Computer Science Ph.D. Applicant

Better and Longer Video Understanding

ENXIN SONG

Research Overview

Efficient Long-Sequence Modeling

Long Video

- MovieChat: From Dense Token to Sparse Memory for Long Video Understanding, CVPR 2024
- MovieChat+: Question-aware Sparse Memory for Long Video Question Answering, TPAMI 2025

Detailed Caption

- AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark, ICLR 2025

Efficient Architecture

- AuroraLong: Bringing RNNs Back to Efficient Open-Ended Video Understanding, ICCV 2025
- VideoNSA: Native Sparse Attention Scales Video Understanding, Undereview

Benchmarking and Evaluation

Knowledge

- Video-MMLU: A Massive Multi-Discipline Lecture Understanding Benchmark, ICCV 2025 Findings

Applications of Generative Models

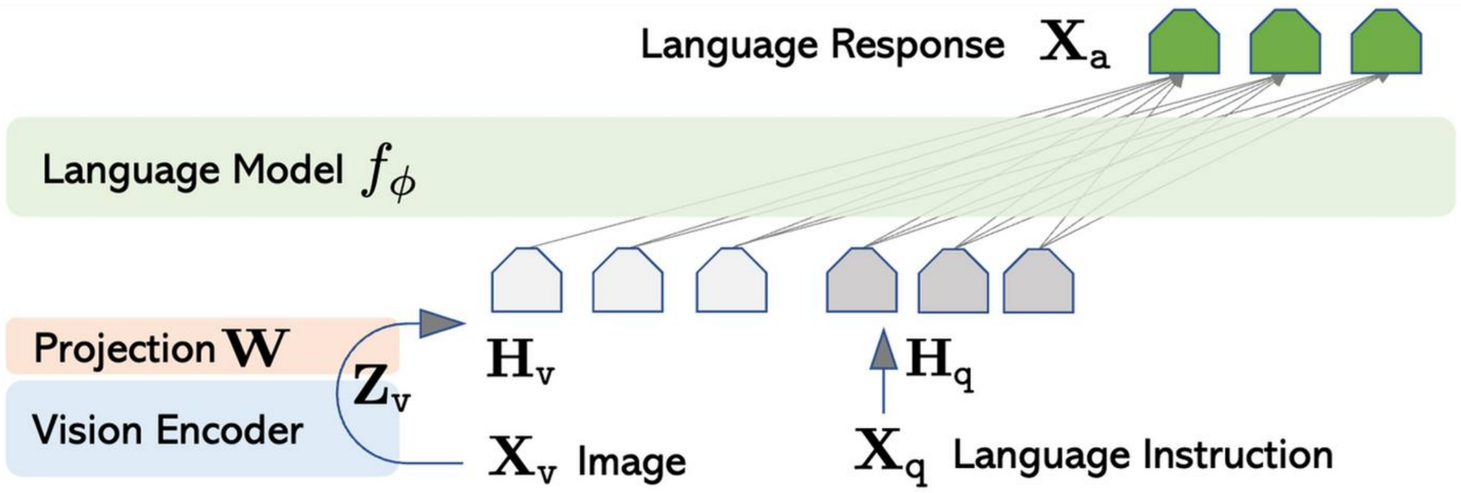
Masked Image Modeling

- Fantasy: Transformer Meets Transformer in Text-to-Image Generation
- Meissonic: Revitalizing Masked Generative Transformers for Efficient High-Resolution Text-to-Image Synthesis, ICLR 2025

Video LLMs

How We Connect?

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

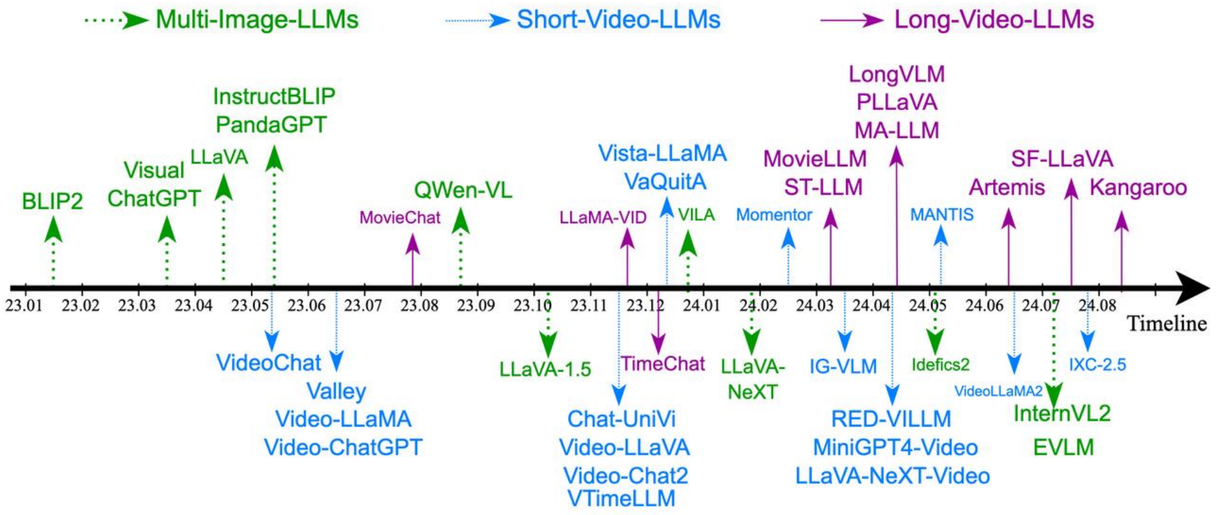


Video LLMs

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.*

Long-form Video Understanding



Why we need long-form video understand

Temporal Complexity and Granularity, Narrative Comprehension, Real-World Applications,*etc*

What are the current challenges?

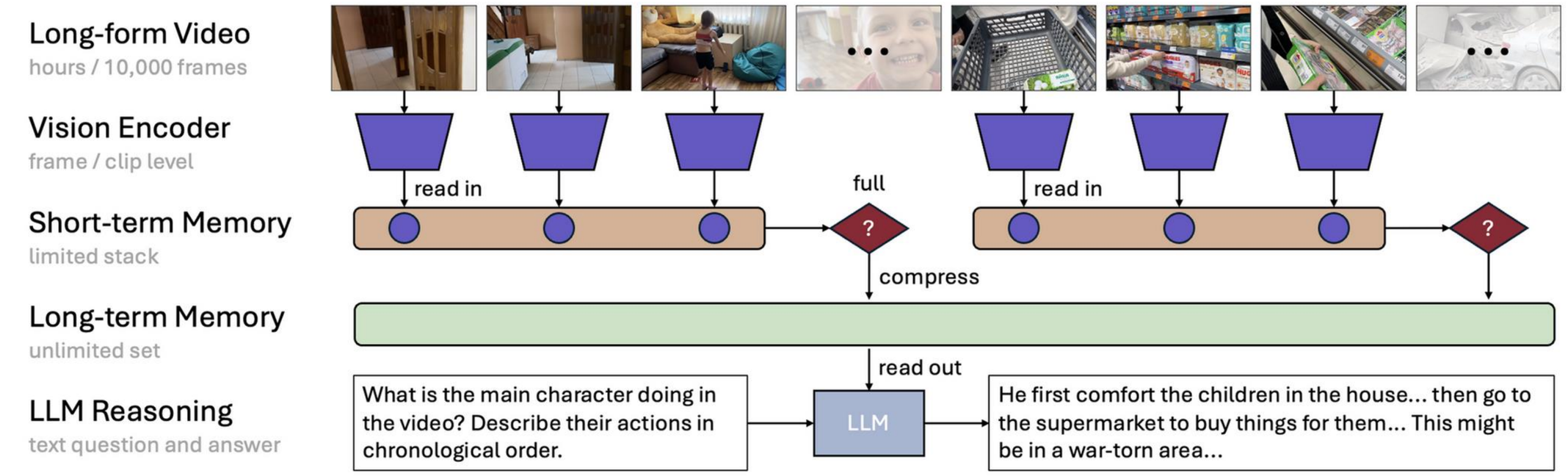
Efficiency, Training Data, *etc*

Can we do that with current LMMs?

Yes! We found that the LMMs trained on images and short videos can be adapted to long-form video tasks even without further fine-tuning.

MovieChat

First ever video understanding system that can take over 10,000 frames as input.



