AIM: Adaptive Inference of Multi-Modal LLMs via Token Merging and Pruning

Yiwu Zhong[1], Zhuoming Liu[2], Yin Li[2], Liwei Wang[1*]

[1]The Chinese University of Hong Kong, [2]University of Wisconsin-Madison (*Corresponding author)

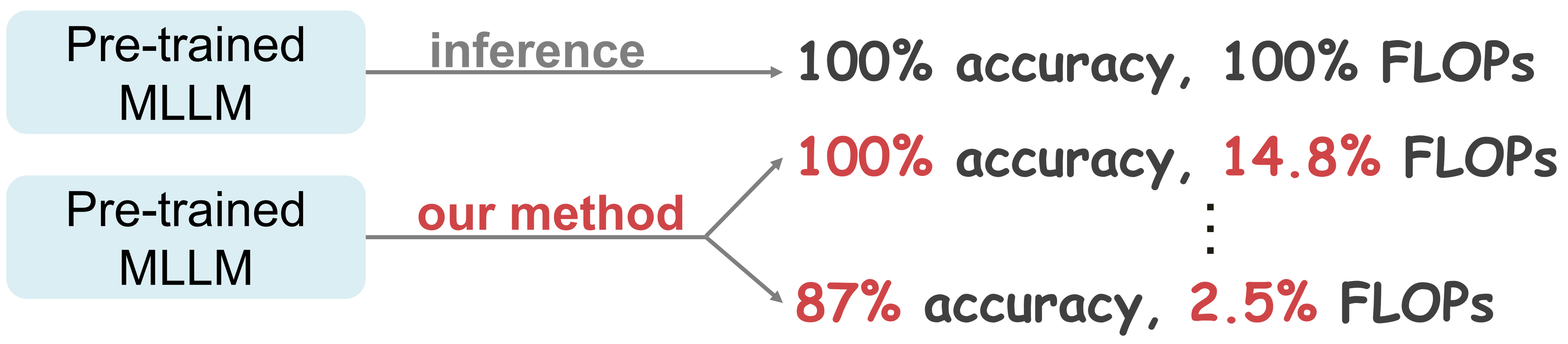
香港中文大學 The Chinese University of Hong Kong

