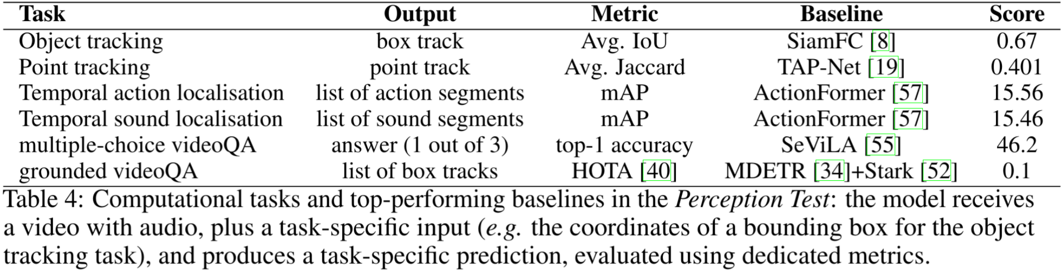
Annotation type	# classes	# annot	# videos	Rate (fps)
Objects tracks	5101	189940	11609	1
Point tracks	NA	8647	145	30
Action segments	63	73503	11353	30
Sound segments	16	137128	11433	30
mc-vQA	132	38060	10361	NA
g-vQA	34	6086	3063	1

Area	# videoQA
Memory	7256 (36)
Abstraction	12737 (58)
Physics	23741 (80)
Semantics	24965 (82)

Reasoning	# videoQA
Descriptive	31536 (106)
Explanatory	4513 (14)
Predictive	1278 (7)
Counterfactual	733 (5)

Table 3: **Top**: Annotations in the *Perception Test*. Each object or point track contains frame-level annotations at a certain *frame rate*, *e.g.* each point is annotated on every frame, at 30 fps. Action and sound segments are annotated at the original video frame rate. # classes refers to the number of unique object names for object tracks and the number of unique questions for multiple-choice videoQA (mc-vQA) and grounded videoQA (g-vQA). **Bottom**: Number of videoQA pairs and (unique questions) per area and type of reasoning. Note that one question may be counted in multiple areas if it tests more than one skill. Each question is assigned a unique type of reasoning.

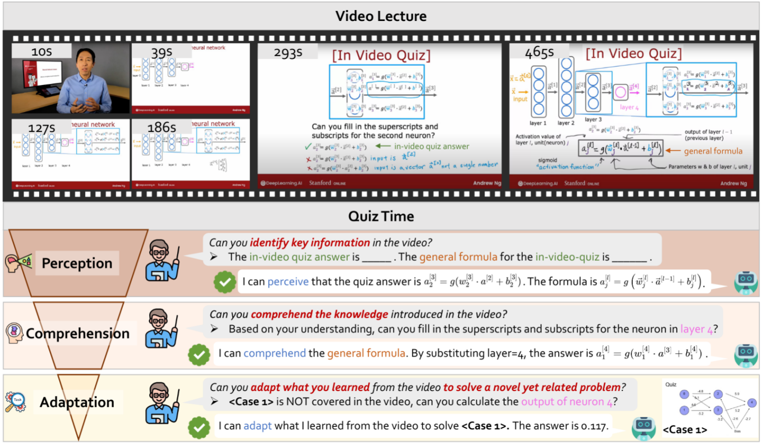
PerceptionTest



Task	Output	Metric	Baseline	Score
Object tracking	box track	Avg. IoU	SiamFC [8]	0.67
Point tracking	point track	Avg. Jaccard	TAP-Net [19]	0.401
Temporal action localisation	list of action segments	mAP	ActionFormer [57]	15.56
Temporal sound localisation	list of sound segments	mAP	ActionFormer [57]	15.46
multiple-choice videoQA	answer (1 out of 3)	top-1 accuracy	SeViLA [55]	46.2
grounded videoQA	list of box tracks	HOTA [40]	MDETR [34]+Stark [52]	0.1

Table 4: Computational tasks and top-performing baselines in the *Perception Test*: the model receives a video with audio, plus a task-specific input (*e.g.* the coordinates of a bounding box for the object tracking task), and produces a task-specific prediction, evaluated using dedicated metrics.

Video-MMMU



Can you fill in the superscripts and subscripts for the second neuron?

*Can you **identify key information** in the video?*

- The in-video quiz answer is _____ . The general formula for the in-video-quiz is ______ .

✓ I can perceive that the quiz answer is $a_2^{[3]} = g(w_2^{[3]} \cdot a^{[2]} + b_2^{[3]})$. The formula is $a_j^{[l]} = g\left(\vec{w}_j^{[l]} \cdot \vec{a}^{[l-1]} + b_j^{[l]}\right)$.

*Can you **comprehend the knowledge** introduced in the video?*

- Based on your understanding, can you fill in the superscripts and subscripts for the neuron in layer 4?

✓ I can comprehend the general formula. By substituting layer=4, the answer is $a_1^{[4]} = g(w_1^{[4]} \cdot a^{[3]} + b_1^{[4]})$.

*Can you **adapt what you learned** from the video **to solve a novel yet related problem**?*

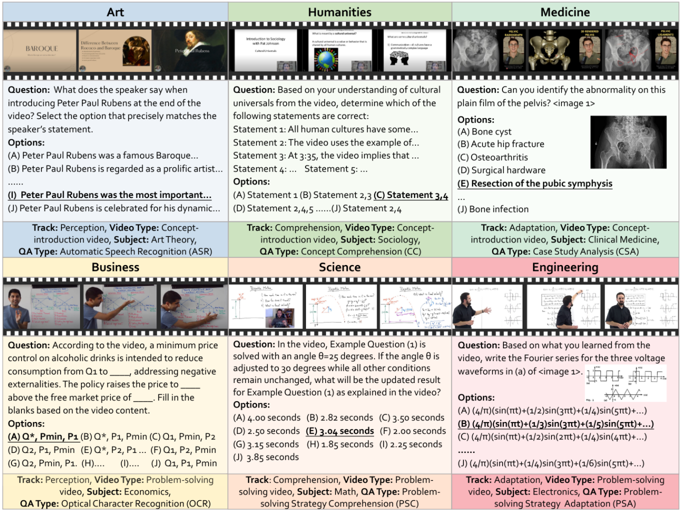
- **<Case 1>** is NOT covered in the video, can you calculate the output of neuron 4?

✓ I can adapt what I learned from the video to solve **<Case 1>.** The answer is 0.117.

Figure 1. An illustration of **Video-MMMU**: Evaluating the knowledge acquisition capability from videos through three cognitive stages: **1) Perception:** if models can identify key information related to knowledge; **2) Comprehension:** if models can interpret the underlying concepts; **3) Adaptation:** if models can adapt the knowledge from videos to novel scenarios.

Hu et al. Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. arXiv:2501.13826

Video-MMMU



Art

Question: What does the speaker say when introducing Peter Paul Rubens at the end of the video? Select the option that precisely matches the speaker's statement.

Options:
(A) Peter Paul Rubens was a famous Baroque...
(B) Peter Paul Rubens is regarded as a prolific artist...
......
(I) Peter Paul Rubens was the most important...
(J) Peter Paul Rubens is celebrated for his dynamic...

Track: Perception, **Video Type:** Concept-introduction video, **Subject:** Art Theory, **QA Type:** Automatic Speech Recognition (ASR)

Humanities

Question: Based on your understanding of cultural universals from the video, determine which of the following statements are correct:
Statement 1: All human cultures have some...
Statement 2: The video uses the example of...
Statement 3: At 3:35, the video implies that ...
Statement 4: ... Statement 5: ...

Options:
(A) Statement 1 (B) Statement 2,3 **(C) Statement 3,4**
(D) Statement 2,4,5 ......(J) Statement 2,4

Track: Comprehension, **Video Type:** Concept-introduction video, **Subject:** Sociology, **QA Type:** Concept Comprehension (CC)

Medicine

Question: Can you identify the abnormality on this plain film of the pelvis? <image 1>

Options:
(A) Bone cyst
(B) Acute hip fracture
(C) Osteoarthritis
(D) Surgical hardware
(E) Resection of the pubic symphysis
...
(J) Bone infection

Track: Adaptation, **Video Type:** Concept-introduction video, **Subject:** Clinical Medicine, **QA Type:** Case Study Analysis (CSA)

Business

Question: According to the video, a minimum price control on alcoholic drinks is intended to reduce consumption from Q1 to ____, addressing negative externalities. The policy raises the price to ____ above the free market price of ____. Fill in the blanks based on the video content.

Options:
(A) Q*, Pmin, P1 (B) Q*, P1, Pmin (C) Q1, Pmin, P2
(D) Q2, P1, Pmin (E) Q*, P2, P1 ... (F) Q1, P2, Pmin
(G) Q2, Pmin, P1. (H).... (I).... (J) Q1, P1, Pmin

Track: Perception, **Video Type:** Problem-solving video, **Subject:** Economics, **QA Type:** Optical Character Recognition (OCR)

Science

Question: In the video, Example Question (1) is solved with an angle θ=25 degrees. If the angle θ is adjusted to 30 degrees while all other conditions remain unchanged, what will be the updated result for Example Question (1) as explained in the video?

Options:
(A) 4.00 seconds (B) 2.82 seconds (C) 3.50 seconds
(D) 2.50 seconds **(E) 3.04 seconds** (F) 2.00 seconds
(G) 3.15 seconds (H) 1.85 seconds (I) 2.25 seconds
(J) 3.85 seconds

Track: Comprehension, **Video Type:** Problem-solving video, **Subject:** Math, **QA Type:** Problem-solving Strategy Comprehension (PSC)

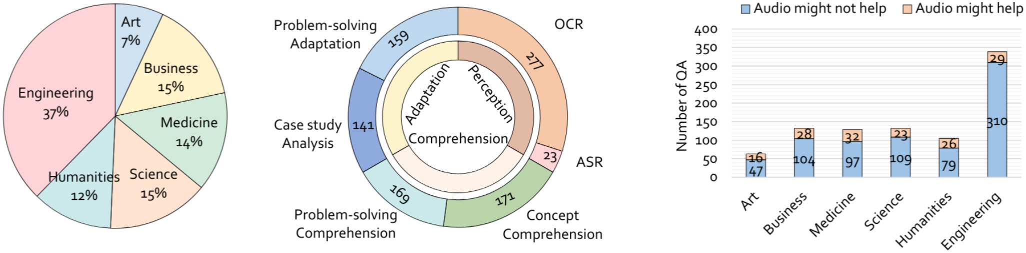
Engineering

Question: Based on what you learned from the video, write the Fourier series for the three voltage waveforms in (a) of <image 1>.