MovieChat

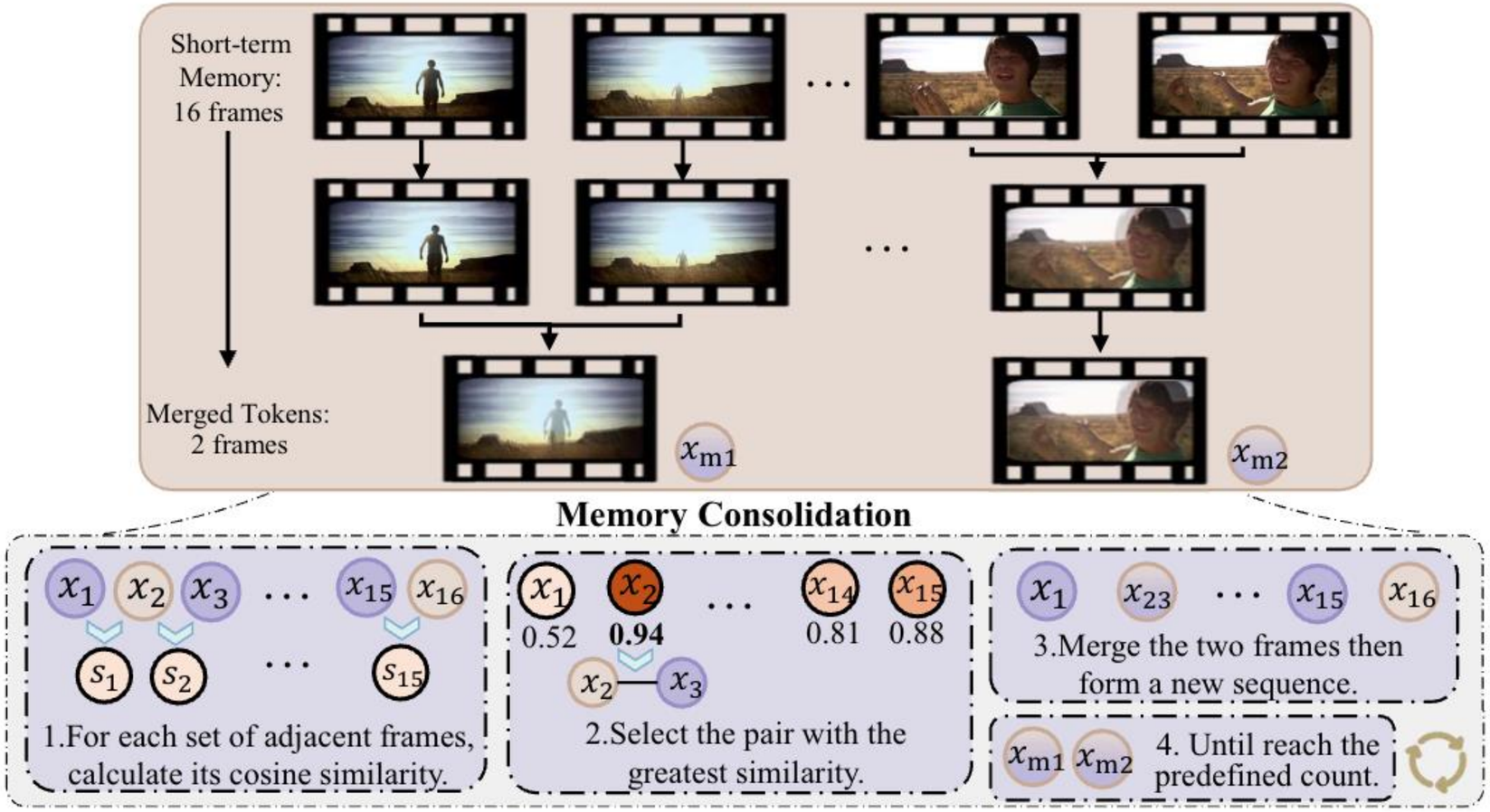
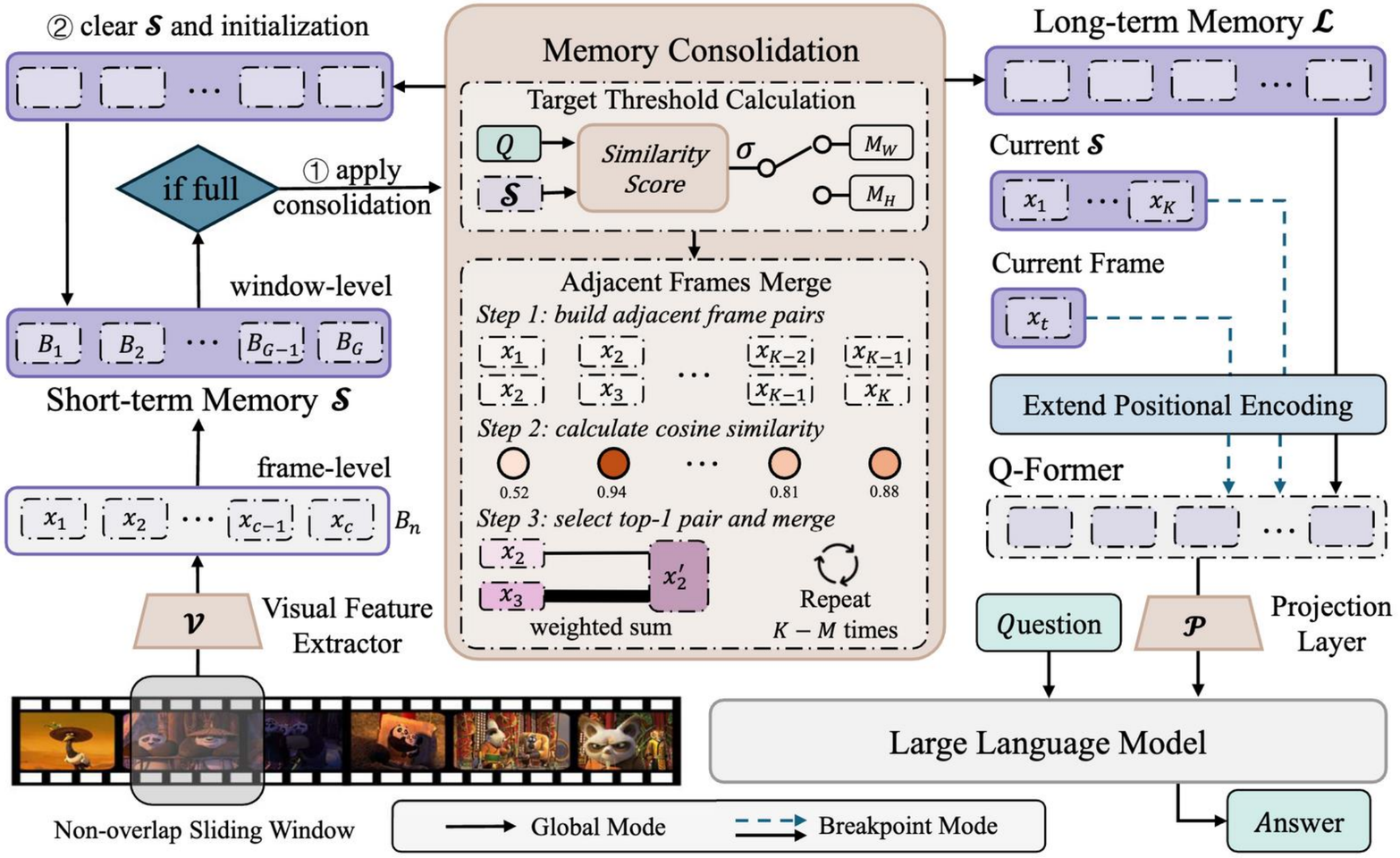


Figure: Memory Compression in MovieChat

MovieChat+


② clear $\mathcal{S}$ and initialization
Memory Consolidation
Target Threshold Calculation
Q
$\mathcal{S}$
Similarity Score
σ
M_W
M_H
Long-term Memory $\mathcal{L}$
Current $\mathcal{S}$
x_1 ⋯ x_K
Current Frame
x_t
if full
① apply consolidation
window-level
B_1 B_2 ⋯ B_{G-1} B_G
Short-term Memory $\mathcal{S}$
Adjacent Frames Merge
Step 1: build adjacent frame pairs
x_1 x_2 | x_2 x_3 | ⋯ | x_{K-2} x_{K-1} | x_{K-1} x_K
Step 2: calculate cosine similarity
0.52 0.94 ⋯ 0.81 0.88
Step 3: select top-1 pair and merge
x_2 x_3 → x_2'
weighted sum
Repeat $K - M$ times
Extend Positional Encoding
Q-Former
frame-level
x_1 x_2 ⋯ x_{c-1} x_c B_n
$\mathcal{V}$
Visual Feature Extractor
Question
$\mathcal{P}$
Projection Layer
Large Language Model
Answer
Non-overlap Sliding Window
Global Mode
Breakpoint Mode

MovieChat+

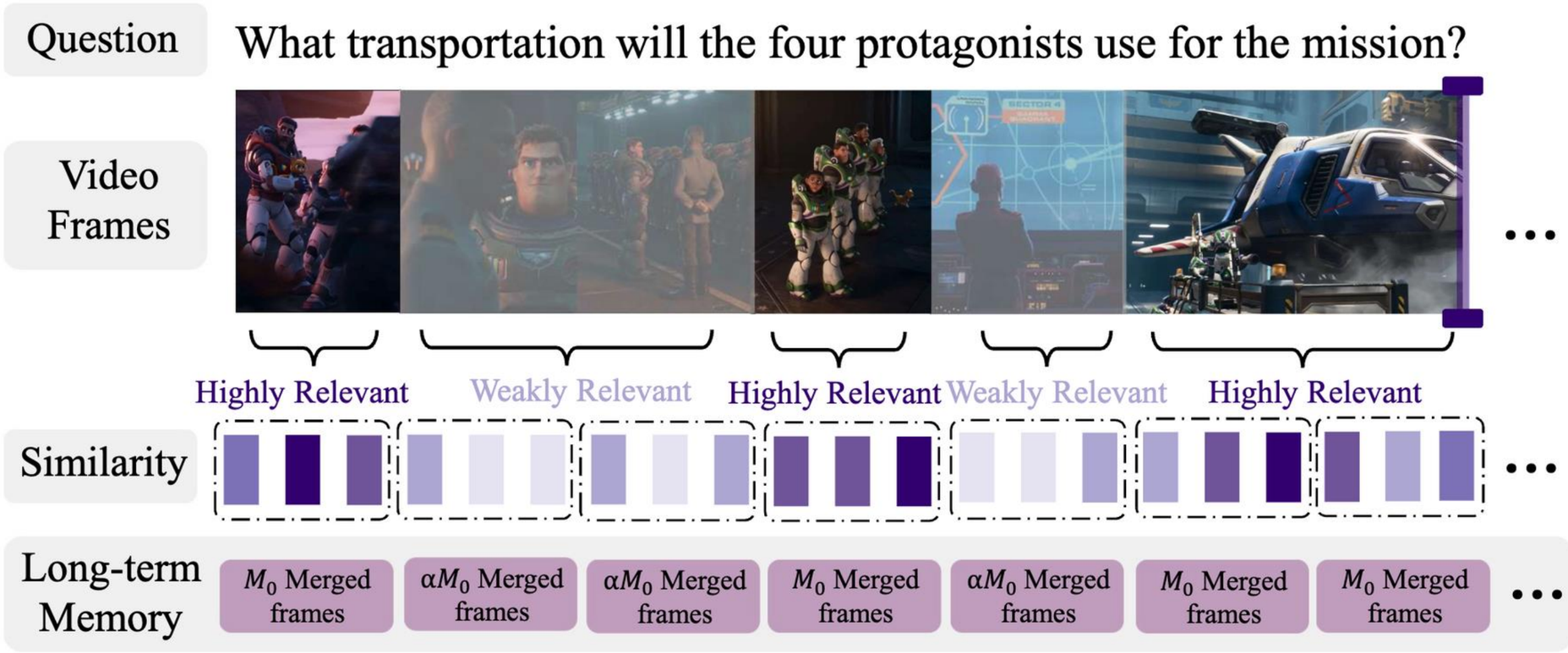


Figure: Question-aware memory selection in MovieChat+

MovieChat+

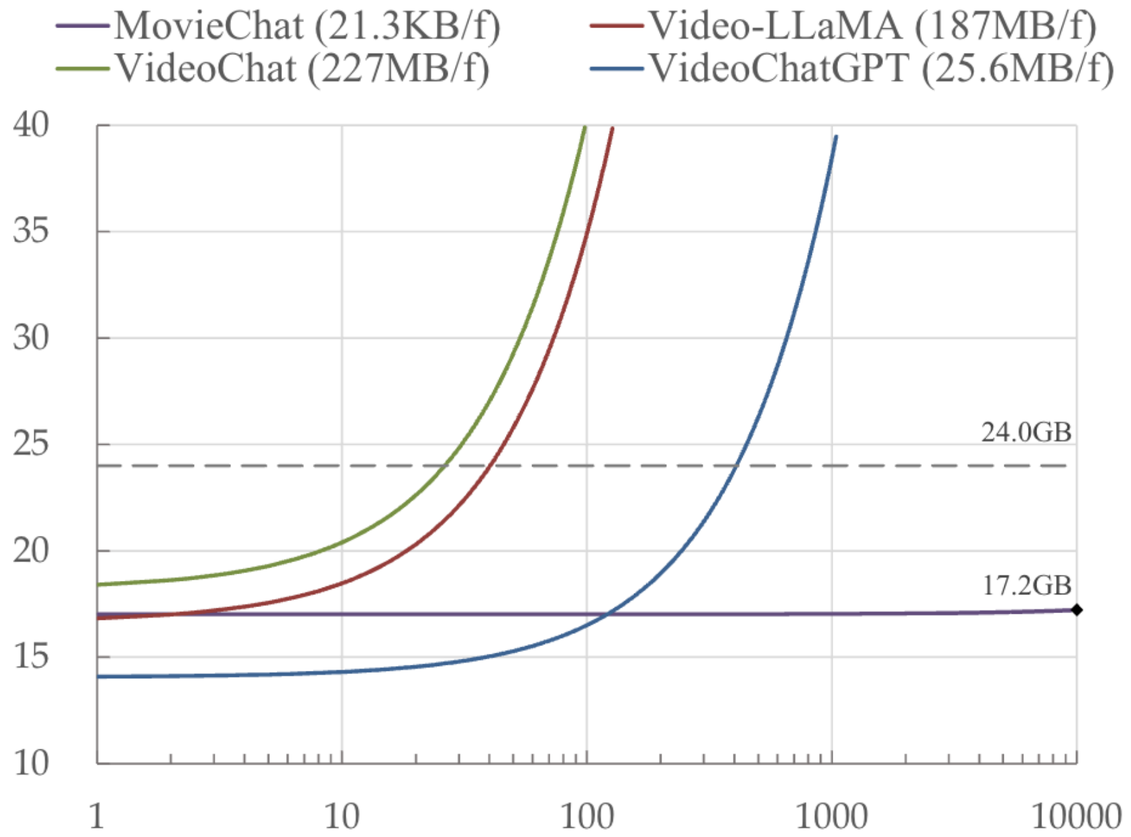


Figure: Video random-access memory (VRAM) cost under gigabyte (GB) (y-axis) v.s. frame number (x-axis) comparison.

