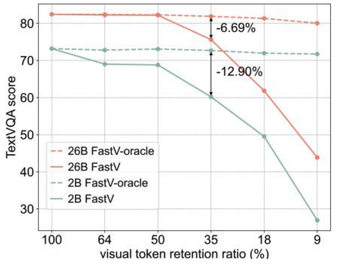
A Stitch in Time Saves Nine:
Small VLM is a Precise Guidance for Accelerating Large VLMs

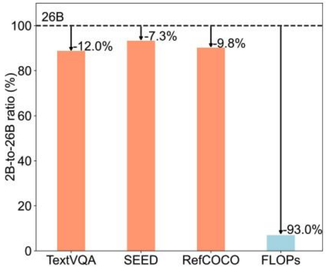
Wangbo Zhao[1*] Yizeng Han[2*] Jiasheng Tang[2,3] Zhikai Li[1]
Yibing Song[2,3] Kai Wang[1†] Zhangyang Wang[4] Yang You[1]
[1]National University of Singapore [2]DAMO Academy, Alibaba Group
[3]Hupan Lab [4]The University of Texas at Austin

CVPR Nashville JUNE 11-15, 2025

Motivation

- A single-layer attention map is suboptimal compared to the global attention maps aggregated from all layers
- Global attention information, such as the attention map aggregated across all layers, more effectively preserves essential tokens
- The global attention map aggregated from a small VLM closely resembles that of a large VLM, suggesting an efficient alternative.
- The performance gap between small and large VLM is minimal compared to their computation disparity.



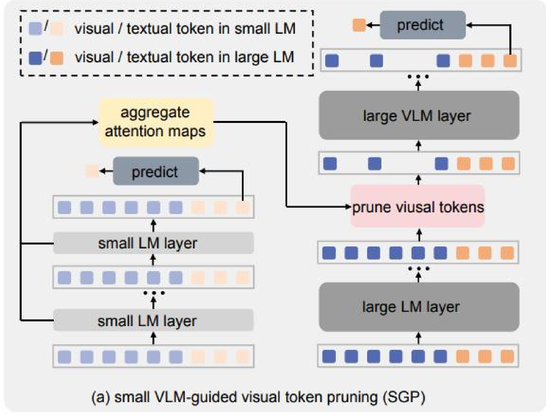


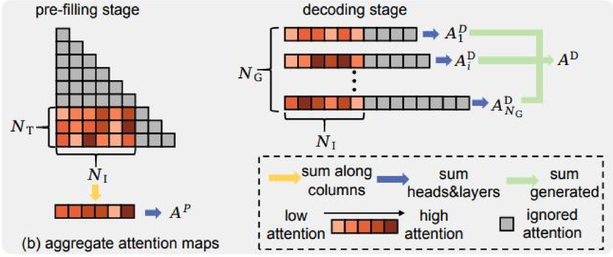
Methodlogy

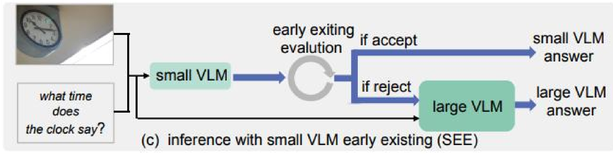
(a) Small VLM-guided visual token pruning in a large VLM (SGP)

.(b) Aggregation of attention maps in SGP

(c) Inference with Small VLM Early Exiting (SEE)







Experiment

method	token ratio	visual question answering				comprehensive benchmark				visual grounding			score ratio
		TextVQA	ChartQA	DocVQA	GQA	SEED	MMBench	MM-Vet	MME	RC	RC+	RC-g	
InternVL2-26B [7]	100%	82.45	84.92	92.14	64.89	76.78	83.46	64.00	2270	91.24	86.67	88.44	100.00%
InternVL2-2B [7]	100%	73.19	76.24	85.93	61.16	71.62	72.93	43.30	1878	82.26	73.53	77.55	87.25%
26B w/ ToMe [2]	64%	80.22	76.24	79.51	64.49	75.60	82.74	60.10	2235	84.02	78.91	80.35	94.24%
	35%	75.74	62.44	66.79	63.61	73.84	81.28	52.50	2178	71.08	64.97	68.08	85.20%
	9%	51.69	28.60	28.46	57.52	65.19	73.09	37.70	1933	20.33	17.74	19.36	54.28%
26B w/ FastV [5]	64%	82.26	85.08	92.20	64.80	76.81	83.24	63.20	2270	91.30	86.66	88.30	99.84%
	35%	75.62	71.68	68.32	61.20	71.64	78.31	45.00	2140	85.06	77.61	81.39	88.28%
	9%	43.84	26.20	26.81	44.90	54.56	62.33	31.60	1799	19.65	16.66	17.22	46.99%
26B w/ SGP (ours)	64%	82.41	85.04	92.12	65.07	76.71	83.30	65.60	2259	91.07	86.71	88.05	100.14%
	35%	81.97	81.68	91.14	64.62	75.72	82.17	63.20	2258	89.38	84.35	86.07	98.36%
	9%	78.98	72.96	87.26	62.10	72.23	75.56	52.10	2004	80.36	72.22	77.45	89.58%