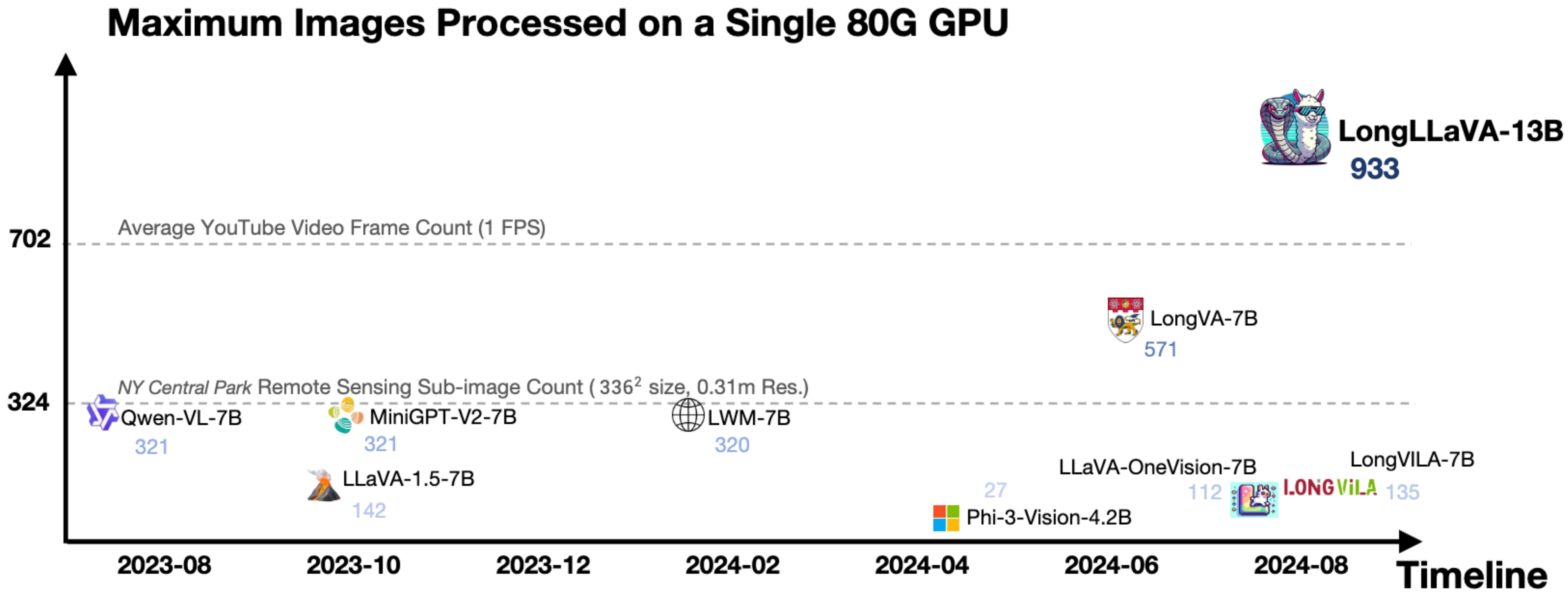
looongLLaVA

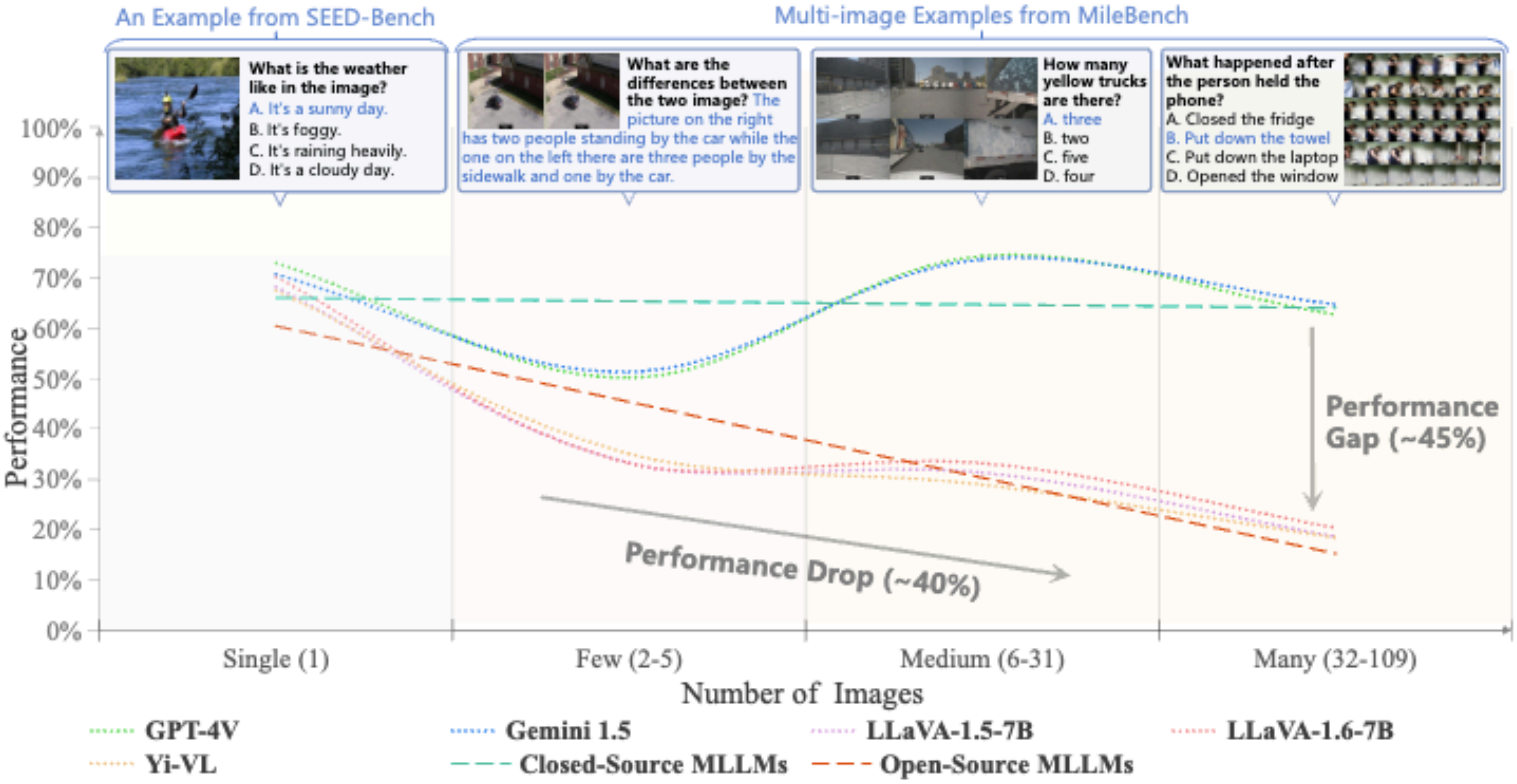
Scaling Multi-modal LLMs to 1000 Images Efficiently via Hybrid Architecture

Kumar Krishna Agrawal , 09/23/2024

Toward multi-image MLMs



Challenges



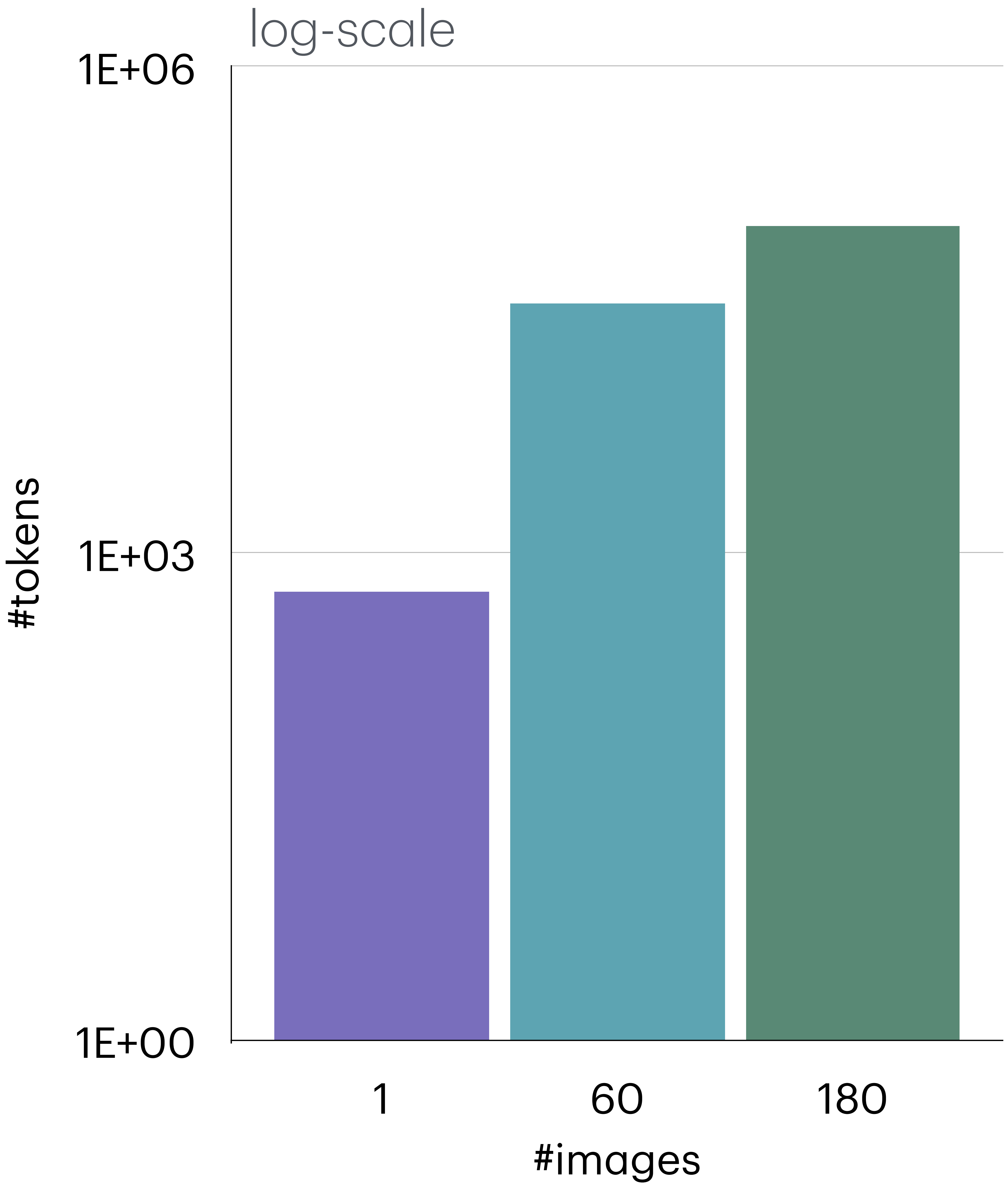
An Example from SEED-Bench
Multi-image Examples from MileBench
What is the weather like in the image?
A. It's a sunny day.
B. It's foggy.
C. It's raining heavily.
D. It's a cloudy day.
What are the differences between the two image? The picture on the right has two people standing by the car while the one on the left there are three people by the sidewalk and one by the car.
How many yellow trucks are there?
A. three
B. two
C. five
D. four
What happened after the person held the phone?
A. Closed the fridge
B. Put down the towel
C. Put down the laptop
D. Opened the window
Performance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Performance Gap (~45%)
Performance Drop (~40%)
Single (1)
Few (2-5)
Medium (6-31)
Many (32-109)
Number of Images
GPT-4V
Gemini 1.5
LLaVA-1.5-7B
LLaVA-1.6-7B
Yi-VL
Closed-Source MLLMs
Open-Source MLLMs