MovieChat-1K

Benchmark	*Labels*	*#Eval Videos*	*#Eval QAs*	*Avg Duration (s)*	*Released Date*
MSVD-QA [5]	Auto	520	13,157	10	2011
MSRVTT-QA [6]	Auto	2,990	72,821	15	2017
ActivityNet-QA [7]	Human	800	8,000	180	2019
NeXT-QA [8]	Human	1,000	8,564	44	2021
MovieChat-1K [3]	Human	130	1,950	**564**	**2023.7**
EgoSchema [9]	Auto	5,031	5,031	180	2023.8
MVBench [10]	Auto	4,000	4,000	16	2023.11
LongVideoBench [11]	Human	3,763	6,678	473	2024.7

Table: The popular benchmarks for video question answering.

MovieChat

Method	MSVD-QA		MSRVTT-QA		ActivityNet-QA		NExT-QA	
	Acc.	Sco.	Acc.	Sco.	Acc.	Sco.	Acc.	Sco.
FrozenBiLM	2.2	–	16.8	–	24.7	–	–	–
VideoChat	56.3	2.8	45.0	2.5	26.5	2.2	**56.6**	**3.2**
LLaMA Adapter	54.9	3.1	43.8	2.7	34.2	2.7	–	–
VideoLLaMA	51.6	2.5	29.6	1.8	12.4	1.1	–	–
Video-ChatGPT	64.9	3.3	49.3	**2.8**	35.2	2.7	54.6	**3.2**
MovieChat	75.2	3.8	52.7	2.6	45.7	**3.4**	49.9	2.7
MovieChat+	**76.5**	**3.9**	**53.9**	2.7	**48.1**	**3.4**	54.8	3.0

Table: Quantitative evaluation for short video question answering.

MovieChat

Method	Text Decoder	# Frames	Global Mode		Breakpoint Mode	
			Acc.	Sco.	Acc.	Sco.
GIT	non-LLM based	6	28.8	1.83	29.2	1.98
mPLUG-2	non-LLM based	8	31.7	2.13	30.8	1.83
VideoChat	LLM based	32	57.8	3.00	46.1	2.29
VideoLLaMA	LLM based	32	51.7	2.67	39.1	2.04
Video-ChatGPT	LLM based	100	47.6	2.55	48.0	2.45
MovieChat	LLM based	2048	<u>62.3</u>	<u>3.23</u>	<u>48.3</u>	<u>2.57</u>
MovieChat+	LLM based	2048	**71.2**	**3.51**	**49.6**	**2.62**

Table: Quantitative evaluation for long video question answering on MovieChat-1K test set.

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 Detailed Captioning

AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark

ICLR 2025

VDC

Dataset	Theme	# Video	# Clip	# Caption	# Word	# Vocab.	Ave. Length
MSVD	Open	1,970	1,970	70,028	607,339	13,010	8.67
MSR-VTT		7,180	10,000	200,000	1,856,523	29,316	9.28
ActivityNet		20,000	100,000	100,000	1,340,000	15,564	13.40
S-MiT		515,912	515,912	515,912	5,618,064	50,570	10.89
M-VAD	Movie	92	48,986	55,905	519,933	18,269	9.30
MPII-MD		94	68,337	68,375	653,467	24,549	9.56
Youcook2	Cooking	2,000	15,400	15,400	121,418	2,583	7.88
Charades	Human	9,848	10,000	27,380	607,339	13,000	22.18
VATEX		41,300	41,300	413,000	4994,768	44,103	12.09
VDC (ours)	Open	1,027	1,027	1,027	515,441	20,419	**500.91**

Table: Benchmark comparison for video captioning task. Ave. Length indicates the average number of words per caption.

VDC

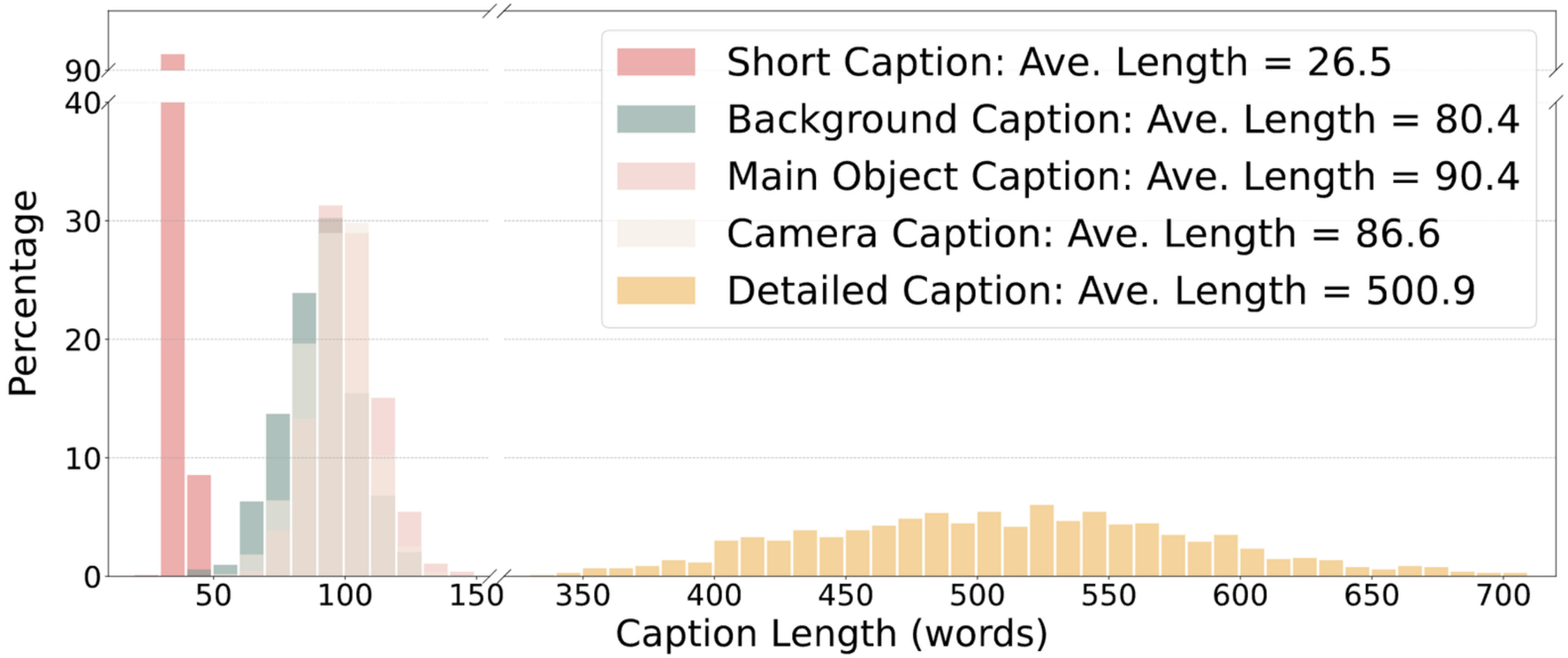
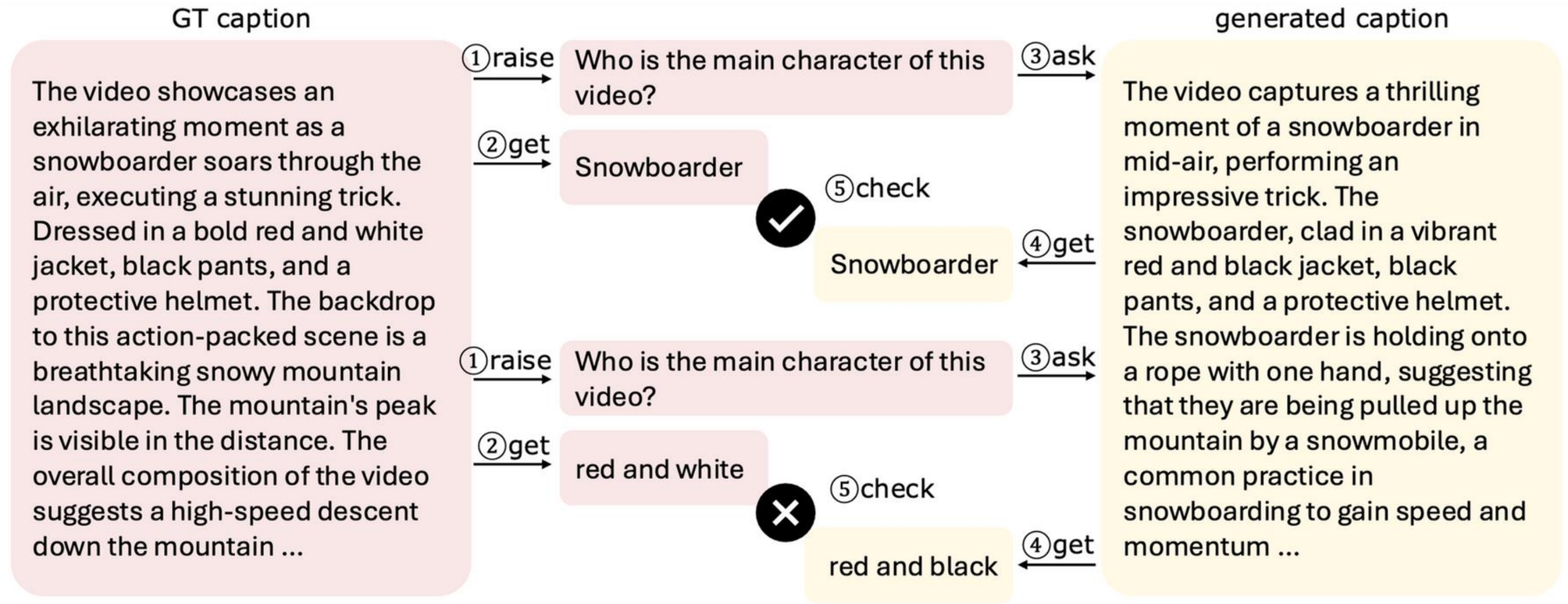


Figure: Distribution of structured caption length.