Options:
(A) (4/π)(sin(πt)+(1/2)sin(3πt)+(1/4)sin(5πt)+...)
(B) (4/π)(sin(πt)+(1/3)sin(3πt)+(1/5)sin(5πt)+...)
(C) (4/π)(sin(πt)+(1/2)sin(2πt)+(1/4)sin(4πt)+...)
......
(J) (4/π)(sin(πt)+(1/4)sin(3πt)+(1/6)sin(5πt)+...)

Track: Adaptation, **Video Type:** Problem-solving video, **Subject:** Electronics, **QA Type:** Problem-solving Strategy Adaptation (PSA)

Video-MMMU



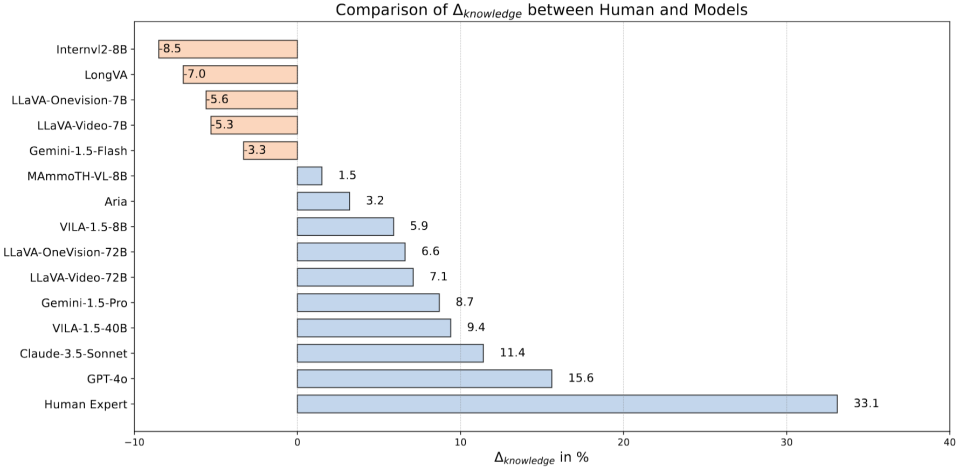
(a) Video distribution across disciplines.

(b) QA distribution across types.

(c) QA distribution with respect to audio.

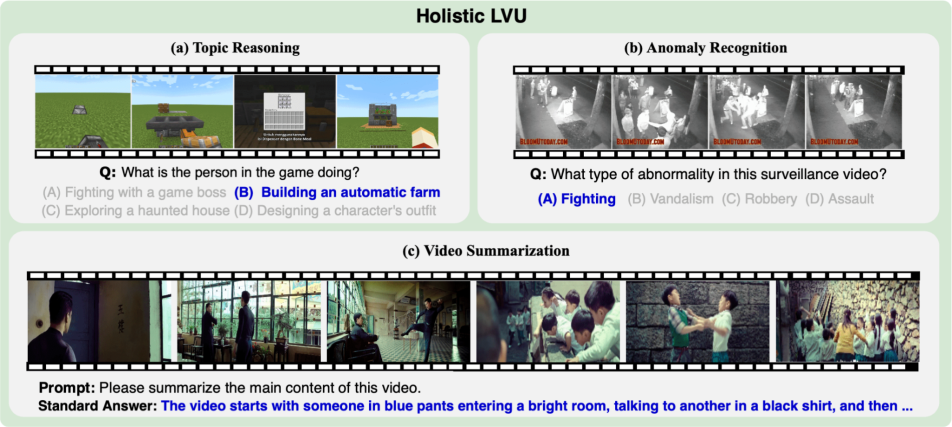
Figure 3. Taxonomy of QA types and video disciplines.

Video-MMMU



(a) Comparison of $\Delta_{\text{knowledge}}$ (performance improvement in the Adaptation track after watching the video compared to before).

MLVU



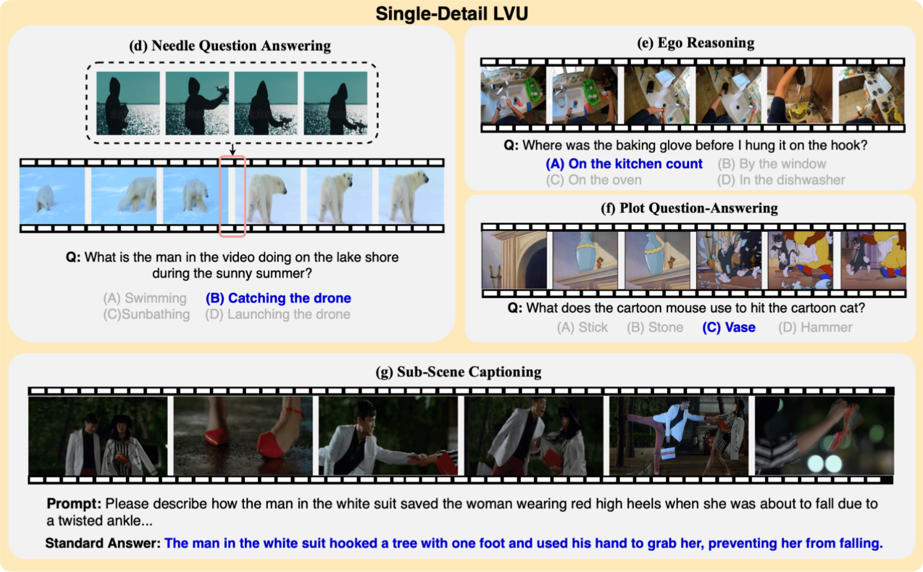
Q: What is the person in the game doing?
(A) Fighting with a game boss **(B) Building an automatic farm**
(C) Exploring a haunted house (D) Designing a character's outfit

Q: What type of abnormality in this surveillance video?
(A) Fighting (B) Vandalism (C) Robbery (D) Assault

Prompt: Please summarize the main content of this video.
Standard Answer: **The video starts with someone in blue pants entering a bright room, talking to another in a black shirt, and then ...**

Zhou et al. MLVU: Benchmarking Multi-task Long Video Understanding. CVPR 2025

MLVU



Single-Detail LVU

(d) Needle Question Answering

Q: What is the man in the video doing on the lake shore during the sunny summer?

(A) Swimming **(B) Catching the drone**
(C)Sunbathing (D) Launching the drone

(e) Ego Reasoning

Q: Where was the baking glove before I hung it on the hook?

(A) On the kitchen count (B) By the window
(C) On the oven (D) In the dishwasher

(f) Plot Question-Answering

Q: What does the cartoon mouse use to hit the cartoon cat?

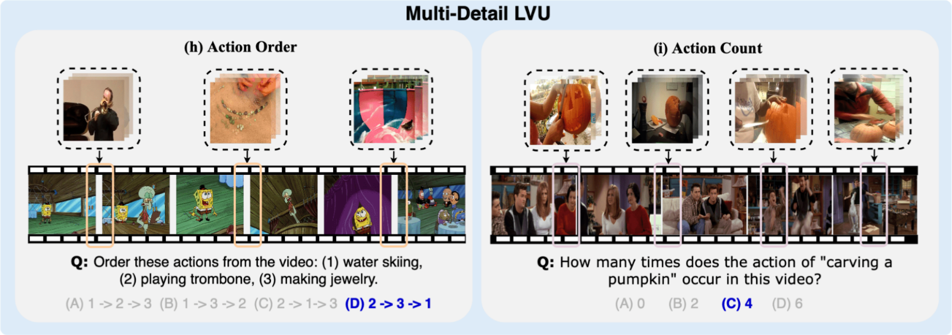
(A) Stick (B) Stone **(C) Vase** (D) Hammer

(g) Sub-Scene Captioning

Prompt: Please describe how the man in the white suit saved the woman wearing red high heels when she was about to fall due to a twisted ankle...

Standard Answer: **The man in the white suit hooked a tree with one foot and used his hand to grab her, preventing her from falling.**

MLVU



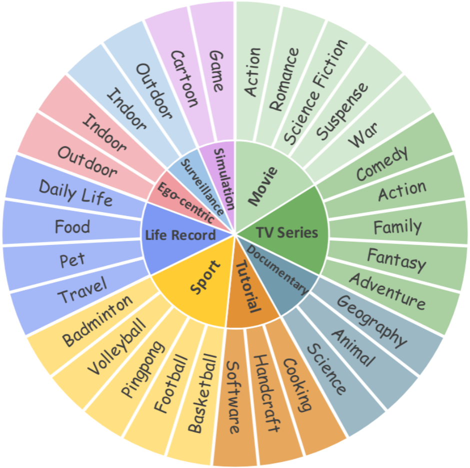
Q: Order these actions from the video: (1) water skiing, (2) playing trombone, (3) making jewelry.

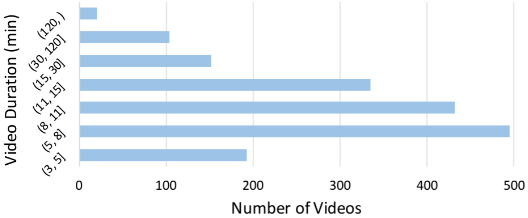
(A) 1 -> 2 -> 3 (B) 1 -> 3 -> 2 (C) 2 -> 1-> 3 **(D) 2 -> 3 -> 1**

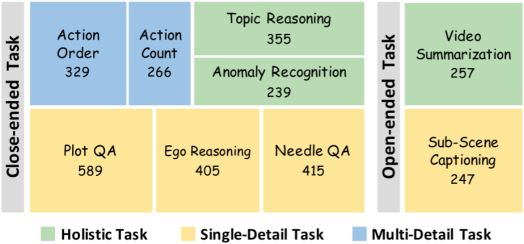
Q: How many times does the action of "carving a pumpkin" occur in this video?