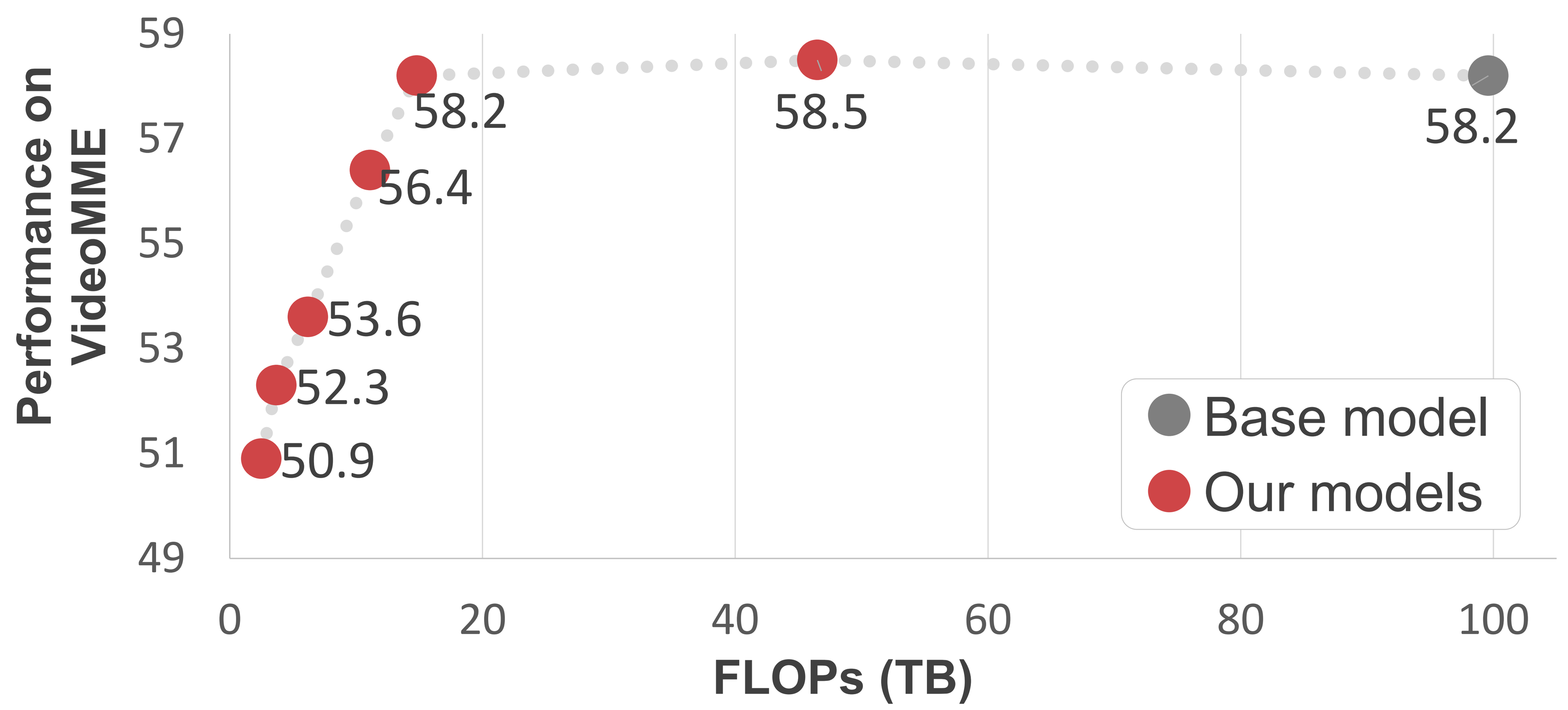
WISCONSIN UNIVERSITY OF WISCONSIN-MADISON

ICCV HONOLULU HAWAII OCT 19-23, 2025

Motivation

Can we speed up MLLM without accuracy loss?