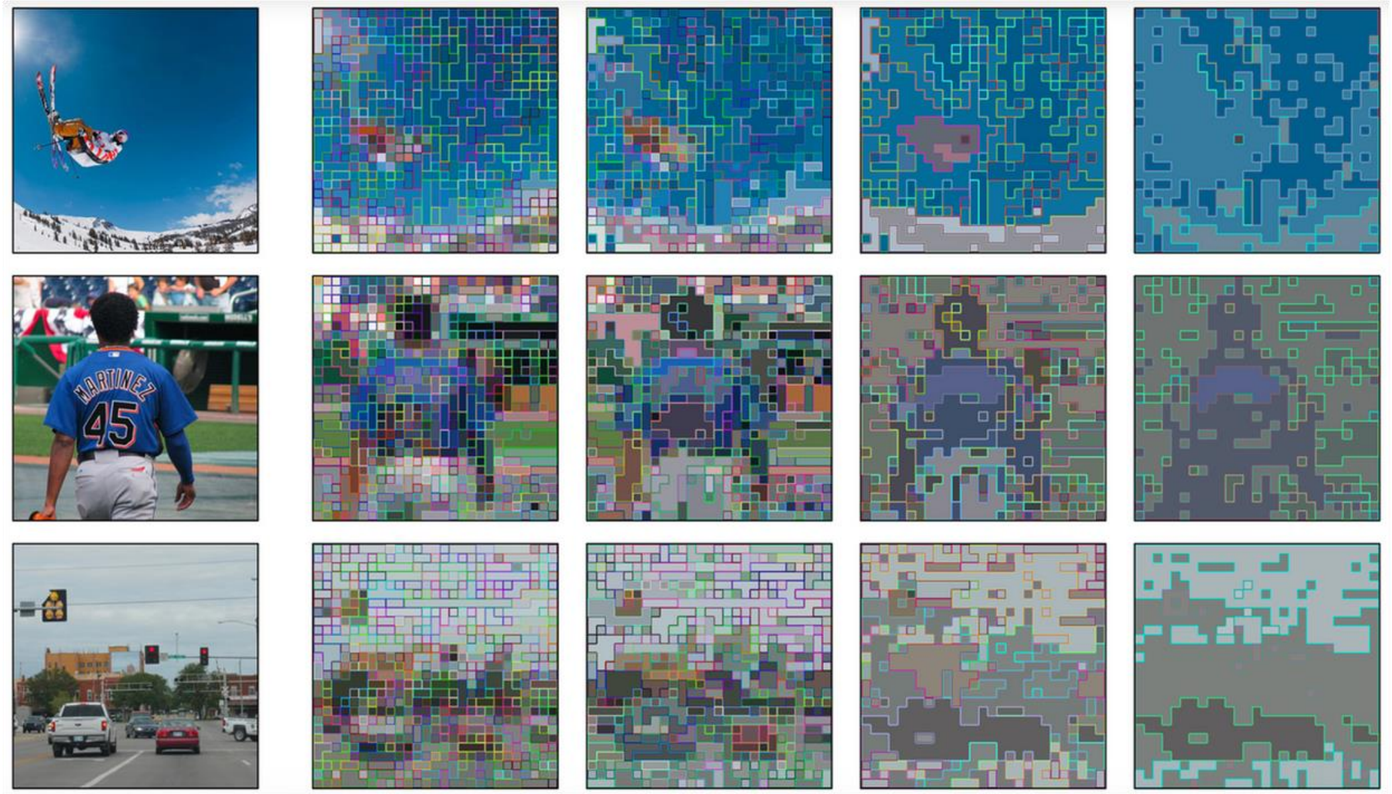
VDC

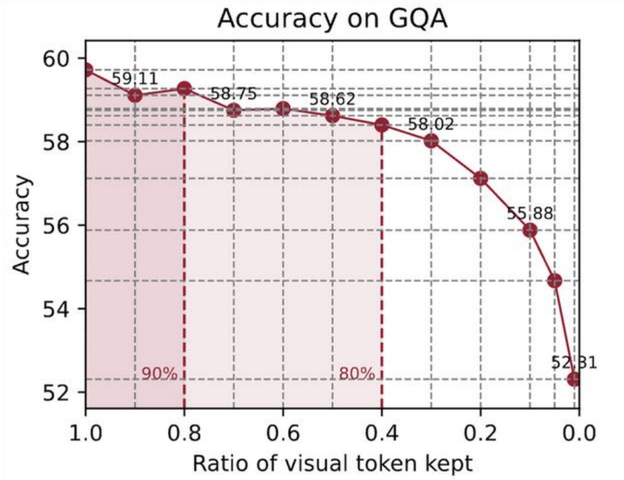
First ever evaluation system for detailed video captioning.



AuroraCap

5% tokens but 90% performace.





AuroraCap

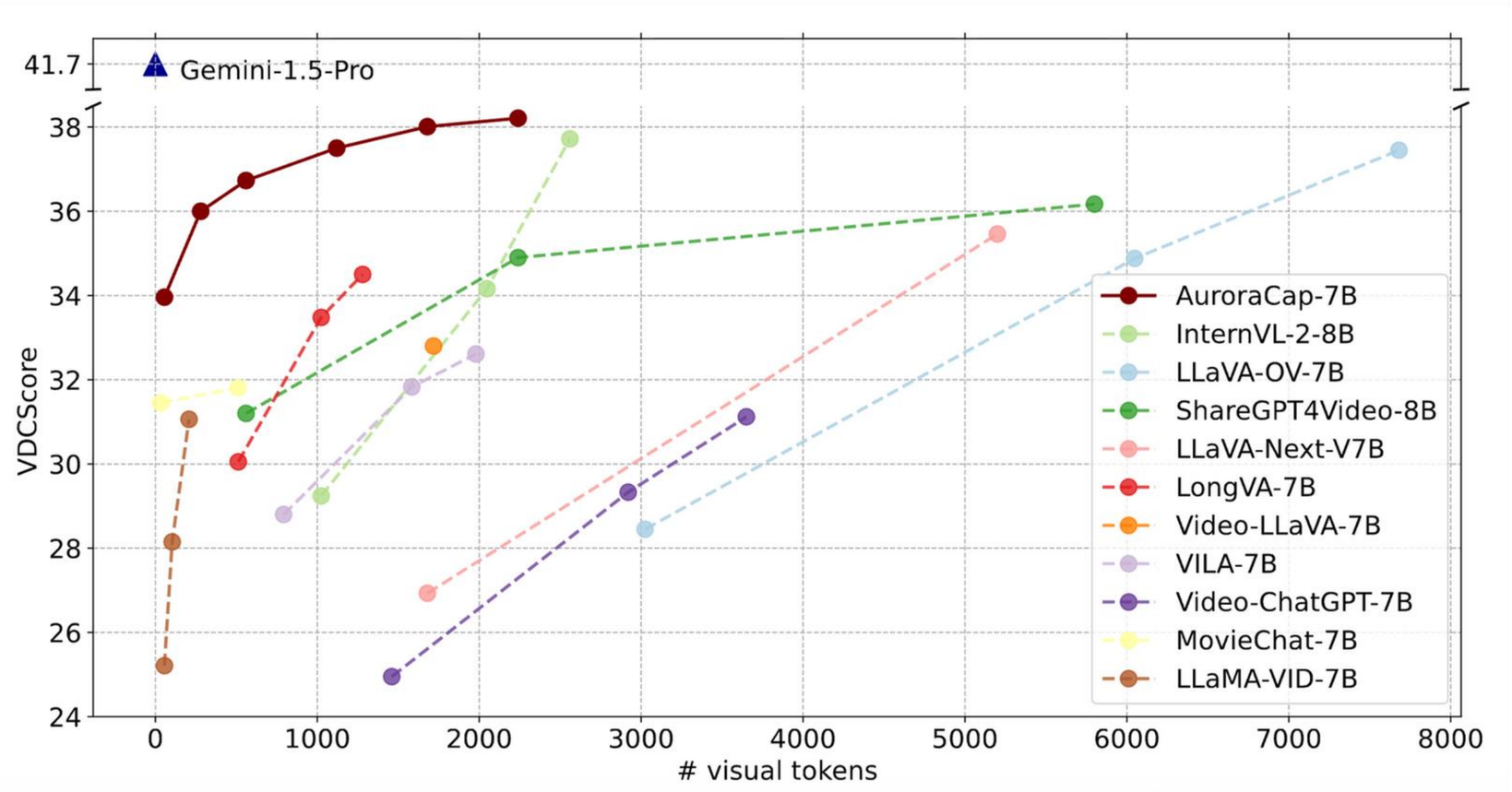
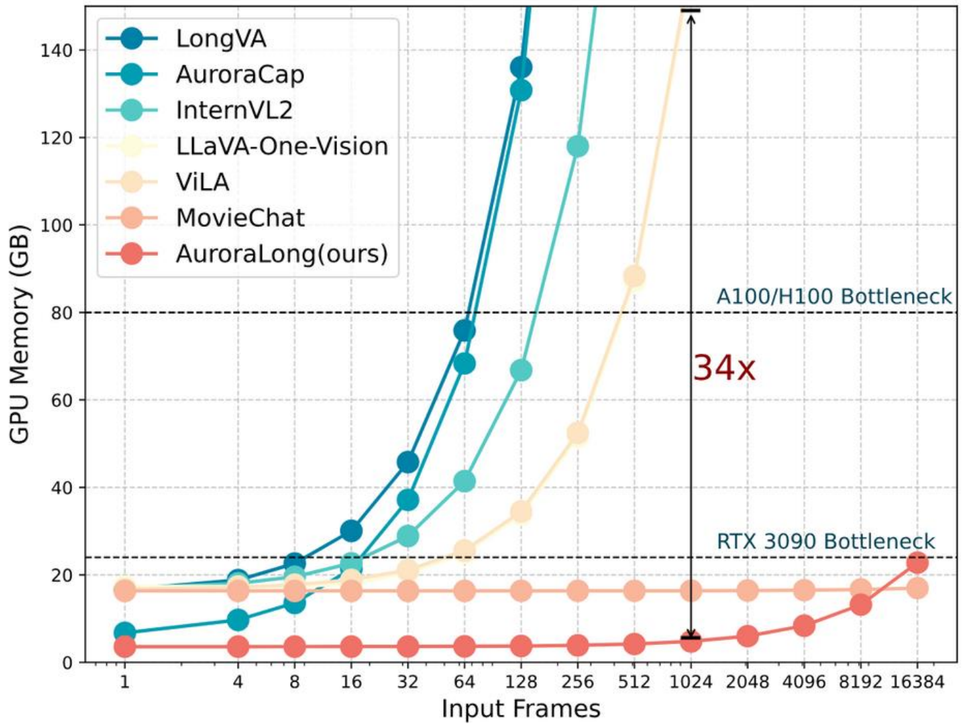


Figure: Comparison between various models with different number of visual tokens input on VDC.

Efficient Video Understanding

AuroraLong: Bringing RNNs Back to Efficient Open-Ended Video Understanding

ICCV 25 submission with score 4, 4, 6



AuroraLong

What can we do with Linear RNN-based LLMs?