Challenges

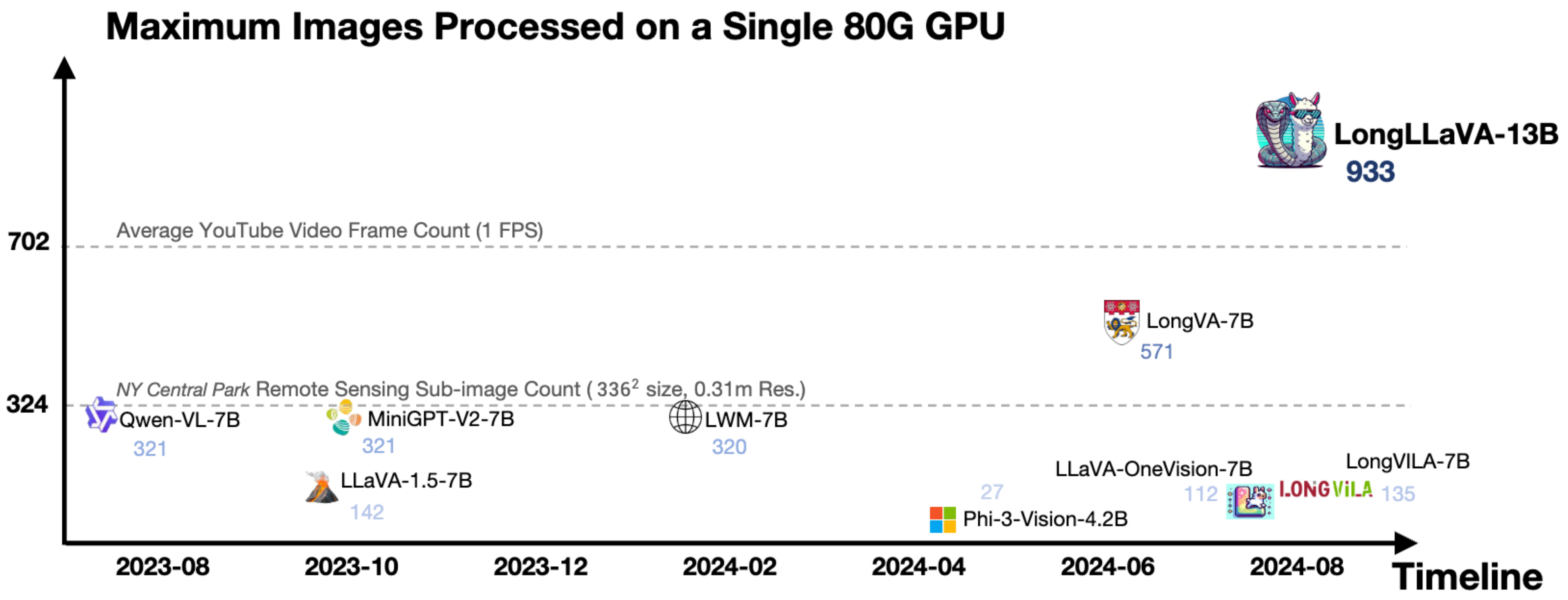
Excessive Input Length

- number of tokens grows linearly with number of images

- computational/memory complexity, particularly for attn computation grows quadratically