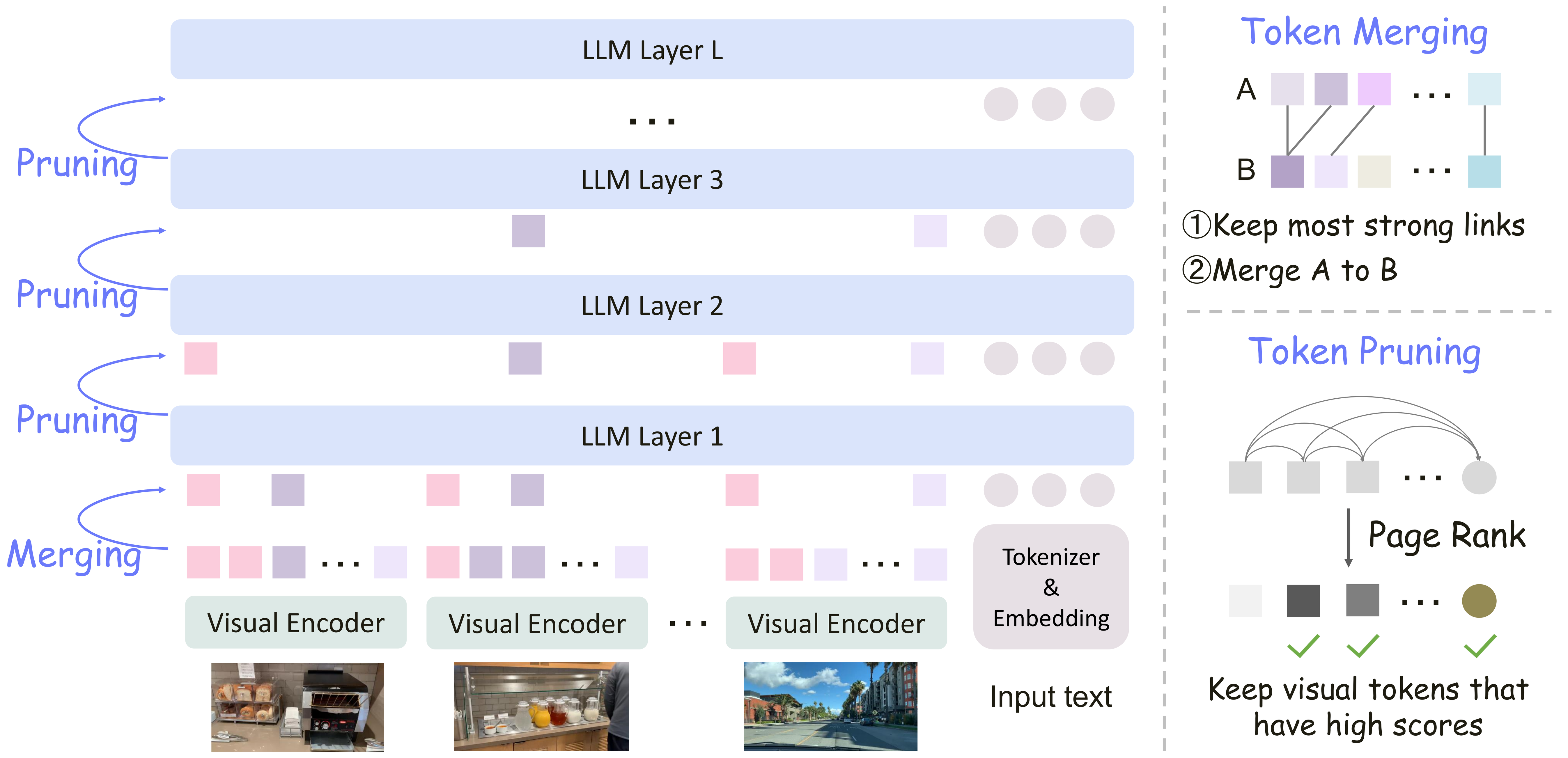




- **Goal**: developing adaptive inference for multi-modal LLMs.
- **Key challenge**: how to reduce redundancy & identify necessary visual tokens that contribute to "accuracy"?
- **Key idea**: introducing a training-free method based on visual token similarity and multi-modal token importance.
- **Key results: 7x speed up & +4.6 accuracy for long videos**

Project Page: https://github.com/LaVi-Lab/AIM

Method



Findings

1. AIM achieves a broad range of accuracy-efficiency trade-offs (**adaptive inference**).
2. AIM's overhead is **negligible**.
3. Visual tokens matter at **early layers** & text tokens are focused at **later layers**.
4. Pruning text tokens (at any layer) **hurts** performance.

Retention Ratio	l_1	l_2	FLOPs (TB)	Prefill Time (ms)	VideoMME wo-subs
100.0%	-	-	99.63	439.58	58.2
50.0%	-	-	46.48	182.65	58.5
25.0%	**14**	**22**	**14.76**	**55.03**	**58.2**
12.5%	14	22	11.14	39.41	56.4
6.3%	14	22	6.17	21.69	53.6
3.1%	14	22	3.72	13.26	52.3
1.6%	14	22	2.51	10.12	50.9

AIM Adaptive Inference

FLOPs (GB)	Video LLM (Qwen2-7B)
Token Merging	88.25
Token Pruning	4.18
Total	92.43
LLM Inference	14757

AIM Overhead

Retention Ratio	FLOPs (TB)	Prefill Time (ms)	VideoMME wo-subs
100.0%	99.63	439.58	58.2
50.0%	46.48	182.65	58.5
25.0%	**22.90**	**83.94**	**58.0**
12.5%	11.64	41.22	56.6
6.3%	6.41	22.54	53.6
3.1%	3.85	13.68	52.3
1.6%	2.57	10.15	50.9