Standard causal attention

$$o_t = \sum_{j=1}^{t} \frac{\exp(q_t^\top k_j)}{\sum_{j'=1}^{t} \exp(q_t^\top k_{j'})} v_j \quad \Rightarrow \quad O = \mathrm{softmax}(QK^\top + \log M)V$$

Approximate by removing softmax and denominator

$$o_t \approx \sum_{j=1}^{t} (q_t^\top k_j) v_j \quad \Rightarrow \quad O \approx (QK^\top \odot M)V$$

Rewriting as a linear recurrent form

$$o_t = \left(\sum_{j=1}^{t} v_j k_j^\top \right) q_t = S_t q_t, \quad S_t = S_{t-1} + v_t k_t^\top$$

AuroraLong

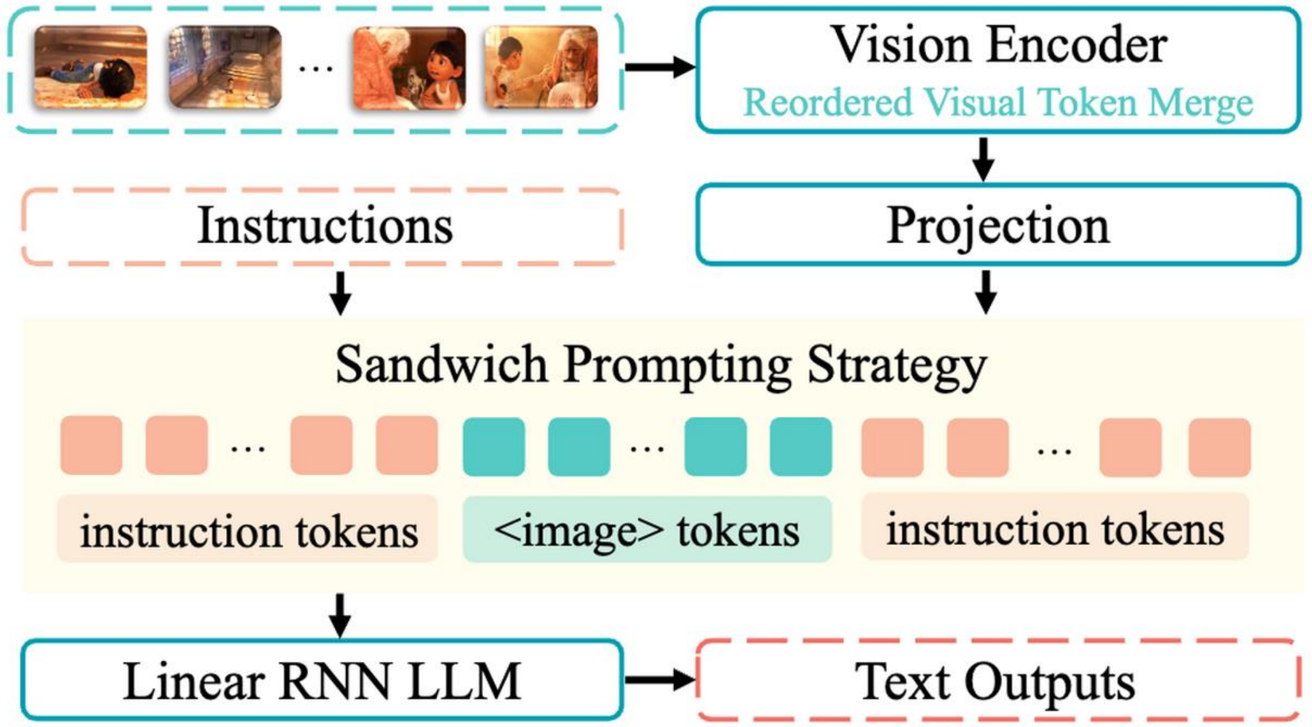


Figure: We explore reordered visual token merge and apply sandwich prompting strategy on the Linear RNN LLM.

AuroraLong

Models	Input	CTX	Size	MLVU								MovieChat-1K	
				AVG	AR	ER	AO	AC	TR	NQA	PQA	Global	Break
GPT4-o	0.5fps	-	-	**54.5**	**68.8**	47.8	46.2	35.0	**83.7**	42.9	57.1	-	-
LLAMA-VID [76]	1 fps	4k	7B	18.1	23.1	11.3	18.6	15.0	20.9	21.7	16.0	51.7	39.1
mPLUG-Owl-V [145]	16 frm	4k	7B	16.7	15.4	13.2	14.3	20.0	25.3	6.7	22.0	62.9	44.1
Video-ChatGPT [91]	16 frm	2k	7B	21.2	17.9	32.1	17.1	13.3	17.6	28.3	22.0	47.6	48.0
MovieChat [113]	2048 frm	4k	7B	16.5	10.3	15.1	17.1	15.0	18.7	23.3	16.0	62.3	48.3
Video-LLAVA [91]	8 frm	4k	7B	30.1	38.5	26.4	20.0	21.7	70.3	13.3	26.0	55.2	53.1
LLaVA-NeXT [83]	16 frm	8k	7B	27.1	17.9	26.4	21.4	16.7	63.7	13.3	30.0	45.8	55.2
ShareGPT4Video [17]	16 frm	8k	8B	34.2	25.6	45.3	17.1	8.3	73.6	31.7	38.0	69.0	60.9
InternVL-1.5 [25]	16 frm	8k	26B	37.9	51.3	24.5	14.3	13.3	80.2	40.0	42.0	57.7	61.1
LongVA [157]	256 frm	224k	7B	42.1	41.0	39.6	17.1	23.3	81.3	46.7	46.0	55.9	56.5
VILA-1.5 [78]	14 frm	276k	40B	46.2	56.4	35.8	34.3	11.7	84.7	38.3	**62.0**	57.2	60.1
Video-XL [110]	256 frm	132k	7B	46.3	28.2	41.5	**48.6**	31.7	78.0	**50.0**	46.0	-	-
LLaVA-OneVision* [65]	32 frm	132k	0.5B	50.3	58.5	52.4	28.6	30.9	67.0	33.3	42.8	-	-
Qwen2-VL* [128]	32 frm	132k	2B	48.7	54.7	47.6	30.9	28.6	73.8	40.4	60.5	-	-
InternVL2* [24]	32 frm	200k	2B	48.2	57.4	**57.1**	35.7	33.4	66.7	28.5	50.0	-	-
AURORALONG (ours)	48 frm	4k	2B	52.7	59.5	**57.1**	33.2	**42.9**	69.0	45.2	61.9	**84.0**	**64.0**

Table: Evaluation on video question-answering and video captioning tasks across different video length.

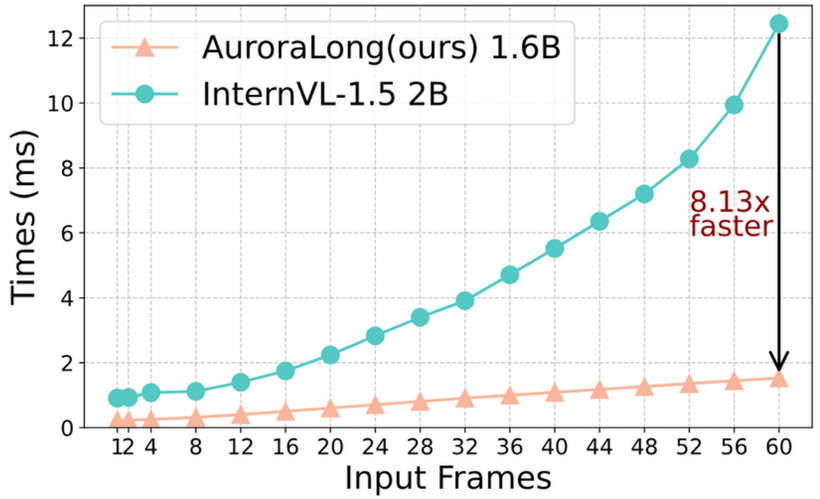


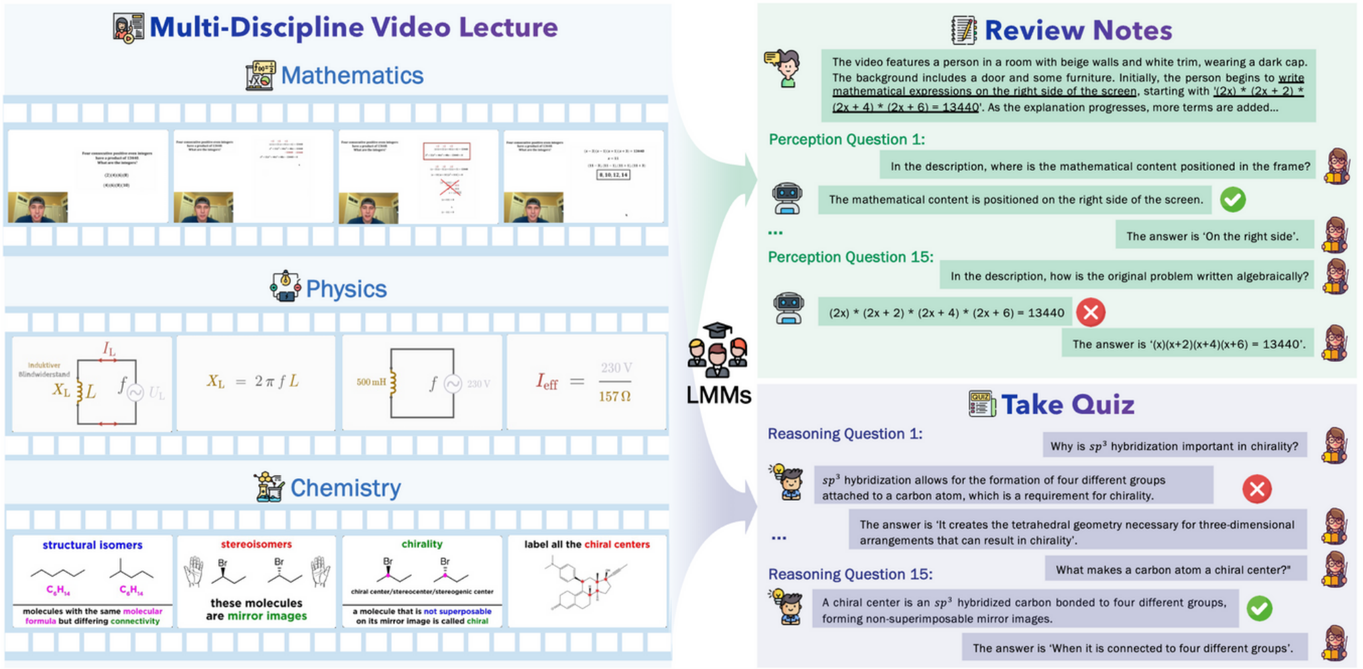
Figure: AuroraLong requires less computation and provides lower latency.

Lecture Video Understanding

Can video LLMs really understand real-world lectures?

NOT YET.