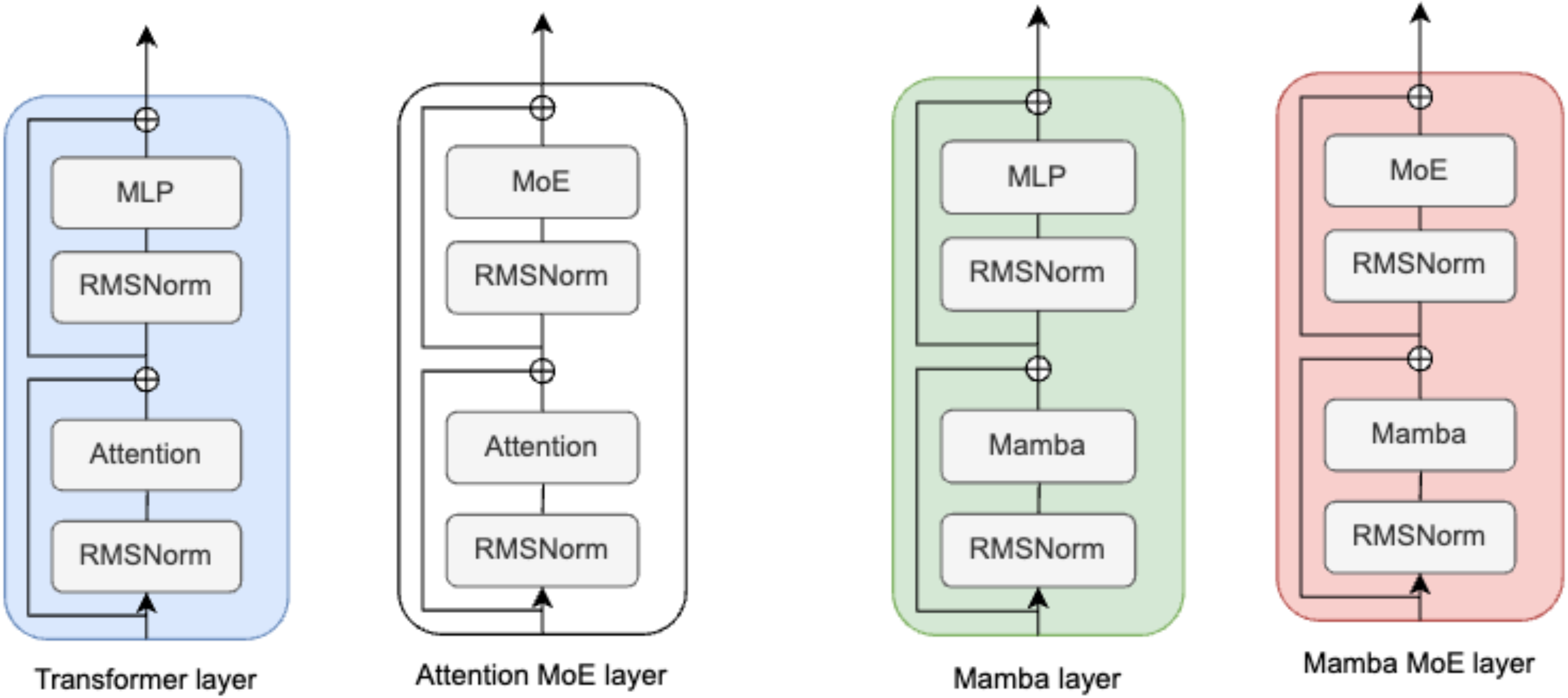


Challenges



Computational cost for inference increases (e.g. storing KV-Cache) grows

Architecture



Architecture

Architecture	Compute Complexity	ICL	Representative models
Transformer	Quadratic	✓	Gemma (Team et al., 2024a), LLaMA (Touvron et al., 2023a)
Mamba	Linear	✗	Mamba (Gu & Dao, 2024), Mamba-2 (Dao & Gu, 2024)
Hybrid	Quasi-Linear	✓	Jamba (Lieber et al., 2024), Zamba (Glorioso et al., 2024)