(A) 0 (B) 2 **(C) 4** (D) 6

MLVU







Close-ended Task

Action Order 329

Action Count 266

Topic Reasoning 355

Anomaly Recognition 239

Plot QA 589

Ego Reasoning 405

Needle QA 415

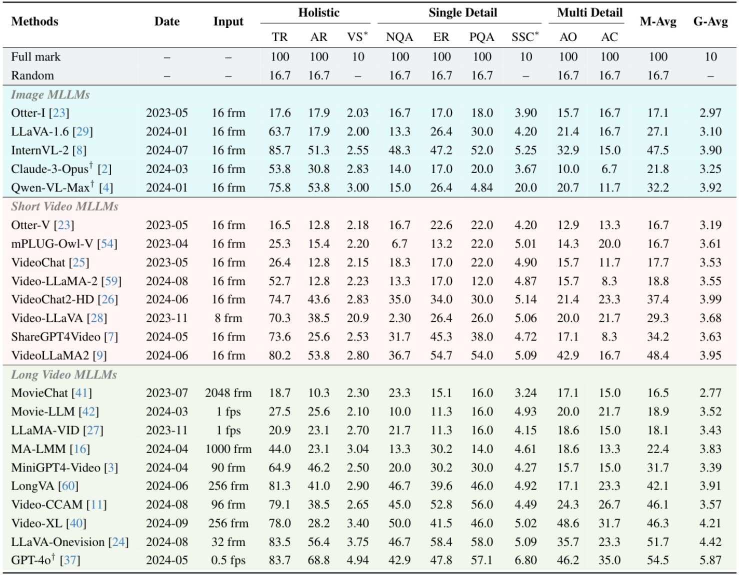
Open-ended Task

Video Summarization 257

Sub-Scene Captioning 247

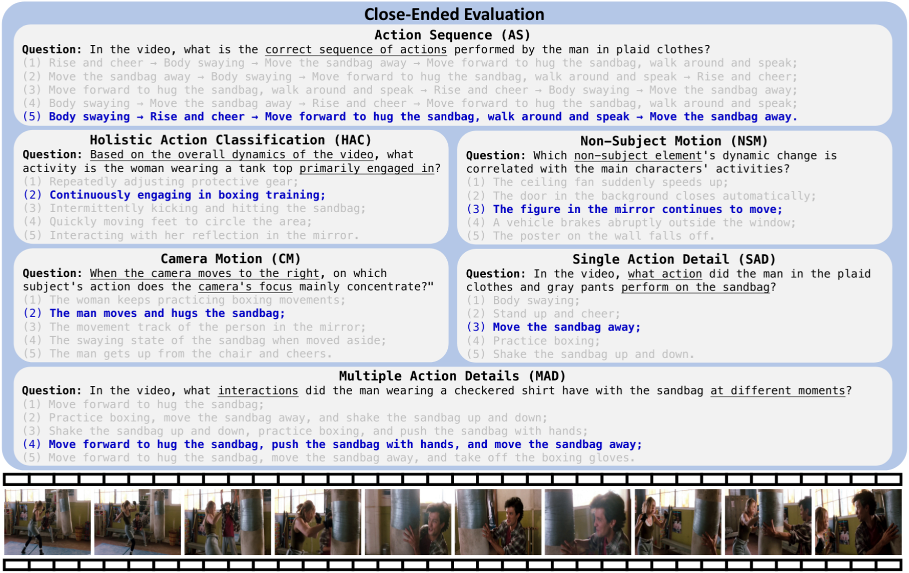
🟩 **Holistic Task** 🟨 **Single-Detail Task** 🟦 **Multi-Detail Task**

MLVU



Methods	Date	Input	Holistic			Single Detail				Multi Detail		M-Avg	G-Avg
			TR	AR	VS*	NQA	ER	PQA	SSC*	AO	AC		
Full mark	–	–	100	100	10	100	100	100	10	100	100	100	10
Random	–	–	16.7	16.7	–	16.7	16.7	16.7	–	16.7	16.7	16.7	–
Image MLLMs													
Otter-I [23]	2023-05	16 frm	17.6	17.9	2.03	16.7	17.0	18.0	3.90	15.7	16.7	17.1	2.97
LLaVA-1.6 [29]	2024-01	16 frm	63.7	17.9	2.00	13.3	26.4	30.0	4.20	21.4	16.7	27.1	3.10
InternVL-2 [8]	2024-07	16 frm	85.7	51.3	2.55	48.3	47.2	52.0	5.25	32.9	15.0	47.5	3.90
Claude-3-Opus† [2]	2024-03	16 frm	53.8	30.8	2.83	14.0	17.0	20.0	3.67	10.0	6.7	21.8	3.25
Qwen-VL-Max† [4]	2024-01	16 frm	75.8	53.8	3.00	15.0	26.4	4.84	20.0	20.7	11.7	32.2	3.92
Short Video MLLMs													
Otter-V [23]	2023-05	16 frm	16.5	12.8	2.18	16.7	22.6	22.0	4.20	12.9	13.3	16.7	3.19
mPLUG-Owl-V [54]	2023-04	16 frm	25.3	15.4	2.20	6.7	13.2	22.0	5.01	14.3	20.0	16.7	3.61
VideoChat [25]	2023-05	16 frm	26.4	12.8	2.15	18.3	17.0	22.0	4.90	15.7	11.7	17.7	3.53
Video-LLaMA-2 [59]	2024-08	16 frm	52.7	12.8	2.23	13.3	17.0	12.0	4.87	15.7	8.3	18.8	3.55
VideoChat2-HD [26]	2024-06	16 frm	74.7	43.6	2.83	35.0	34.0	30.0	5.14	21.4	23.3	37.4	3.99
Video-LLaVA [28]	2023-11	8 frm	70.3	38.5	20.9	2.30	26.4	26.0	5.06	20.0	21.7	29.3	3.68
ShareGPT4Video [7]	2024-05	16 frm	73.6	25.6	2.53	31.7	45.3	38.0	4.72	17.1	8.3	34.2	3.63
VideoLLaMA2 [9]	2024-06	16 frm	80.2	53.8	2.80	36.7	54.7	54.0	5.09	42.9	16.7	48.4	3.95
Long Video MLLMs													
MovieChat [41]	2023-07	2048 frm	18.7	10.3	2.30	23.3	15.1	16.0	3.24	17.1	15.0	16.5	2.77
Movie-LLM [42]	2024-03	1 fps	27.5	25.6	2.10	10.0	11.3	16.0	4.93	20.0	21.7	18.9	3.52
LLaMA-VID [27]	2023-11	1 fps	20.9	23.1	2.70	21.7	11.3	16.0	4.15	18.6	15.0	18.1	3.43
MA-LMM [16]	2024-04	1000 frm	44.0	23.1	3.04	13.3	30.2	14.0	4.61	18.6	13.3	22.4	3.83
MiniGPT4-Video [3]	2024-04	90 frm	64.9	46.2	2.50	20.0	30.2	30.0	4.27	15.7	15.0	31.7	3.39
LongVA [60]	2024-06	256 frm	81.3	41.0	2.90	46.7	39.6	46.0	4.92	17.1	23.3	42.1	3.91
Video-CCAM [11]	2024-08	96 frm	79.1	38.5	2.65	45.0	52.8	56.0	4.49	24.3	26.7	46.1	3.57
Video-XL [40]	2024-09	256 frm	78.0	28.2	3.40	50.0	41.5	46.0	5.02	48.6	31.7	46.3	4.21
LLaVA-Onevision [24]	2024-08	32 frm	83.5	56.4	3.75	46.7	58.4	58.0	5.09	35.7	23.3	51.7	4.42
GPT-4o† [37]	2024-05	0.5 fps	83.7	68.8	4.94	42.9	47.8	57.1	6.80	46.2	35.0	54.5	5.87

FAVOR-Bench



Close-Ended Evaluation

Action Sequence (AS)

Question: In the video, what is the correct sequence of actions performed by the man in plaid clothes?

(1) Rise and cheer → Body swaying → Move the sandbag away → Move forward to hug the sandbag, walk around and speak;
(2) Move the sandbag away → Body swaying → Move forward to hug the sandbag, walk around and speak → Rise and cheer;
(3) Move forward to hug the sandbag, walk around and speak → Rise and cheer → Body swaying → Move the sandbag away;
(4) Body swaying → Move the sandbag away → Rise and cheer → Move forward to hug the sandbag, walk around and speak;
(5) Body swaying → Rise and cheer → Move forward to hug the sandbag, walk around and speak → Move the sandbag away.

Holistic Action Classification (HAC)

Question: Based on the overall dynamics of the video, what activity is the woman wearing a tank top primarily engaged in?

(1) Repeatedly adjusting protective gear;
(2) Continuously engaging in boxing training;
(3) Intermittently kicking and hitting the sandbag;
(4) Quickly moving feet to circle the area;
(5) Interacting with her reflection in the mirror.

Non-Subject Motion (NSM)

Question: Which non-subject element's dynamic change is correlated with the main characters' activities?

(1) The ceiling fan suddenly speeds up;
(2) The door in the background closes automatically;
(3) The figure in the mirror continues to move;
(4) A vehicle brakes abruptly outside the window;
(5) The poster on the wall falls off.

Camera Motion (CM)

Question: When the camera moves to the right, on which subject's action does the camera's focus mainly concentrate?"

(1) The woman keeps practicing boxing movements;
(2) The man moves and hugs the sandbag;
(3) The movement track of the person in the mirror;
(4) The swaying state of the sandbag when moved aside;
(5) The man gets up from the chair and cheers.

Single Action Detail (SAD)

Question: In the video, what action did the man in the plaid clothes and gray pants perform on the sandbag?

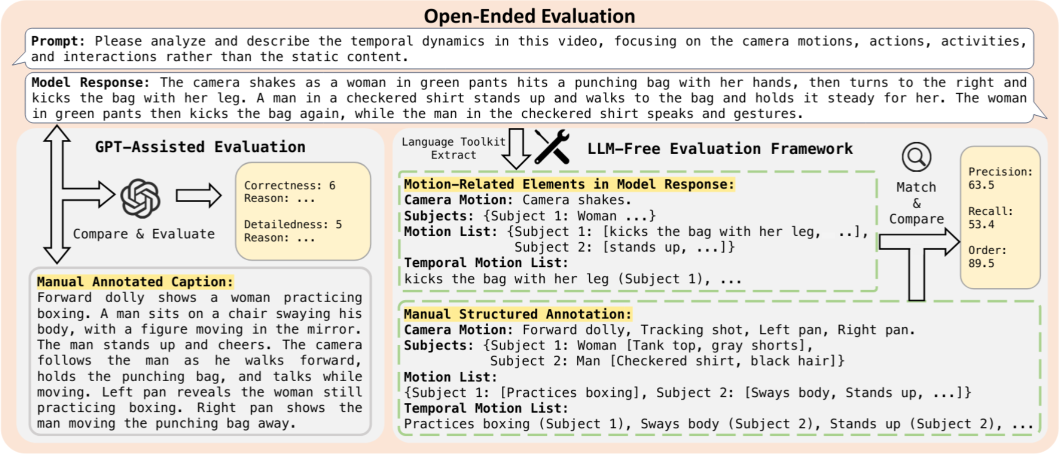
(1) Body swaying;
(2) Stand up and cheer;
(3) Move the sandbag away;
(4) Practice boxing;
(5) Shake the sandbag up and down.