ICCV Findings



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

- 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？

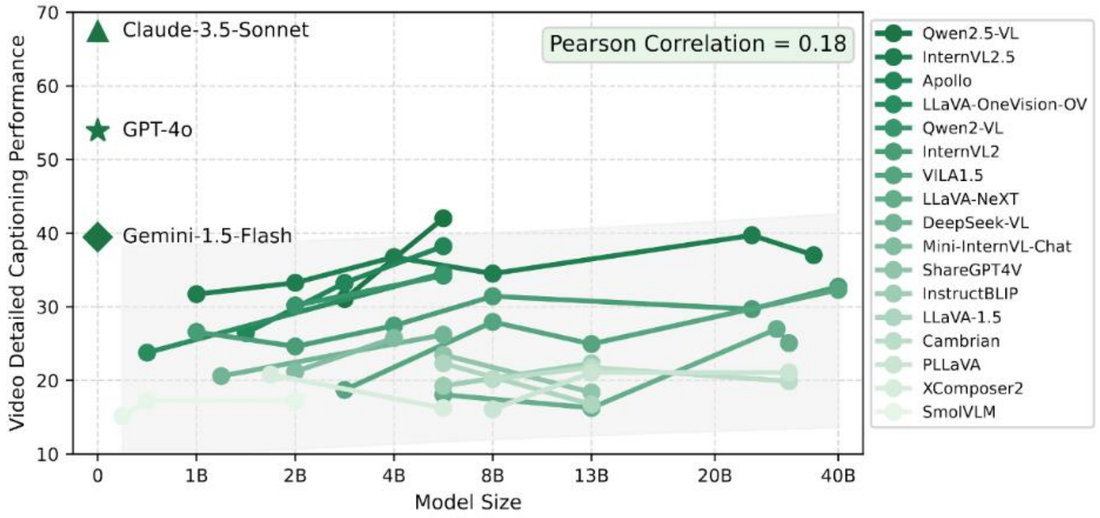
How the vision token influence performance？

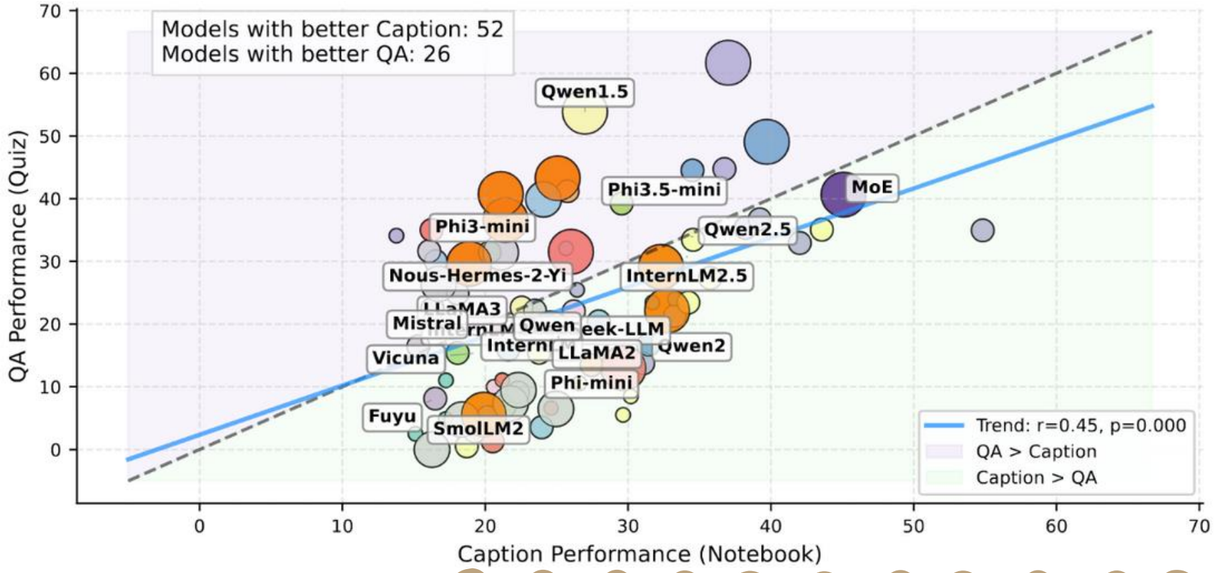
Video-MMLU

How the base LLMs influence performance？

Finding 1. Large scale LMMs do not show clear advantages over smaller ones.

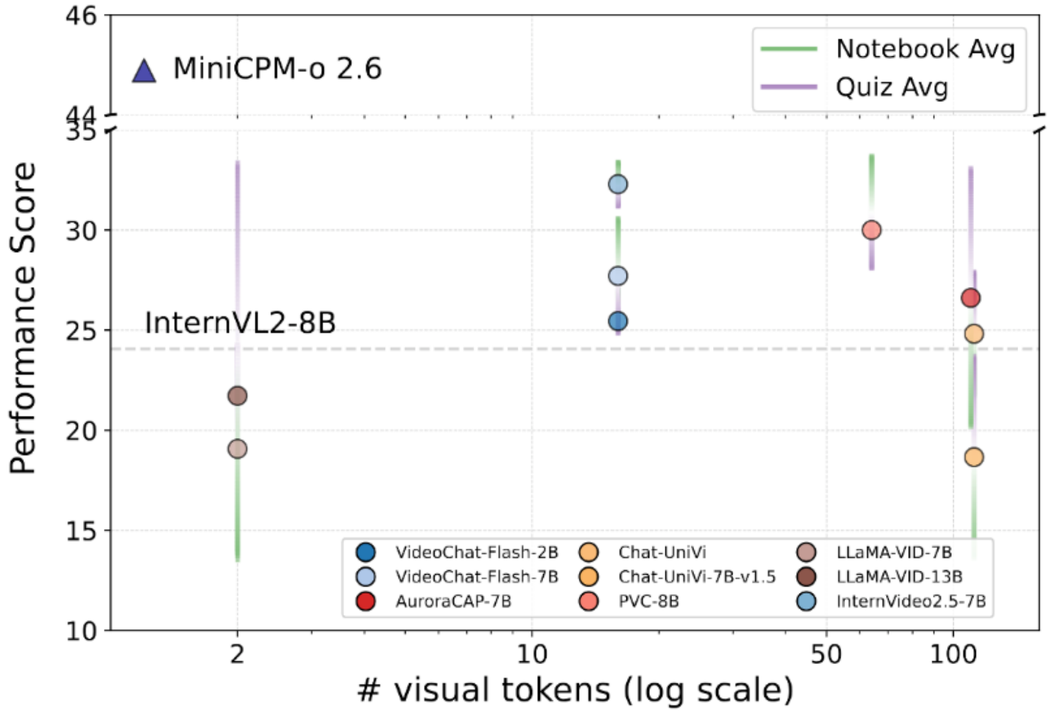
Although LMM scaling laws suggest significant performance improvements with increased model size, this trend is less pronounced in Video-MMLU. Model size shows a stronger correlation with performance in video QA compared to video captioning, implying reasoning benefits more from scaling.





Video-MMLU

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

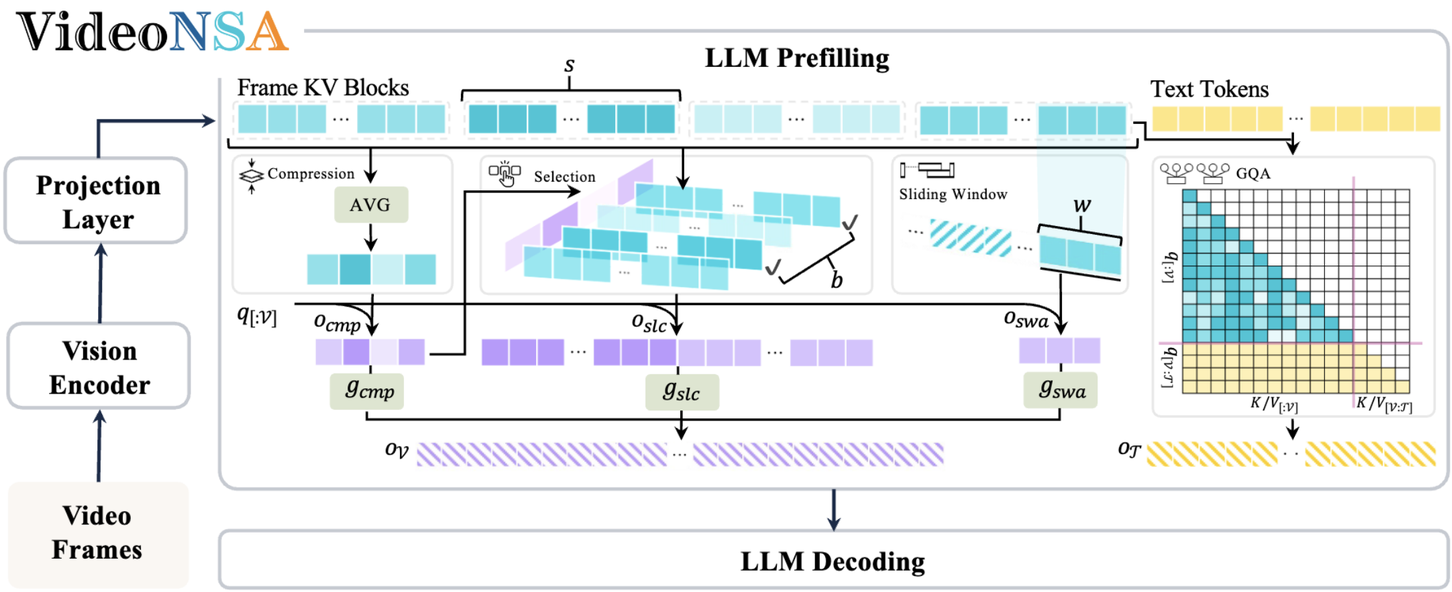


Efficient Video Understanding

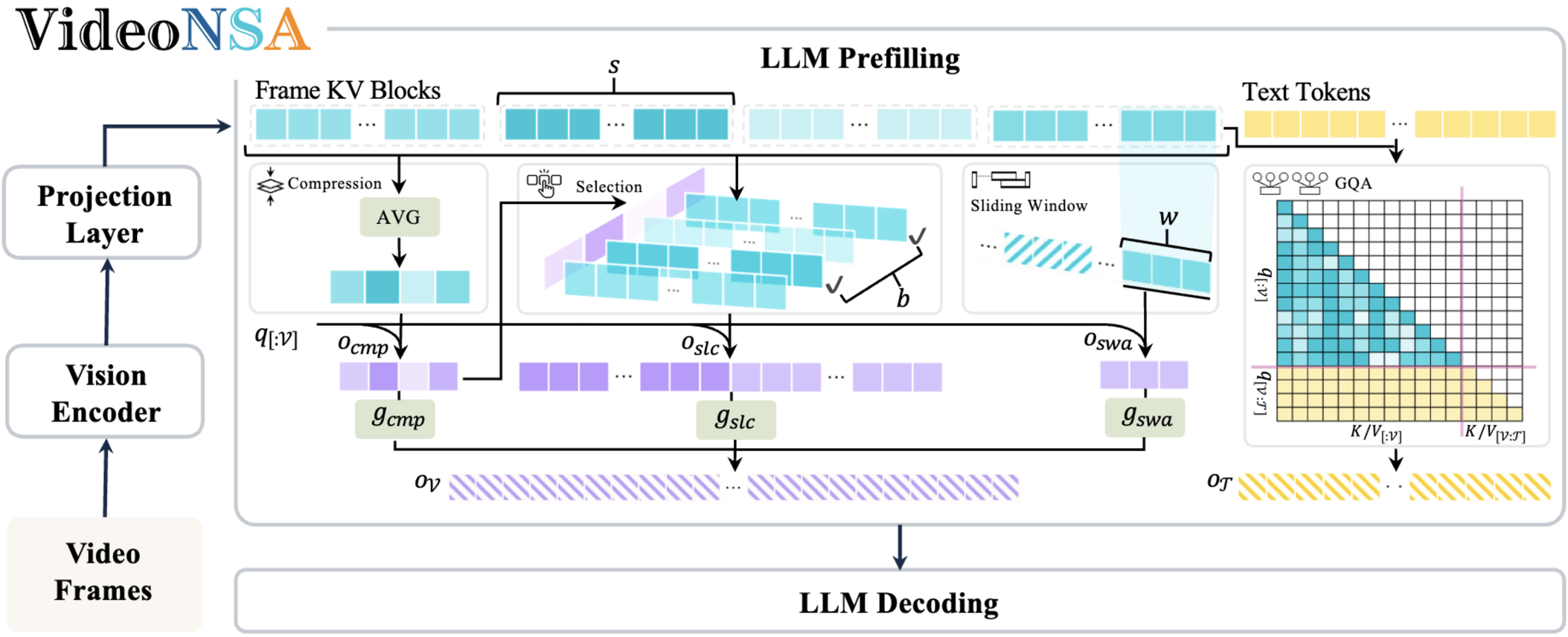
Token Compression causes irreversible information loss.