Architecture

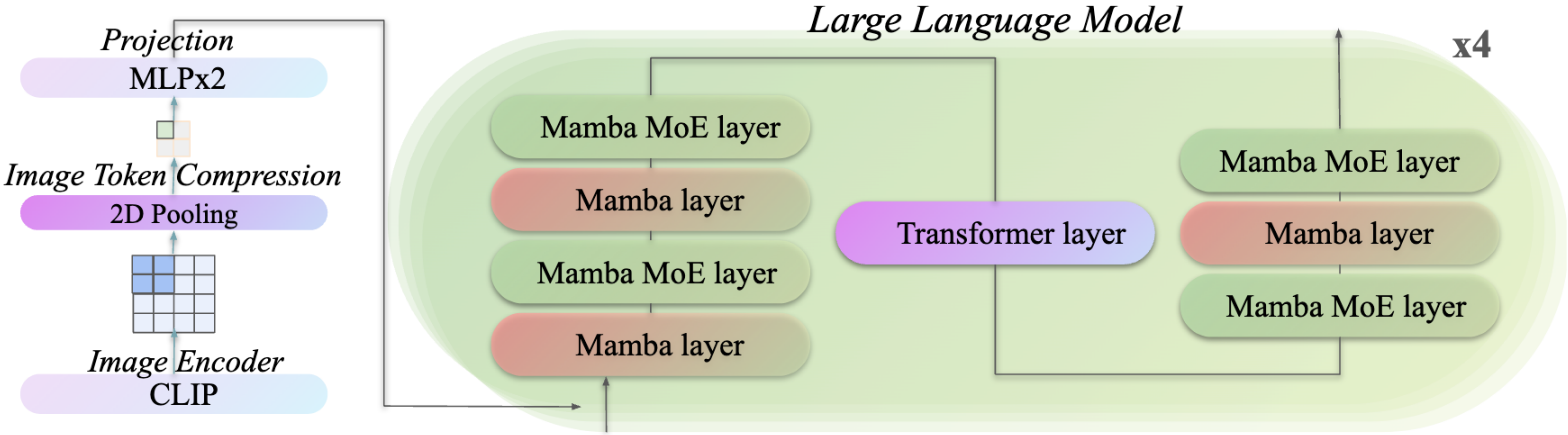


Figure 2: Architecture of LongLLaVA

Data Protocols

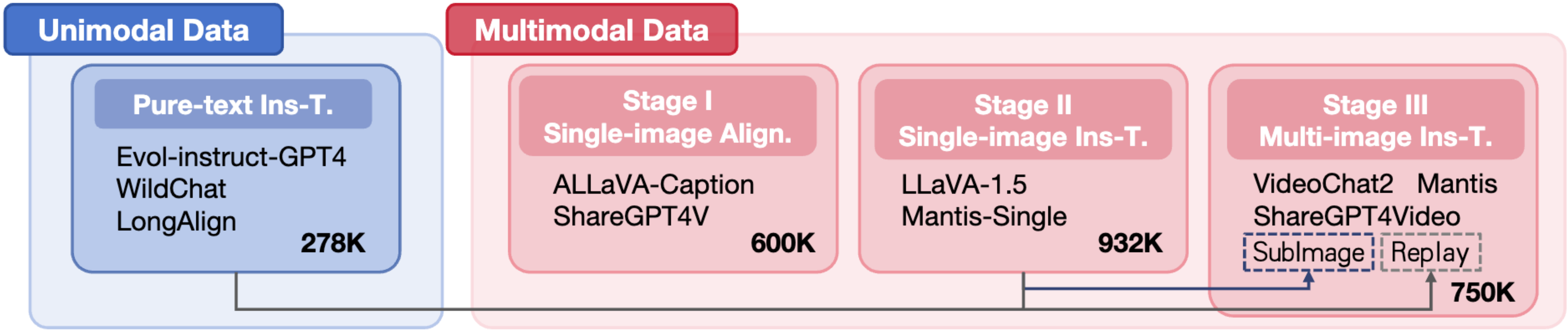
Data Processing Protocol

In the Following Statement: `<Image>=<img><img_token>...</img>`
For Single-image: “`<Image>\n` What is this?”
For Multi-image: “`<Image>\n` This is a cat. `<Image>\n`This is a:”
For Video: “`<vid><Image><t>...<Image></vid>\n` What are they?”
For Patched-image: “`<Image>\n<Image>..\n..<Image>\n` What are they?”

Figure 3: Data Processing Protocol for LongLLaVA.