Multiple Action Details (MAD)

Question: In the video, what interactions did the man wearing a checkered shirt have with the sandbag at different moments?

(1) Move forward to hug the sandbag;
(2) Practice boxing, move the sandbag away, and shake the sandbag up and down;
(3) Shake the sandbag up and down, practice boxing, and push the sandbag with hands;
(4) Move forward to hug the sandbag, push the sandbag with hands, and move the sandbag away;
(5) Move forward to hug the sandbag, move the sandbag away, and take off the boxing gloves.

FAVOR-Bench



Prompt: Please analyze and describe the temporal dynamics in this video, focusing on the camera motions, actions, activities, and interactions rather than the static content.

Model Response: The camera shakes as a woman in green pants hits a punching bag with her hands, then turns to the right and kicks the bag with her leg. A man in a checkered shirt stands up and walks to the bag and holds it steady for her. The woman in green pants then kicks the bag again, while the man in the checkered shirt speaks and gestures.

Correctness: 6
Reason: ...

Detailedness: 5
Reason: ...

Manual Annotated Caption:
Forward dolly shows a woman practicing boxing. A man sits on a chair swaying his body, with a figure moving in the mirror. The man stands up and cheers. The camera follows the man as he walks forward, holds the punching bag, and talks while moving. Left pan reveals the woman still practicing boxing. Right pan shows the man moving the punching bag away.

Motion-Related Elements in Model Response:
Camera Motion: Camera shakes.
Subjects: {Subject 1: Woman ...}
Motion List: {Subject 1: [kicks the bag with her leg, ..],
Subject 2: [stands up, ...]}
Temporal Motion List:
kicks the bag with her leg (Subject 1), ...

Precision:
63.5

Recall:
53.4

Order:
89.5

Manual Structured Annotation:
Camera Motion: Forward dolly, Tracking shot, Left pan, Right pan.
Subjects: {Subject 1: Woman [Tank top, gray shorts],
Subject 2: Man [Checkered shirt, black hair]}
Motion List:
{Subject 1: [Practices boxing], Subject 2: [Sways body, Stands up, ...]}
Temporal Motion List:
Practices boxing (Subject 1), Sways body (Subject 2), Stands up (Subject 2), ...

Tu et al. FAVOR-Bench: A Comprehensive Benchmark for Fine-Grained Video Motion Understanding. arXiv:2503.14935

Outline

1. Tasks and benchmarks

2. Models

LLaVA-OneVision

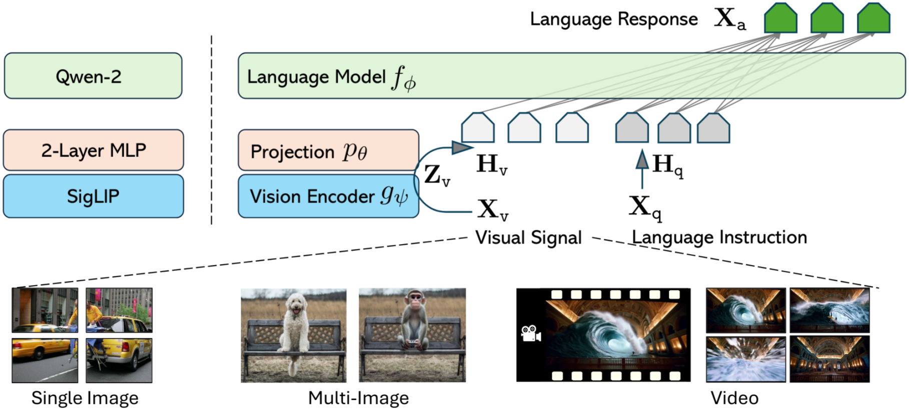
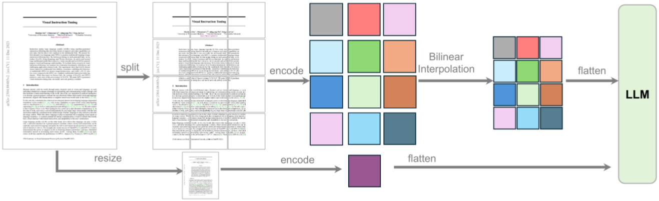


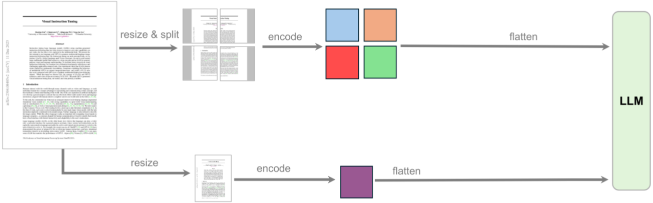
Figure 1: LLaVA-OneVision network architecture. Left: The current model instantiation; Right: the general form of LLaVA architecture in [83], but is extended to support more visual signals.

Li et al. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326

LLaVA-OneVision



(a) Higher AnyRes with Bilinear Interpolation



(b) The original AnyRes

LLaVA-OneVision

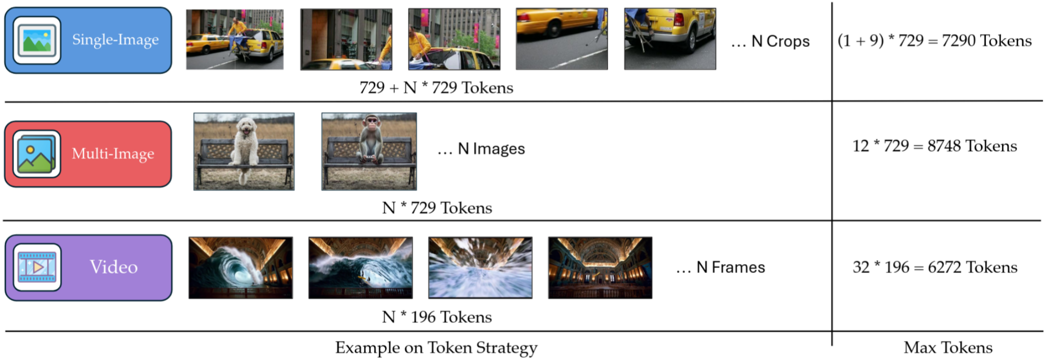
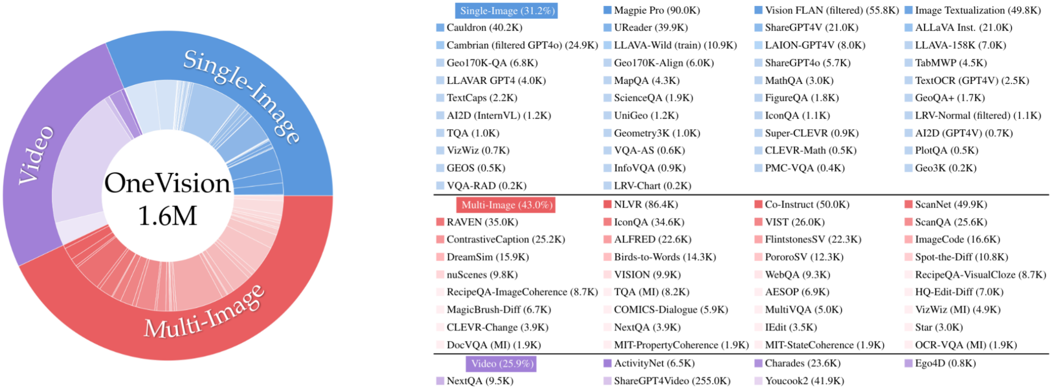


Figure 3: The visual representation strategy to allocate tokens for each scenario in LLaVA-OneVision. The maximum number of visual tokens across different scenarios is designed to be similar, ensuring balanced visual representations to accommodate cross-scenario capability transfer. Note that 729 is the #tokens for SigLIP to encode a visual input of resolution 384×384.