What can we do with sparse attention?

ICLR 26 submission with score 8,6,6,4



Efficient Video Understanding



VideoNSA

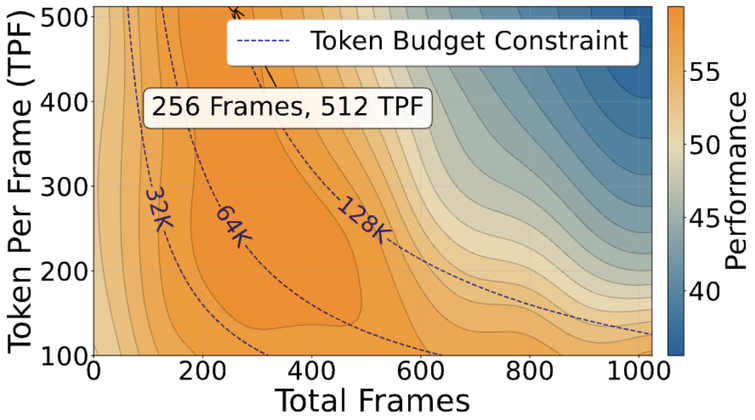
Do learned sparse attention weights remain beneficial in dense attention settings?

Table 3: Ablation study on transferring sparse attention weights to dense attention across tasks.

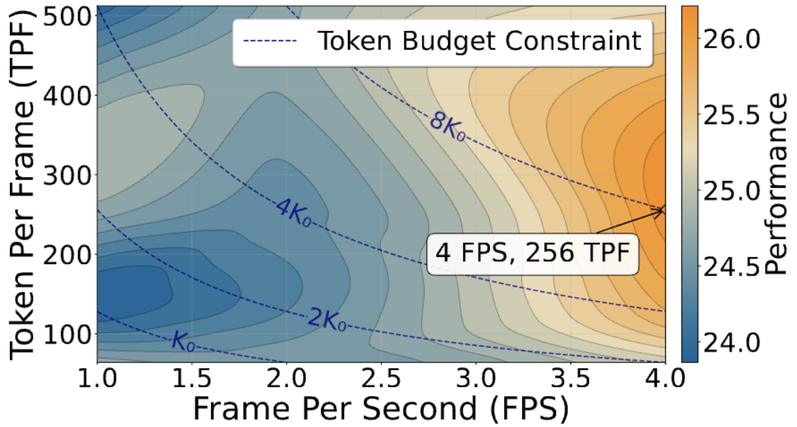
Model	Long Video Understanding				Temporal Reasoning	Spatial Understanding
	LongVideoBench	MLVU_{Test}	TimeScope	LongTimeScope	Tomato	VSIBench
Qwen2.5-VL-7B	58.7	51.2	81.0	40.7	22.6	29.7
Dense-SFT	57.8 (-1.5%)	51.2 (+0.0%)	76.8 (-5.2%)	40.2 (-1.2%)	21.7 (-4.0%)	30.6 (+2.1%)
Dense-NSA	56.1 (-4.4%)	51.6 (+0.8%)	**83.0 (+2.5%)**	40.9 (+0.5%)	23.4 (+3.5%)	33.1 (+10.7%)
VideoNSA	**59.4 (+1.1%)**	**51.8 (+1.2%)**	82.7 (+2.1%)	**44.4 (+9.1%)**	**26.2 (+15.9%)**	**36.1 (+20.3%)**

VideoNSA

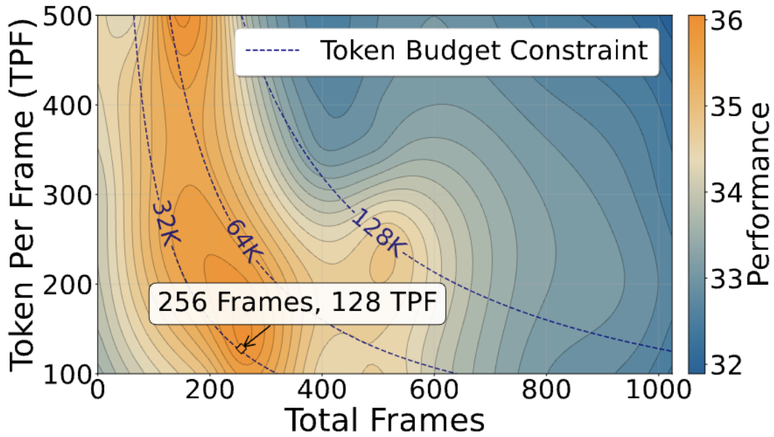
How far can VideoNSA scale in context length?



(a) Information Scaling of LongVideoBench



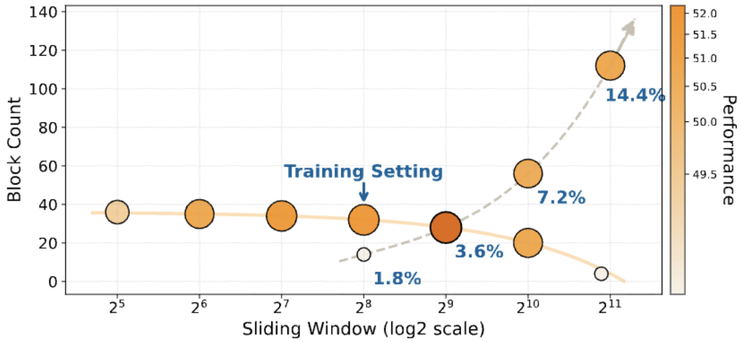
(c) Information Scaling of Tomato



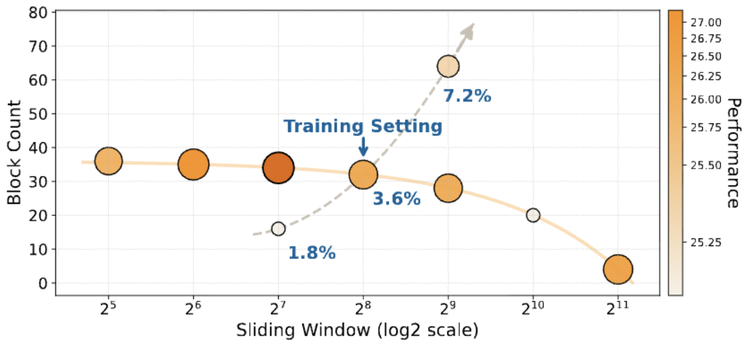
(d) Information Scaling of VSIBench

VideoNSA

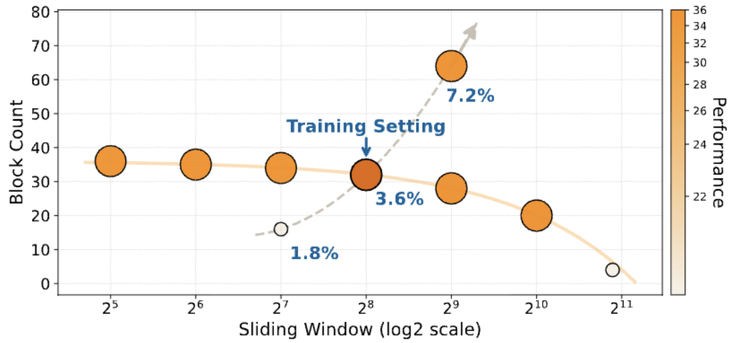
How to allocate the attention budget?