Regular Single and Multiple Images: use <img> </img> tokens
Video: use <vid> </vid> to enclose image tokens, <t> to separate tokens
High Resolution Image: for multiple patches, use <\n> to indicate same image patching

Data Protocols



Benchmarks

Model	PFLOPs	MileBench				VideoMME w/o subs				MVBench
		Temporal	Semantic	IR	Avg.	Short	Medium	Long	Avg.	
Proprietary Models										
GPT-4V	-	45.6	58.9	86.7	63.7	70.5	55.8	53.5	59.9	**43.5**
GPT-4o	-	**56.2**	**63.5**	**88.8**	**69.5**	72.5	63.1	58.6	64.7	-
Gemini-1.5-Pro	-	50.2	58.3	88.0	65.5	**78.8**	**68.8**	**61.1**	**69.6**	-
Claude3-Opus	-	37.4	48.1	25.0	36.8	70.5	57.4	51.2	59.7	-
Open-source MLLMs										
Video-LLaMA2	3.71	-	-	-	-	55.9	45.4	42.1	47.8	34.1
VideoChat2	0.24	25.5	25.5	9.2	20.1	48.3	37.0	33.2	39.5	51.9
LongVILA	3.90	-	-	-	-	**61.8**	**49.7**	39.7	50.5	-
Phi-3-Vision	2.68	46.9	50.0	18.7	38.5	-	-	-	-	-
OmChat	3.90	51.4	52.0	34.2	45.9	-	-	-	-	50.2
LongLLaVA*	0.22	**52.7**	**52.1**	**67.5**	**57.4**	60.9	**49.7**	**44.1**	**51.6**	**54.6**

Diagnostic Evaluation

Video MLLM	PFLOPs	Retrieval E	Retrieval I-1	Retrieval I-2	Ordering E	Ordering I-1	Ordering I-2	Counting E-1	Counting E-2	Counting I	Average
Proprietary Models											
Gemini-1.5	-	100.0	96.0	76.0	90.7	95.3	32.7	60.7	7.3	42.0	66.7
GPT-4o	-	100.0	98.0	87.3	88.4	86.6	45.2	36.8	0.0	36.1	64.4
GPT-4V	-	100.0	99.3	82.0	42.6	22.8	23.0	37.6	0.0	32.4	48.9
Open-source MLLMs											
Video-LLama2	0.85	1.2	26.0	6.0	0.0	0.0	0.0	2.0	4.7	0.7	4.5
VideoChat2	0.08	43.4	40.0	14.6	0.0	0.0	1.3	4.4	8.0	12.4	12.4
LongLLaVA*	0.09	**100**	**73.3**	**100.0**	**37.5**	**35.3**	**34.8**	**36.0**	**23.7**	**28.0**	**52.1**

Ablations

Method	#Token	GQA	MMMU	SQA^I	$SEED^{v1}_{img}$	$Mile^*_{avg}$
LLaVA-1.5-13B	576	63.3	34.4	71.6	68.2	27.6
+Jamba	576	63.2	41.4	75.4	69.8	38.2
+*1D Pooling*	144	60.4	42.0	73.9	66.3	36.2
+2D Pooling	144	61.3	42.1	75.2	67.4	37.7
+Single-image Data	144	62.2	42.1	75.9	68.9	50.0
+Multi-image Data	144	59.9	39.2	73.4	65.3	57.4

In Context Learning

Model	Arch.	Active Param.	#Few-shot of VL-ICL 1	2	4	5	100K Token (Efficiency) Prefill (s)	TP (tokens/s)	Mem.(GB)
Cobra	Mamba	3B	48.7	50.3	51.0	51.5	10.2	42.7	29.9
LLaVA-1.6*	Transformer	13B	50.0	52.3	54.6	58.9	34.0	14.7	79.4
LongLLaVA*	Hybrid	13B	52.3	59.0	59.0	61.3	25.5	37.6	79.1

Table 5: ICL Capability and Efficiency Analysis across different Architectures. * means the Model is evaluated using `Int8` precision. We select Cobra 3B since it is the largest Mamba-based LLM to date.