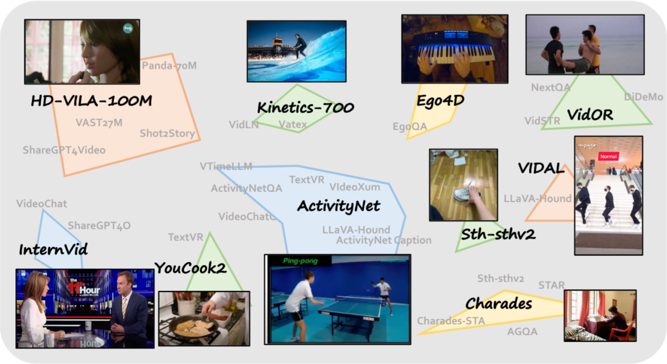
LLaVA-OneVision



Single-Image (31.2%)	🟦 Magpie Pro (90.0K)	🟦 Vision FLAN (filtered) (55.8K)	🟦 Image Textualization (49.8K)
🟦 Cauldron (40.2K)	🟦 UReader (39.9K)	🟦 ShareGPT4V (21.0K)	🟦 ALLaVA Inst. (21.0K)
🟦 Cambrian (filtered GPT4o) (24.9K)	🟦 LLAVA-Wild (train) (10.9K)	🟦 LAION-GPT4V (8.0K)	🟦 LLAVA-158K (7.0K)
🟦 Geo170K-QA (6.8K)	🟦 Geo170K-Align (6.0K)	🟦 ShareGPT4o (5.7K)	🟦 TabMWP (4.5K)
🟦 LLAVAR GPT4 (4.0K)	🟦 MapQA (4.3K)	🟦 MathQA (3.0K)	🟦 TextOCR (GPT4V) (2.5K)
🟦 TextCaps (2.2K)	🟦 ScienceQA (1.9K)	🟦 FigureQA (1.8K)	🟦 GeoQA+ (1.7K)
🟦 AI2D (InternVL) (1.2K)	🟦 UniGeo (1.2K)	🟦 IconQA (1.1K)	🟦 LRV-Normal (filtered) (1.1K)
🟦 TQA (1.0K)	🟦 Geometry3K (1.0K)	🟦 Super-CLEVR (0.9K)	🟦 AI2D (GPT4V) (0.7K)
🟦 VizWiz (0.7K)	🟦 VQA-AS (0.6K)	🟦 CLEVR-Math (0.5K)	🟦 PlotQA (0.5K)
🟦 GEOS (0.5K)	🟦 InfoVQA (0.9K)	🟦 PMC-VQA (0.4K)	🟦 Geo3K (0.2K)
🟦 VQA-RAD (0.2K)	🟦 LRV-Chart (0.2K)		
Multi-Image (43.0%)	🟥 NLVR (86.4K)	🟥 Co-Instruct (50.0K)	🟥 ScanNet (49.9K)
🟥 RAVEN (35.0K)	🟥 IconQA (34.6K)	🟥 VIST (26.0K)	🟥 ScanQA (25.6K)
🟥 ContrastiveCaption (25.2K)	🟥 ALFRED (22.6K)	🟥 FlintstonesSV (22.3K)	🟥 ImageCode (16.6K)
🟥 DreamSim (15.9K)	🟥 Birds-to-Words (14.3K)	🟥 PororoSV (12.3K)	🟥 Spot-the-Diff (10.8K)
🟥 nuScenes (9.8K)	🟥 VISION (9.9K)	🟥 WebQA (9.3K)	🟥 RecipeQA-VisualCloze (8.7K)
🟥 RecipeQA-ImageCoherence (8.7K)	🟥 TQA (MI) (8.2K)	🟥 AESOP (6.9K)	🟥 HQ-Edit-Diff (7.0K)
🟥 MagicBrush-Diff (6.7K)	🟥 COMICS-Dialogue (5.9K)	🟥 MultiVQA (5.0K)	🟥 VizWiz (MI) (4.9K)
🟥 CLEVR-Change (3.9K)	🟥 NextQA (3.9K)	🟥 IEdit (3.5K)	🟥 Star (3.0K)
🟥 DocVQA (MI) (1.9K)	🟥 MIT-PropertyCoherence (1.9K)	🟥 MIT-StateCoherence (1.9K)	🟥 OCR-VQA (MI) (1.9K)
Video (25.9%)	🟪 ActivityNet (6.5K)	🟪 Charades (23.6K)	🟪 Ego4D (0.8K)
🟪 NextQA (9.5K)	🟪 ShareGPT4Video (255.0K)	🟪 Youcook2 (41.9K)	

Figure 5: **OneVision 1.6M.** A high-quality single-image, multi-image and video dataset collection. Left: Data Distribution within each category. The outer circle shows the distribution of all data categories and the inner circle shows the distribution of data subsets. Right: The detailed quantities of datasets. "MI" means it is the multi-image version dataset proposed by DEMON [69].

LLaVA-Video

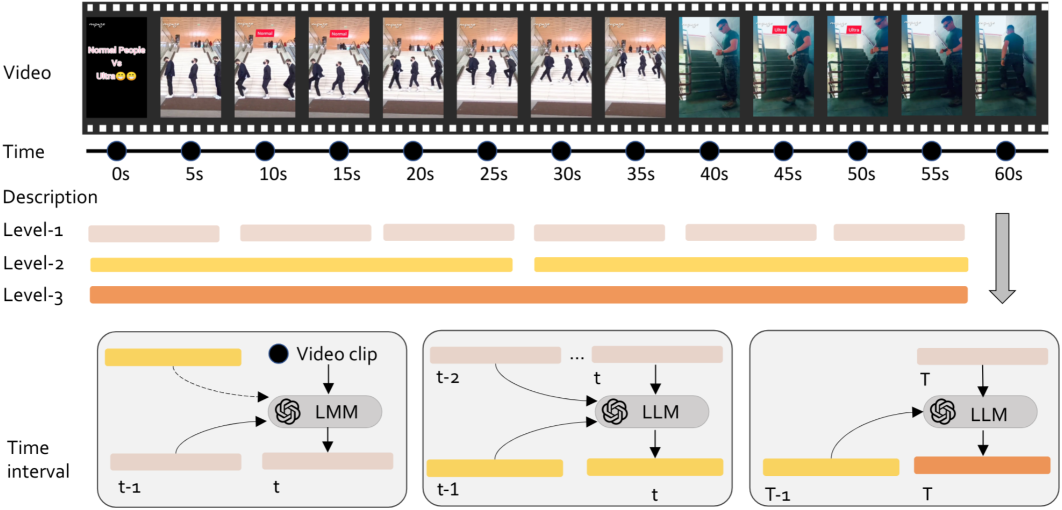


Source	Filtering Logic
HD-VILA-100M	①②③④
InternVid	①②③④
ActivityNet	③⑤
VIDAL	①②③④
Charades	③⑤
Ego4D	③⑤
VidOR	③⑤
YouCook2	③⑤⑥
Kinetics-700	③⑤⑥
Sth-sthv2	③⑤⑥

Figure 1: **Video sources in the proposed *LLaVA-Video-178K*.** (Left) The relationship between 10 video sources we have utilized and other existing video-language datasets. (Right) Filtering logic for video sources. The detail of filtering logic: ① Sorted by Views, ② Number of scenes greater than 2, ③ Video duration between 5 seconds and 180 seconds, ④ Ratio of scenes to video duration less than or equal to 0.5, ⑤ Resolution greater than 480p, ⑥ 50 samples for each category.

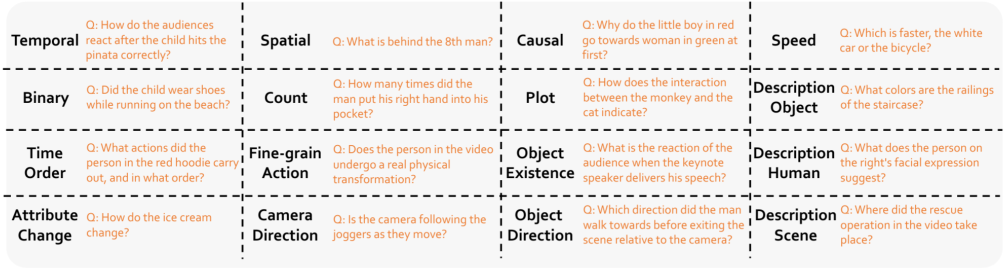
Zhang et al. Video Instruction Tuning With Synthetic Data. arXiv:2410.02713

LLaVA-Video



(a) Level-1 Description (b) Level-2 Description (c) Level-3 Description

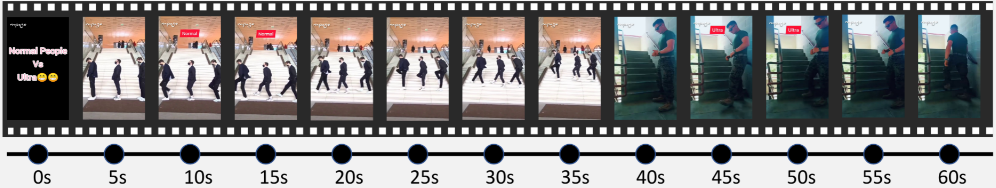
LLaVA-Video



Temporal	Q: How do the audiences react after the child hits the pinata correctly?	**Spatial**	Q: What is behind the 8th man?	**Causal**	Q: Why do the little boy in red go towards woman in green at first?	**Speed**	Q: Which is faster, the white car or the bicycle?
Binary	Q: Did the child wear shoes while running on the beach?	**Count**	Q: How many times did the man put his right hand into his pocket?	**Plot**	Q: How does the interaction between the monkey and the cat indicate?	**Description Object**	Q: What colors are the railings of the staircase?
Time Order	Q: What actions did the person in the red hoodie carry out, and in what order?	**Fine-grain Action**	Q: Does the person in the video undergo a real physical transformation?	**Object Existence**	Q: What is the reaction of the audience when the keynote speaker delivers his speech?	**Description Human**	Q: What does the person on the right's facial expression suggest?
Attribute Change	Q: How do the ice cream change?	**Camera Direction**	Q: Is the camera following the joggers as they move?	**Object Direction**	Q: Which direction did the man walk towards before exiting the scene relative to the camera?	**Description Scene**	Q: Where did the rescue operation in the video take place?

Figure 3: Question types for video question answering in data creation. For each type, we provide its name and an example question.

LLaVA-Video



Annotation type 1: detailed description

The video begins with a black screen displaying the text 'Normal People Vs Ultra' in pink and white letters, accompanied by two smiling face emojis. The scene transitions to a modern building with a staircase. Three individuals, dressed in black suits and white sneakers, stand in a line on the stairs. The text 'Normal' appears in a red box at the top left corner. The individuals start walking up the stairs in a synchronized manner, maintaining their formation. The background shows a few people walking and an escalator on the right side of the stairs. The individuals continue to walk up the stairs in a coordinated manner. The scene then transitions to... <omited>

Annotation type 2: open-ended question

Question: How many steps does "normal people" climb?

Answer: "Normal people" climb 7 steps in the video.

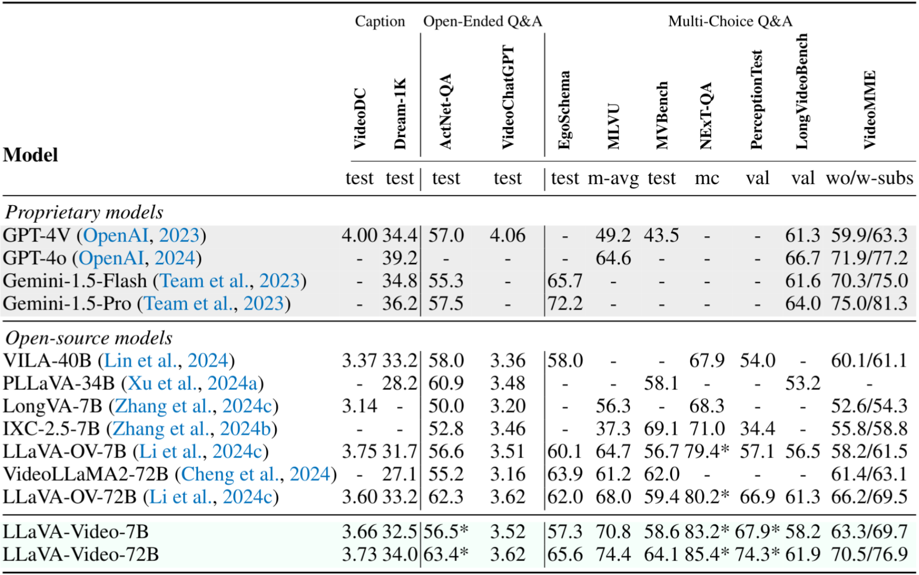
Annotation type 3: multi-choice question