(a) Attention Scaling of MLVU



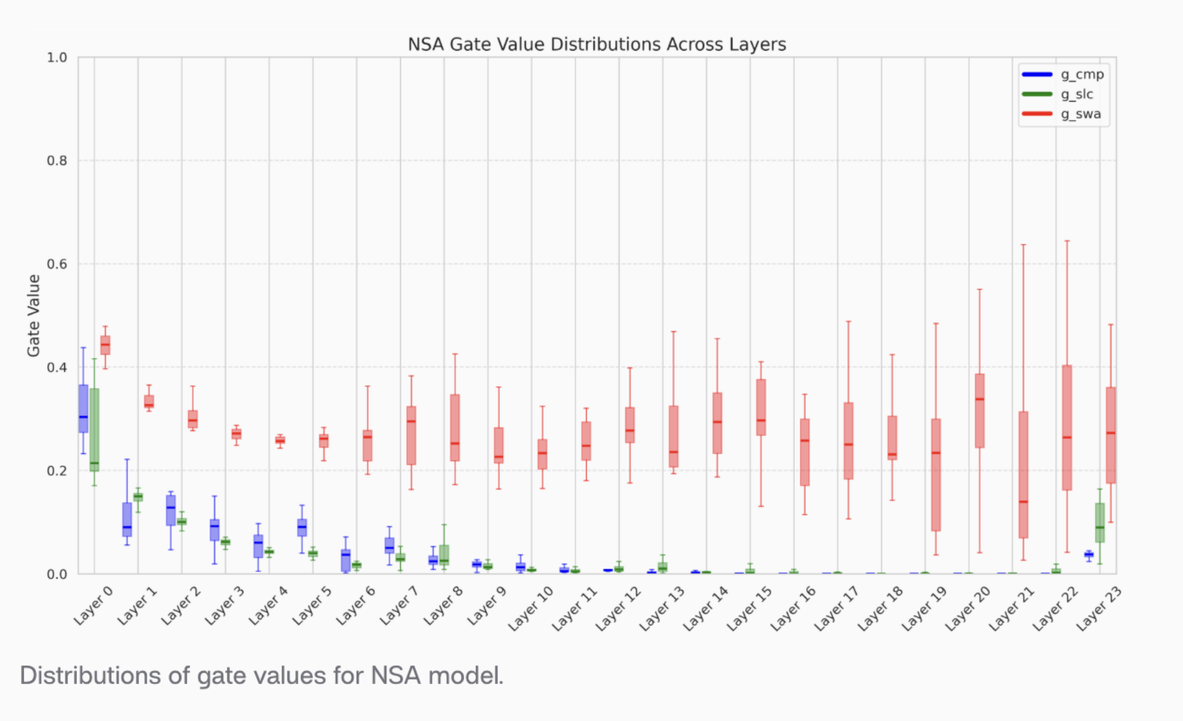
(c) Attention Scaling of Tomato



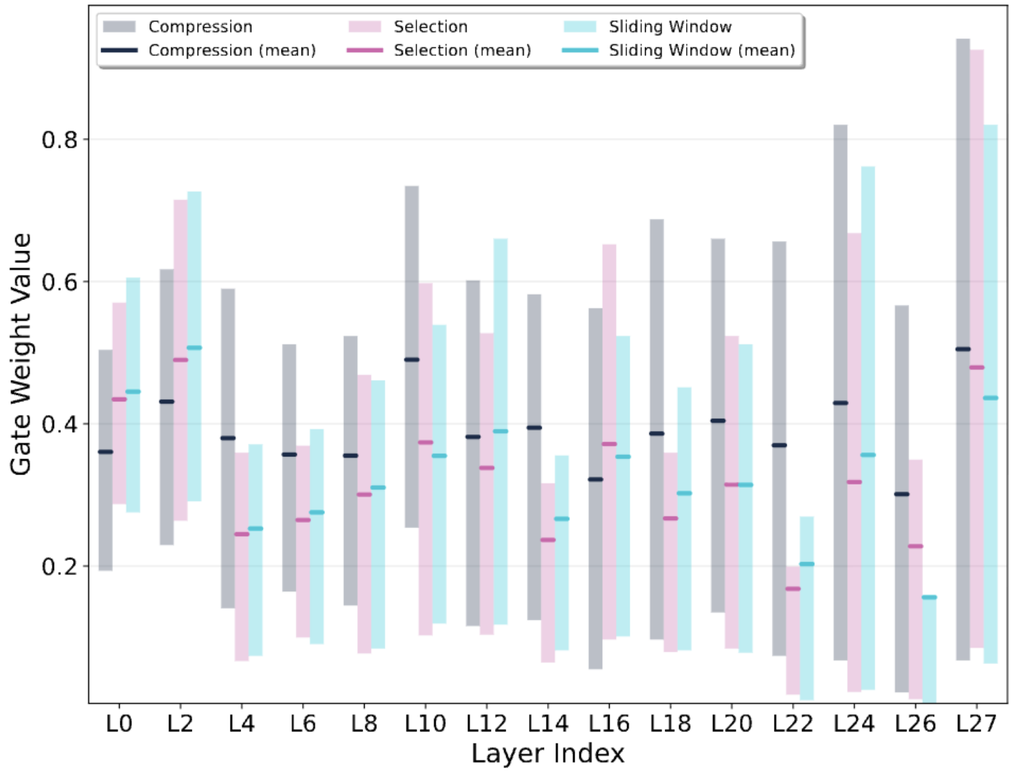
(d) Attention Scaling of VSIBench

VideoNSA

What roles do compression, selection, and sliding-window gates play in VideoNSA?

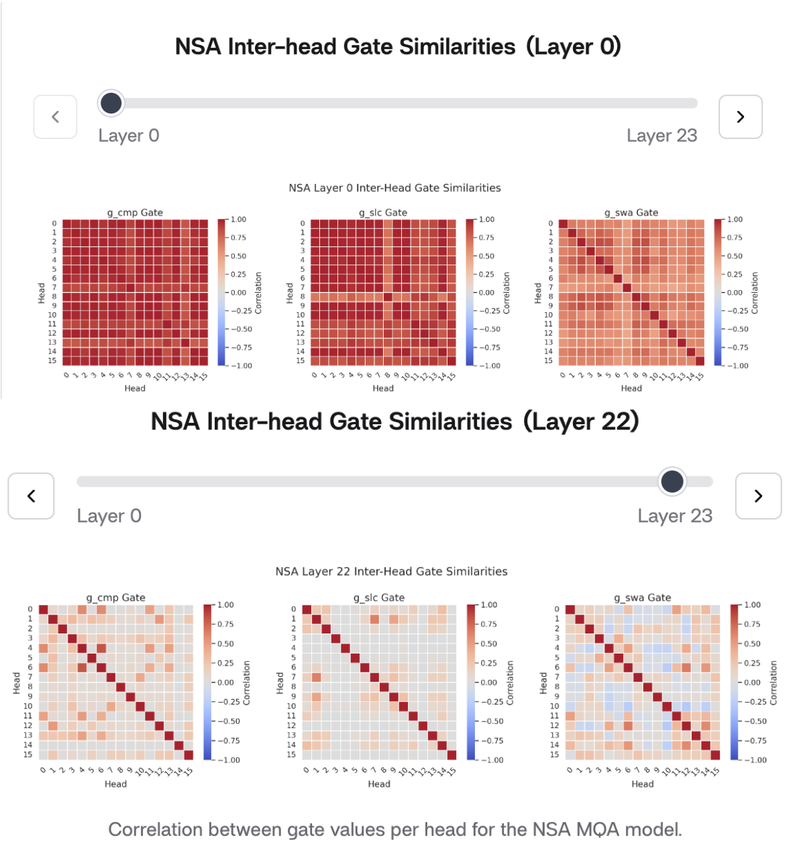


Distributions of gate values for NSA model.

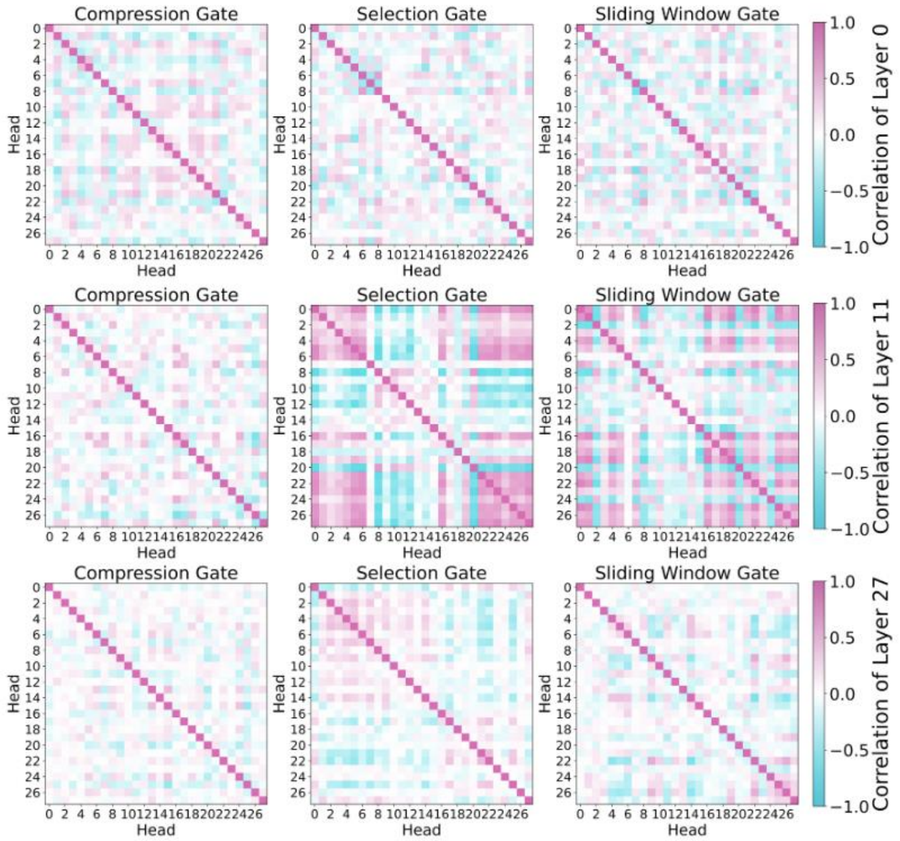


VideoNSA

What roles do compression, selection, and sliding-window gates play in VideoNSA?

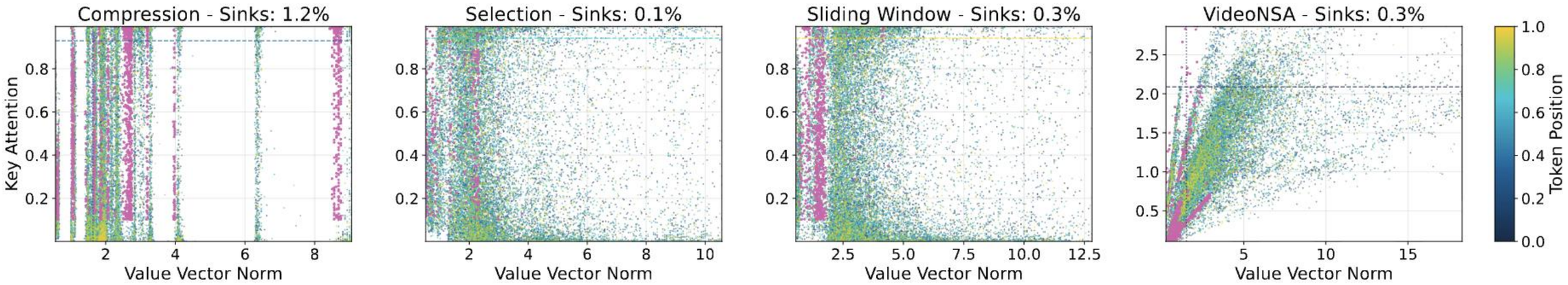


Correlation between gate values per head for the NSA MQA model.



VideoNSA

Do learnable sparse mechanisms induce dynamic attention sinks?



Research Interest

Long-context Encoding

1. How an AI system can perceive and understand extremely long multimodal contexts, such as an entire day of human activity or the full history of a project.
2. Currently we do compression, selection and sliding window. What is Next?
3. How we process the conflict between training time and test time?

Long-context Decoding

1. How an AI system should evolve its internal state over time to generate coherent long-horizon multimodal trajectories？
2. Why do models drift or contradict themselves even with sufficient context?
3. How to generate effective long-context multimodal outputs？

Reliable Evaluation

1. How can we design reliable and persistent evaluation metrics for vision and multimodal models, analogous to perplexity (PPL) in NLP?
2. How can we build dynamic benchmarks that evolve over time to prevent memorization and overfitting to static test sets?

Graduate Plan

1.Research Direction

- Try research on **long-context generation** during the PhD.

2.Research Output

- Aim to submit 1–2 papers per year during the PhD.
- Start writing blogs to organize research thoughts and findings.

3.Open-source Project

- Build a large open-source project, such as a training or evaluation framework, to support long-context research and reproducible experiments.

4.Internships

- leverage industry resources to work on real and meaningful problems, and use these experiences to clarify what I want to work on next

Questions

1. After my first year in the PhD program, when you look back, what concrete outcomes or progress would make you feel that admitting me was the right decision?
2. In your view, what kinds of research questions or work can only or best be done in academia? And how do you see the current trend of PhD students leaving academia for industry roles?
3. Are there kinds of work that you generally discourage students from doing? Is that more about problem choice, methodology, or how the work is framed?
4. Computing Recourse
5. Funding
6. Graduation requirements / Advise Style

THANK YOU

02 May, 2024