Scaling with more images

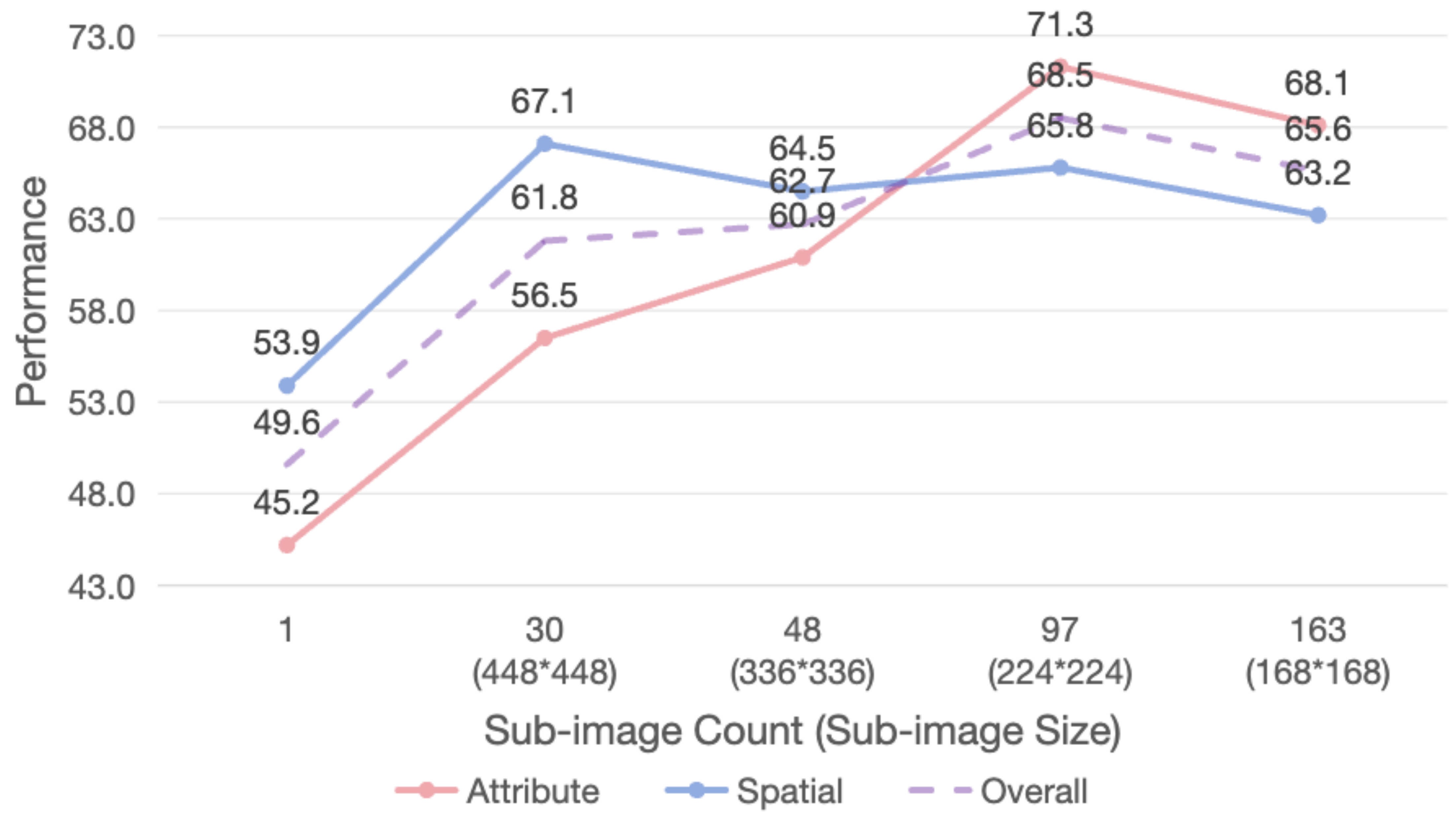


Figure 5: Performance of LongLLaVA with increasing sub-image counts on V*

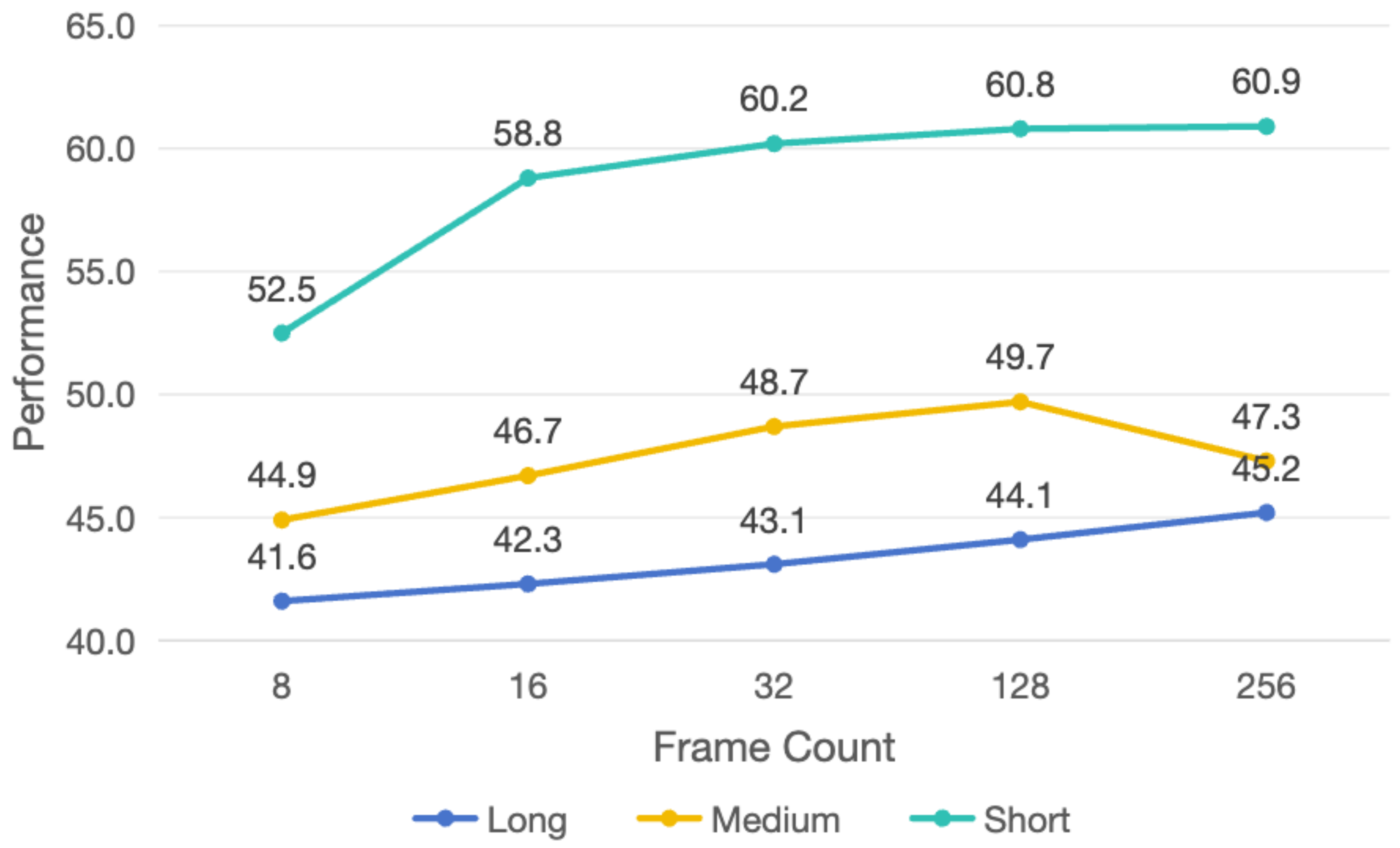
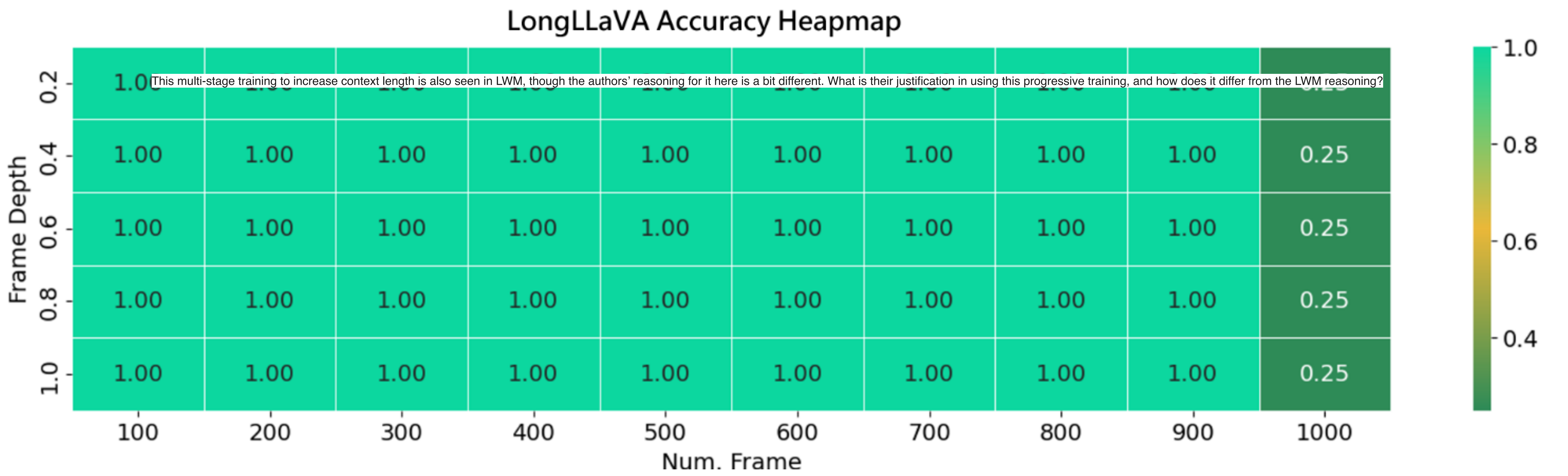


Figure 6: Performance of LongLLaVA with increasing frame counts on Video-MME

Needle in Haystack Evals



This multi-stage training to increase context length is also seen in LWM, though the authors' reasoning for it here is a bit different. What is their justification in using this progressive training, and how does it differ from the LWM reasoning?

Discussion

>> This multi-stage training to increase context length is also seen in LWM, though the authors’ reasoning for it here is a bit different. What is their justification in using this progressive training, and how does it differ from the LWM reasoning?

>> How should one figure out which architecture to go with? We've seen Mamba and the KAN architecture as well but Im curious how people decide what to go with (is it just what everyone is talking about at the time?) Also we just saw a couple weeks ago that SigLIP is better. Why use clip for the encoder then??

Also, many of these papers show multi stage training pipelines. In the future, could one envision dumping all the data into a single folder and then having another mechanism that suggests/picks which samples to train on and when?