Question: How many steps does "normal people" climb? A. 5 B. 6 C. 7 D.8

Answer: C.7

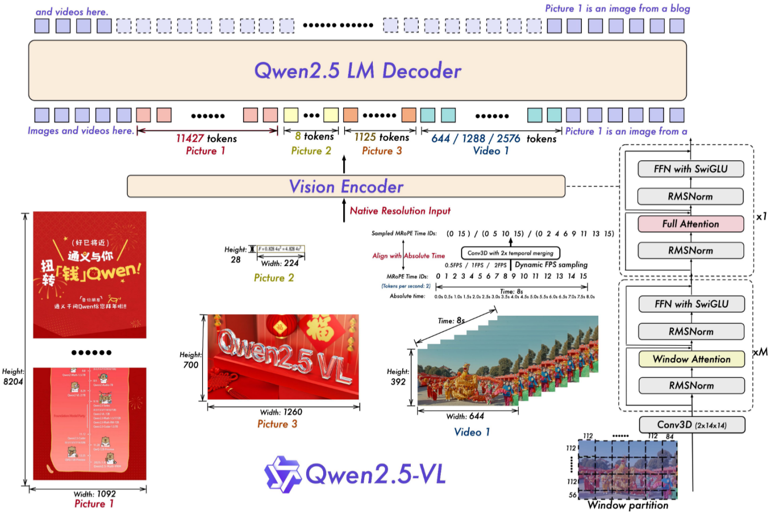
Figure 4: One example to illustrate the video instruction-following data.

LLaVA-Video

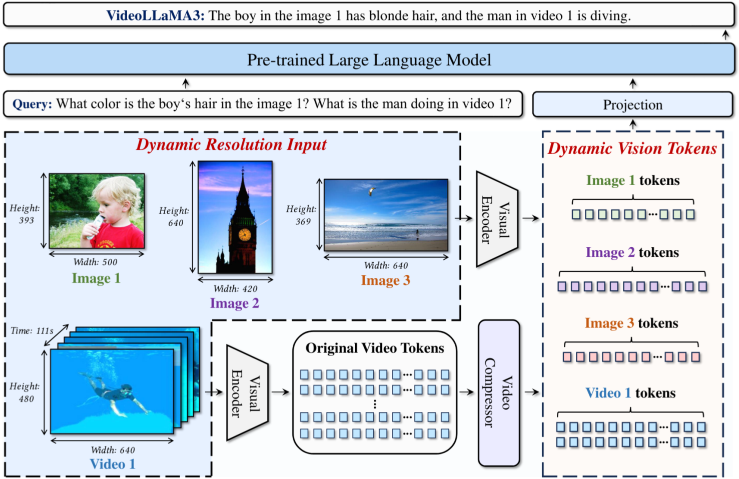


Model	Caption		Open-Ended Q&A		Multi-Choice Q&A						
	VideoDC	**Dream-1K**	**ActNet-QA**	**VideoChatGPT**	**EgoSchema**	**MLVU**	**MVBench**	**NExT-QA**	**PerceptionTest**	**LongVideoBench**	**VideoMME**
	test	test	test	test	test	m-avg	test	mc	val	val	wo/w-subs
Proprietary models											
GPT-4V (OpenAI, 2023)	4.00	34.4	57.0	4.06	-	49.2	43.5	-	-	61.3	59.9/63.3
GPT-4o (OpenAI, 2024)	-	39.2	-	-	-	64.6	-	-	-	66.7	71.9/77.2
Gemini-1.5-Flash (Team et al., 2023)	-	34.8	55.3	-	65.7	-	-	-	-	61.6	70.3/75.0
Gemini-1.5-Pro (Team et al., 2023)	-	36.2	57.5	-	72.2	-	-	-	-	64.0	75.0/81.3
Open-source models											
VILA-40B (Lin et al., 2024)	3.37	33.2	58.0	3.36	58.0	-	-	67.9	54.0	-	60.1/61.1
PLLaVA-34B (Xu et al., 2024a)	-	28.2	60.9	3.48	-	-	58.1	-	-	53.2	-
LongVA-7B (Zhang et al., 2024c)	3.14	-	50.0	3.20	-	56.3	-	68.3	-	-	52.6/54.3
IXC-2.5-7B (Zhang et al., 2024b)	-	-	52.8	3.46	-	37.3	69.1	71.0	34.4	-	55.8/58.8
LLaVA-OV-7B (Li et al., 2024c)	3.75	31.7	56.6	3.51	60.1	64.7	56.7	79.4*	57.1	56.5	58.2/61.5
VideoLLaMA2-72B (Cheng et al., 2024)	-	27.1	55.2	3.16	63.9	61.2	62.0	-	-	-	61.4/63.1
LLaVA-OV-72B (Li et al., 2024c)	3.60	33.2	62.3	3.62	62.0	68.0	59.4	80.2*	66.9	61.3	66.2/69.5
LLaVA-Video-7B	3.66	32.5	56.5*	3.52	57.3	70.8	58.6	83.2*	67.9*	58.2	63.3/69.7
LLaVA-Video-72B	3.73	34.0	63.4*	3.62	65.6	74.4	64.1	85.4*	74.3*	61.9	70.5/76.9

QWen2.5-VL


and videos here.
Picture 1 is an image from a blog
Qwen2.5 LM Decoder
Images and videos here.
11427 tokens
Picture 1
8 tokens
Picture 2
1125 tokens
Picture 3
644 / 1288 / 2576 tokens
Video 1
Picture 1 is an image from a
Vision Encoder
Native Resolution Input
FFN with SwiGLU
RMSNorm
Full Attention
RMSNorm
x1
FFN with SwiGLU
RMSNorm
Window Attention
RMSNorm
xM
Conv3D (2x14x14)
Sampled MRoPE Time IDs: (0 15) / (0 5 10 15) / (0 2 4 6 9 11 13 15)
Align with Absolute Time
Conv3D with 2x temporal merging
0.5FPS / 1FPS / 2FPS
Dynamic FPS sampling
MRoPE Time IDs: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
(Tokens per second: 2)
Time: 8s
Absolute time: 0.0s 0.5s 1.0s 1.5s 2.0s 2.5s 3.0s 3.5s 4.0s 4.5s 5.0s 5.5s 6.0s 6.5s 7.0s 7.5s 8.0s
Height: 28
Width: 224
Picture 2
Height: 8204
Width: 1092
Picture 1
Height: 700
Width: 1260
Picture 3
Time: 8s
Height: 392
Width: 644
Video 1
Qwen2.5-VL
112
112
84
112
112
56
Window partition

VideoLLaMA 3



Zhang et al. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. arXiv:2501.13106

VideoLLaMA 3

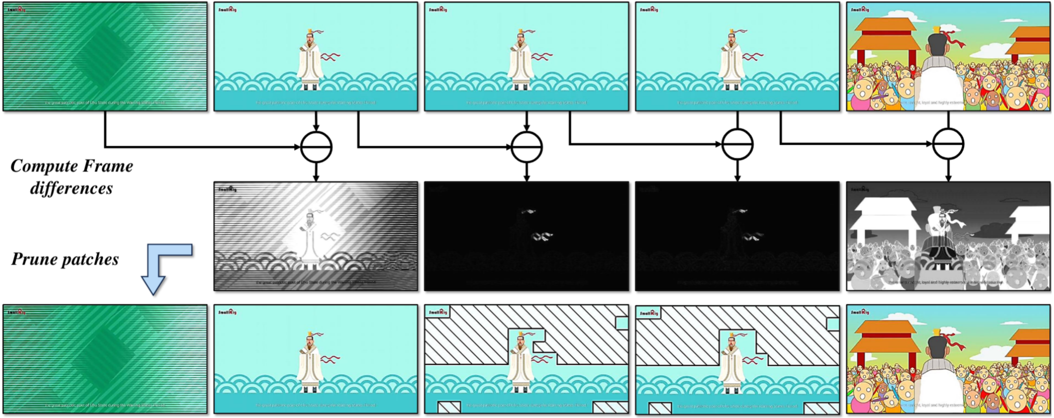


Figure 4: **The calculation flow of our DiffFP.** We prune video tokens based on patch similarities in pixel space, removing patches with smaller distances to the previous frame.

VideoLLaMA 3

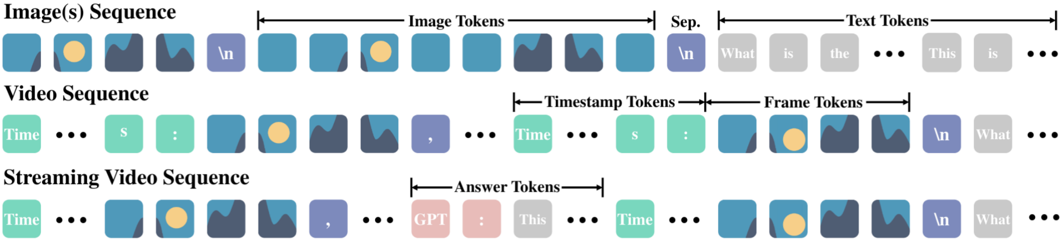


Figure 5: **Data formats for different data types.** ❶ For image sequence, we use "\n" to separate image tokens from different image; ❷ For video sequence, we use "Time: xxs" to indicate timestamps of each frame, "," to separate different frames, and "\n" to separate tokens from different videos; ❸ For streaming video sequence, videos and texts are organized in an interleaved format.

VideoLLaMA 3

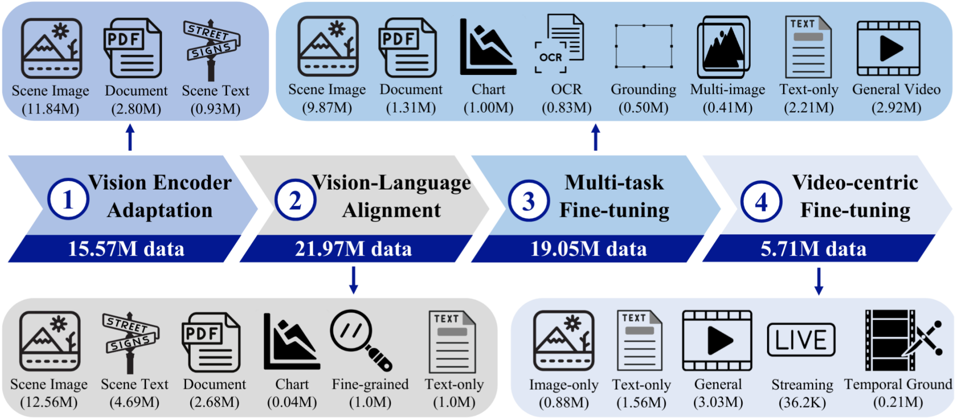


Figure 2: **Training paradigm of VideoLLaMA3.** The training of VideoLLaMA3 has four stages: (1) Vision Encoder Adaptation, (2) Vision-Language Alignment, (3) Multi-task Fine-tuning, and (4) Video-centric Fine-tuning.