Token Merging

Exp.	l_1	l_2	FLOPs (TB)	Prefill Time (ms)	VideoMME wo-subs
1	28	29	22.90	83.94	58.0
2	21	29	20.15	73.61	58.0
3	14	29	17.41	63.34	57.7
4	7	29	14.66	53.08	57.4
5	21	22	17.50	65.35	58.1
6	**14**	**22**	**14.76**	**55.03**	**58.2**
7	7	22	12.01	44.75	56.8
8	14	15	12.10	46.77	54.3
9	7	15	9.36	36.44	52.9
10	7	8	6.71	28.18	41.9

Token Pruning

Video Benchmarks

Model	FLOPs (TB)	Prefill Time (ms)	VideoMME wo / w-subs	MVBench test	MLVU m-avg	EgoSchema test	NextQA mc	PerceptionTest val
Video LLMs								
LongVA-7B [92]	381.09	2186.04	52.6 / 54.3	-	56.3	-	68.3	-
LLaVA-OV-7B [33]	99.63	439.58	58.2 / 61.5	56.7	64.7	60.1	79.4	57.1
Training-free Method Applied during Inference								
VTW [40]	22.38	101.93	41.0 / 50.0	44.3	39.6	38.0	52.1	41.3
PDrop [79]	24.22	104.88	51.7 / 56.6	52.3	55.6	51.8	74.2	52.8
FastV [5]	21.24	79.56	55.9 / 60.0	55.9	61.1	57.5	77.5	**56.3**
LLaVA-Prumerge [62]	23.65	86.89	57.0 / 59.9	56.5	60.6	**61.0**	77.6	55.8
Ours	**14.76**	**55.03**	**58.2 / 61.3**	**57.1**	**63.7**	59.6	**78.4**	56.0

General Video Understanding Benchmarks

Model	Number of Frames	FLOPs (TB)	Prefill Time (ms)	VideoMME wo / w-subs	MLVU m-avg	EgoSchema test
Video LLMs						
LLaVA-OV-7B [33]	32	99.63	439.58	58.2 / 61.5	64.7	60.1
Training-free Method Applied during Inference						
Ours	32	**14.76**	**55.03**	58.2 / 61.3	63.7	59.6
Ours	192	99.27	471.20	**59.2 / 62.3**	**69.3**	**60.8**

Long Video Understanding Benchmarks

AIM **significantly reduces** FLOPs / prefill time with small accuracy loss, and even **improves on long video benchmarks**.

Image Benchmarks

Model	FLOPs (TB)	Prefill Time (ms)	VQA-v2 (107,394)	GQA (12,578)	MME (2,374)	TextVQA (5,000)	SQA-IMG (2,017)	MMB (4,377)	POPE (8,910)
Image LLMs									
Qwen-VL-Chat-7B [1]	6.44	22.51	78.2	57.5	1487.5	61.5	68.2	60.6	-
LLaVA-1.5-7B [41]	8.18	29.30	78.5	62.0	1510.7	58.2	66.8	73.7	85.9
Training-free Method Applied during Inference									
VTW [40]	2.43	13.88	49.4	42.5	916.4	45.7	66.1	63.1	17.9
PDrop [79]	2.36	13.31	58.1	47.3	999.0	50.4	**68.7**	63.5	46.6
FastV [5]	2.58	10.34	74.1	56.6	1438.5	**57.3**	68.0	72.1	73.6
LLaVA-Prumerge+ [62]	2.41	9.73	74.6	57.4	1391.9	55.2	67.9	71.6	82.2
Ours	2.22	10.92	**75.4**	**58.6**	**1443.5**	53.8	68.4	**72.5**	**85.7**
VTW [40]	1.24	10.66	42.3	38.9	683.7	43.0	65.6	36.5	25.2
FastV [5]	1.12	9.56	55.4	45.5	960.4	51.3	66.0	61.5	33.4
LLaVA-Prumerge [62]	1.04	8.99	66.7	51.3	1242.5	**53.8**	**68.0**	67.1	76.2
Ours	1.00	8.98	**69.0**	**54.6**	**1277.7**	48.4	67.1	**69.4**	**79.5**

General Image Understanding Benchmarks: With **less computation** cost, our method **outperforms baselines** on most benchmarks.