VideoLLaMA 3

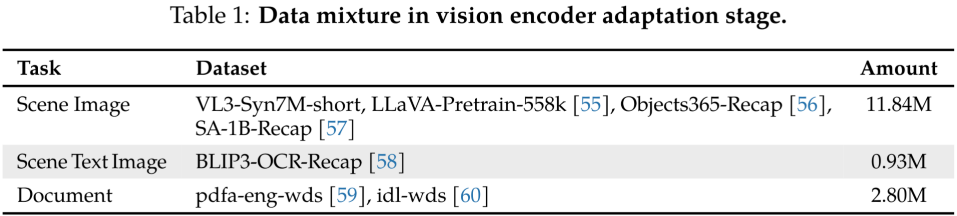


Table 1: **Data mixture in vision encoder adaptation stage.**

Task	**Dataset**	**Amount**
Scene Image	VL3-Syn7M-short, LLaVA-Pretrain-558k [55], Objects365-Recap [56], SA-1B-Recap [57]	11.84M
Scene Text Image	BLIP3-OCR-Recap [58]	0.93M
Document	pdfa-eng-wds [59], idl-wds [60]	2.80M

VideoLLaMA 3

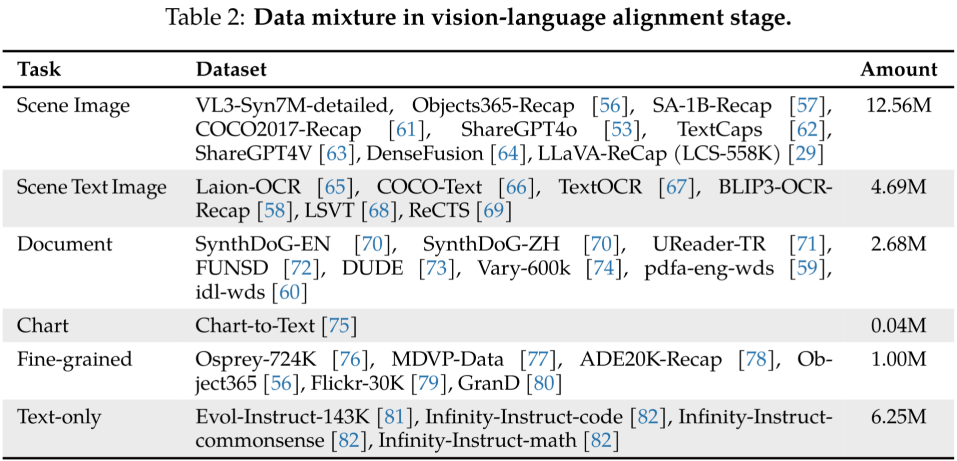
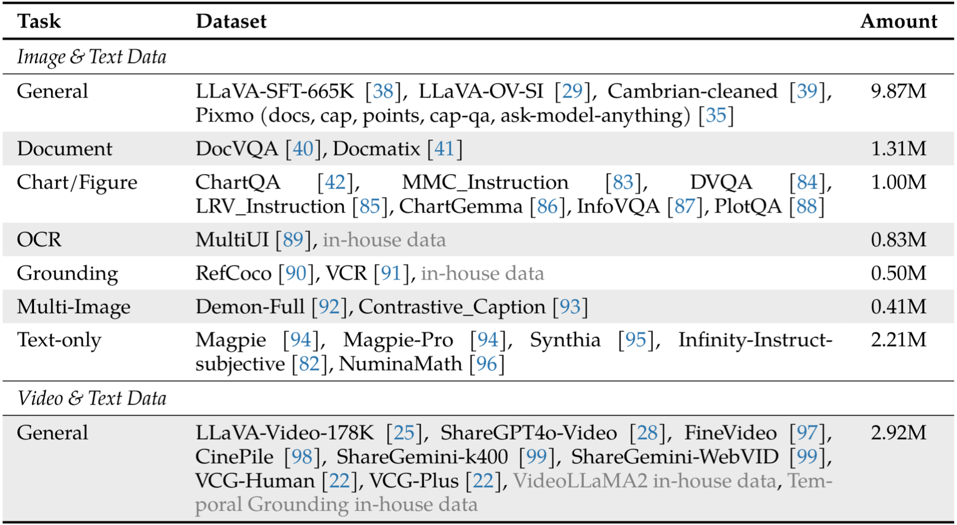


Table 2: **Data mixture in vision-language alignment stage.**

Task	Dataset	Amount
Scene Image	VL3-Syn7M-detailed, Objects365-Recap [56], SA-1B-Recap [57], COCO2017-Recap [61], ShareGPT4o [53], TextCaps [62], ShareGPT4V [63], DenseFusion [64], LLaVA-ReCap (LCS-558K) [29]	12.56M
Scene Text Image	Laion-OCR [65], COCO-Text [66], TextOCR [67], BLIP3-OCR-Recap [58], LSVT [68], ReCTS [69]	4.69M
Document	SynthDoG-EN [70], SynthDoG-ZH [70], UReader-TR [71], FUNSD [72], DUDE [73], Vary-600k [74], pdfa-eng-wds [59], idl-wds [60]	2.68M
Chart	Chart-to-Text [75]	0.04M
Fine-grained	Osprey-724K [76], MDVP-Data [77], ADE20K-Recap [78], Object365 [56], Flickr-30K [79], GranD [80]	1.00M
Text-only	Evol-Instruct-143K [81], Infinity-Instruct-code [82], Infinity-Instruct-commonsense [82], Infinity-Instruct-math [82]	6.25M

VideoLLaMA 3

Table 3: **Data mixture in massive multi-task fine-tuning stage.**



Task	**Dataset**	**Amount**
Image & Text Data		
General	LLaVA-SFT-665K [38], LLaVA-OV-SI [29], Cambrian-cleaned [39], Pixmo (docs, cap, points, cap-qa, ask-model-anything) [35]	9.87M
Document	DocVQA [40], Docmatix [41]	1.31M
Chart/Figure	ChartQA [42], MMC_Instruction [83], DVQA [84], LRV_Instruction [85], ChartGemma [86], InfoVQA [87], PlotQA [88]	1.00M
OCR	MultiUI [89], in-house data	0.83M
Grounding	RefCoco [90], VCR [91], in-house data	0.50M
Multi-Image	Demon-Full [92], Contrastive_Caption [93]	0.41M
Text-only	Magpie [94], Magpie-Pro [94], Synthia [95], Infinity-Instruct-subjective [82], NuminaMath [96]	2.21M
Video & Text Data		
General	LLaVA-Video-178K [25], ShareGPT4o-Video [28], FineVideo [97], CinePile [98], ShareGemini-k400 [99], ShareGemini-WebVID [99], VCG-Human [22], VCG-Plus [22], VideoLLaMA2 in-house data, Temporal Grounding in-house data	2.92M

VideoLLaMA 3

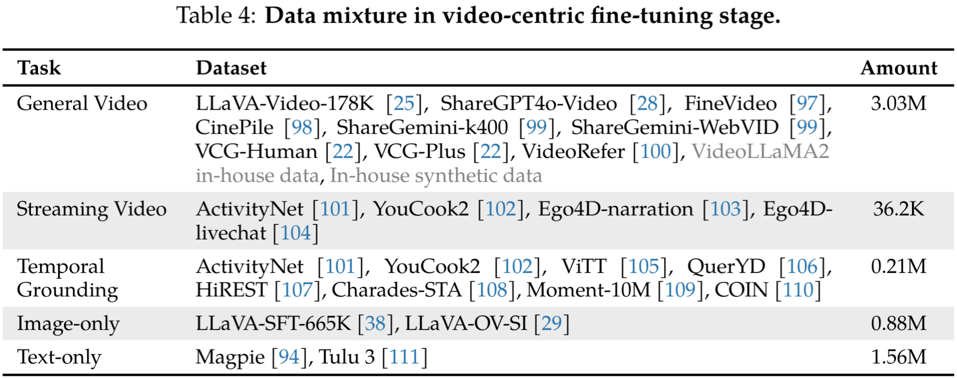


Table 4: **Data mixture in video-centric fine-tuning stage.**

Task	**Dataset**	**Amount**
General Video	LLaVA-Video-178K [25], ShareGPT4o-Video [28], FineVideo [97], CinePile [98], ShareGemini-k400 [99], ShareGemini-WebVID [99], VCG-Human [22], VCG-Plus [22], VideoRefer [100], VideoLLaMA2 in-house data, In-house synthetic data	3.03M
Streaming Video	ActivityNet [101], YouCook2 [102], Ego4D-narration [103], Ego4D-livechat [104]	36.2K
Temporal Grounding	ActivityNet [101], YouCook2 [102], ViTT [105], QuerYD [106], HiREST [107], Charades-STA [108], Moment-10M [109], COIN [110]	0.21M
Image-only	LLaVA-SFT-665K [38], LLaVA-OV-SI [29]	0.88M
Text-only	Magpie [94], Tulu 3 [111]	1.56M

VideoLLaMA 3

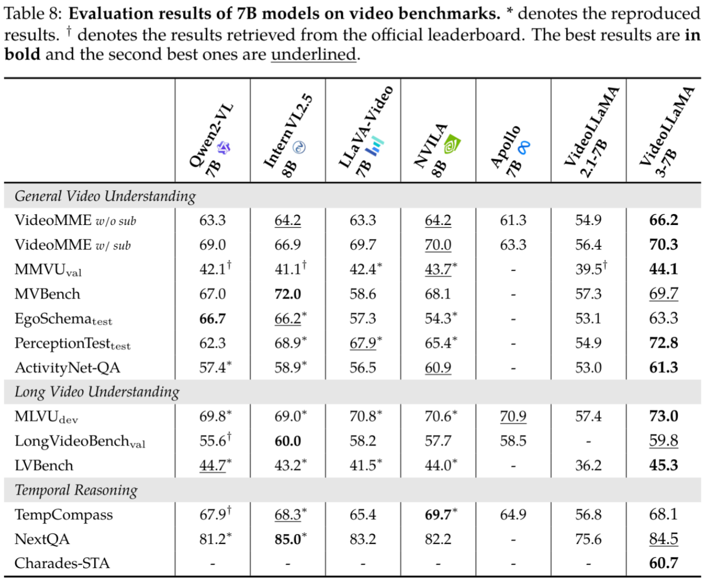


Table 8: **Evaluation results of 7B models on video benchmarks.** * denotes the reproduced results. † denotes the results retrieved from the official leaderboard. The best results are **in bold** and the second best ones are underlined.

	Qwen2-VL 7B	InternVL2.5 8B	LLaVA-Video 7B	NVILA 8B	Apollo 7B	VideoLLaMA 2.1-7B	VideoLLaMA 3-7B
General Video Understanding							
VideoMME *w/o sub*	63.3	64.2	63.3	64.2	61.3	54.9	**66.2**
VideoMME *w/ sub*	69.0	66.9	69.7	70.0	63.3	56.4	**70.3**
MMVU$_{\text{val}}$	$42.1^{\dagger}$	$41.1^{\dagger}$	42.4^{*}	43.7^{*}	-	$39.5^{\dagger}$	**44.1**
MVBench	67.0	**72.0**	58.6	68.1	-	57.3	69.7
EgoSchema$_{\text{test}}$	**66.7**	66.2^{*}	57.3	54.3^{*}	-	53.1	63.3
PerceptionTest$_{\text{test}}$	62.3	68.9^{*}	67.9^{*}	65.4^{*}	-	54.9	**72.8**
ActivityNet-QA	57.4^{*}	58.9^{*}	56.5	60.9	-	53.0	**61.3**
Long Video Understanding							
MLVU$_{\text{dev}}$	69.8^{*}	69.0^{*}	70.8^{*}	70.6^{*}	70.9	57.4	**73.0**
LongVideoBench$_{\text{val}}$	$55.6^{\dagger}$	**60.0**	58.2	57.7	58.5	-	59.8
LVBench	44.7^{*}	43.2^{*}	41.5^{*}	44.0^{*}	-	36.2	**45.3**
Temporal Reasoning							
TempCompass	$67.9^{\dagger}$	68.3^{*}	65.4	**69.7^{*}**	64.9	56.8	68.1
NextQA	81.2^{*}	**85.0^{*}**	83.2	82.2	-	75.6	84.5
Charades-STA	-	-	-	-	-	-	**60.7**

AdaReTaKe

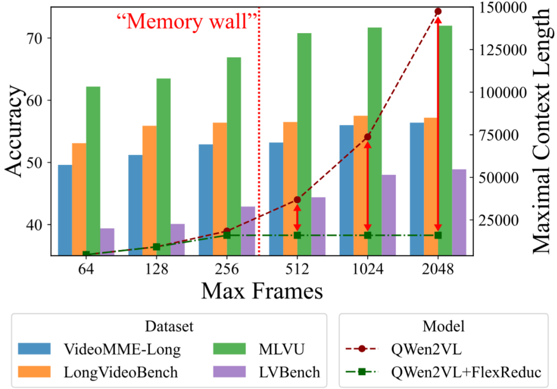


Figure 1: AdaReTaKe enables MLLM to perceive longer with fixed context length for video-language understanding.

AdaReTaKe

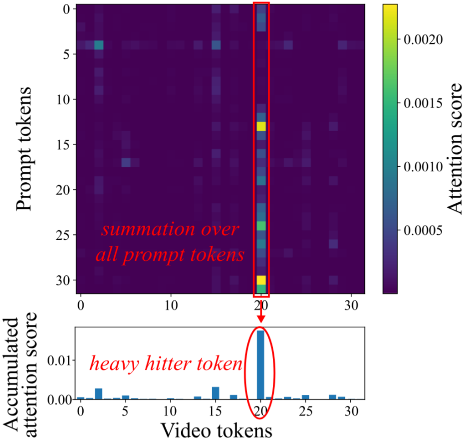


Figure 2: Illustrating example of a heavy hitter. We adopt the heavy hitter ratio to measure the redundancy

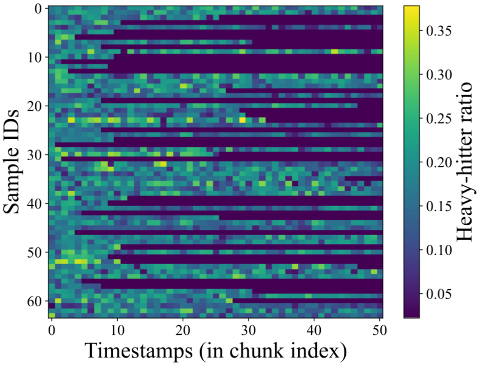
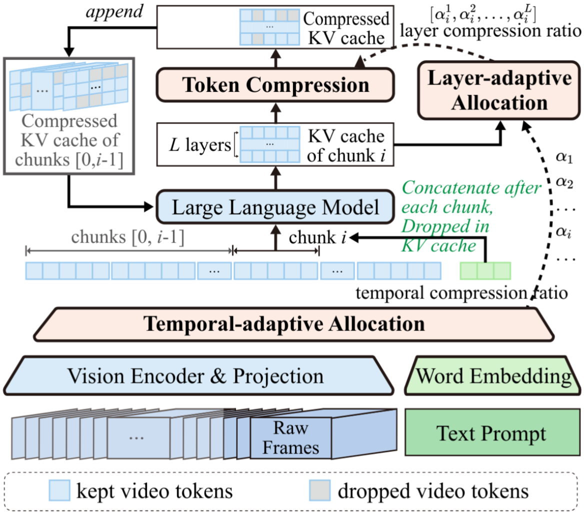


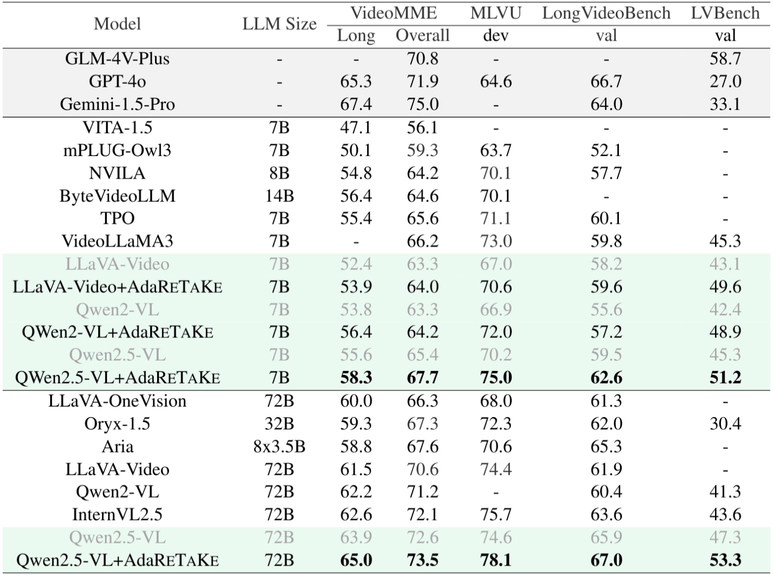
Figure 3: Heavy-hitter ratio among timestamps, showing the unevenly distributed temporal redundancy. The horizontal shaded bars indicate timestamps where the video has ended.

AdaReTaKe



append
Compressed KV cache
$[\alpha_i^1, \alpha_i^2, \ldots, \alpha_i^L]$
layer compression ratio
Compressed KV cache of chunks [0,*i*-1]
Token Compression
Layer-adaptive Allocation
L layers
KV cache of chunk *i*
α_1
α_2
$\ldots$
α_i
$\ldots$
Large Language Model
Concatenate after each chunk, Dropped in KV cache
chunks [0, *i*-1]
chunk *i*
temporal compression ratio
Temporal-adaptive Allocation
Vision Encoder & Projection
Word Embedding
Raw Frames
Text Prompt
kept video tokens
dropped video tokens

AdaReTaKe



Model	LLM Size	VideoMME		MLVU	LongVideoBench	LVBench
		Long	Overall	dev	val	val
GLM-4V-Plus	-	-	70.8	-	-	58.7
GPT-4o	-	65.3	71.9	64.6	66.7	27.0
Gemini-1.5-Pro	-	67.4	75.0	-	64.0	33.1
VITA-1.5	7B	47.1	56.1	-	-	-
mPLUG-Owl3	7B	50.1	59.3	63.7	52.1	-
NVILA	8B	54.8	64.2	70.1	57.7	-
ByteVideoLLM	14B	56.4	64.6	70.1	-	-
TPO	7B	55.4	65.6	71.1	60.1	-
VideoLLaMA3	7B	-	66.2	73.0	59.8	45.3
LLaVA-Video	7B	52.4	63.3	67.0	58.2	43.1
LLaVA-Video+AdaReTaKe	7B	53.9	64.0	70.6	59.6	49.6
Qwen2-VL	7B	53.8	63.3	66.9	55.6	42.4
QWen2-VL+AdaReTaKe	7B	56.4	64.2	72.0	57.2	48.9
Qwen2.5-VL	7B	55.6	65.4	70.2	59.5	45.3
QWen2.5-VL+AdaReTaKe	7B	**58.3**	**67.7**	**75.0**	**62.6**	**51.2**
LLaVA-OneVision	72B	60.0	66.3	68.0	61.3	-
Oryx-1.5	32B	59.3	67.3	72.3	62.0	30.4
Aria	8x3.5B	58.8	67.6	70.6	65.3	-
LLaVA-Video	72B	61.5	70.6	74.4	61.9	-
Qwen2-VL	72B	62.2	71.2	-	60.4	41.3
InternVL2.5	72B	62.6	72.1	75.7	63.6	43.6
Qwen2.5-VL	72B	63.9	72.6	74.6	65.9	47.3
Qwen2.5-VL+AdaReTaKe	72B	**65.0**	**73.5**	**78.1**	**67.0**	**53.3**

Conclusion

We reviewed following topics:

- **Benchmarking video recognition models.** Currently there are only static benchmarks with close-ended questions. Open-ended questions are rare and measured with GPT-Score or empirical pipelines.
- **Architectures of video LLMs.** LLaVA-like models are prevailing. Since videos are seq. of frames, a lot of attention is paid to the training of a quality visual encoder. After that various adaptor are used to compress video context into reasonable amounth of tokens.