Accelerating Multimodal Large Language Models via Dynamic Visual-Token Exit and Empirical Findings

Qiong Wu[1,2], Wenhao Lin[1,2], Yiyi Zhou[1,2*], Weihao Ye[1], Zhanpeng Zeng[1], Xiaoshuai Sun[1,2], Rongrong Ji[1,2]

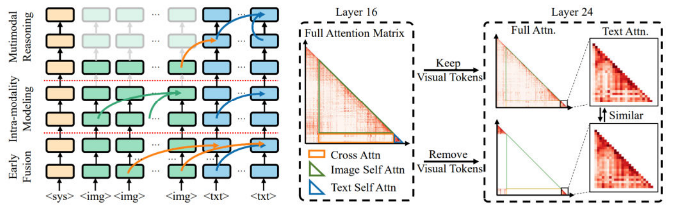
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University, China.

[2]Institute of Artificial Intelligence, Xiamen University, China

INTRODUCTION

Multimodal Large Language Models (MLLMs) integrate visual and textual understanding, yet their inference is computationally expensive due to redundant visual tokens. Existing efficiency methods often prune tokens heuristically, lacking in-depth exploration of the intrinsic behaviors of MLLMs. We propose Dynamic Visual Token Exiting (DyVTE), which adaptively removes redundant visual tokens early, improving efficiency with minimal accuracy loss.

MOTIVATION



Left: Illustration of three main stages observed in MLLMs.

Right: The impact of visual tokens on the text self attention at the multimodal reasoning stage.

CONTRIBUTION

1. We study the problem of visual redundancy from the perspective of MLLMs' behaviors, and reveal the dependency between text and visual tokens.

2. Based on the empirical findings, we propose a novel and effective approach to reduce visual redundancy of MLLMs, termed dynamic visual-token exit (DyVTE), which can dynamically evaluate and schedule the contributions of visual tokens to multimodal reasoning.

3. The extensive experiments on a set of MLLMs well validate the motivation and effectiveness of DyVTE, also providing insights into the principle of MLLMs

Pattern Analysis of MLLMs

Fig 2. The averaged attention scores of four MLLMs in terms of cross, visual and text attentions.

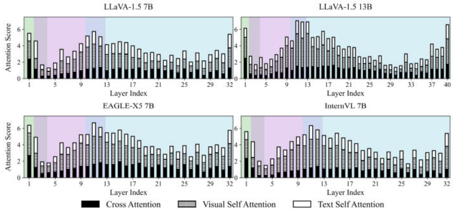


Fig 3. The relationship between manual visual-token exit and performance on LLaVA-1.5 7B and 13B.

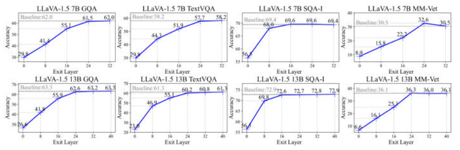
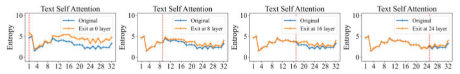


Fig 4. The entropy of text self-attention distribution with different layer to remove all visual tokens.



Dynamic Visual-token Exiting

Method

DyVTE uses lightweight hyper-networks to perceive the text token status of MLLMs and then adaptively judge the right time to remove all visual tokens.

$$\mathbf{p} = Softmax(\mathrm{GELU}([avg(\mathbf{T}^{(k)}_{1:t-1}), \mathbf{T}^{(k)}_{t}]\mathbf{W}_1)\mathbf{W}_2)$$

When the prediction p is exit, DyVTE will remove all visual tokens after this layer, while the text ones are kept in the rest layers of MLLMs, denoted by $P'_l = G_{l+1:L}(\mathbf{T}^{(l)})$

Optimization

The objective of DyVTE can be defined by $\underset{\theta_h}{\operatorname{argmin}}\, d(P, P'_l)$

We compare the discrete outputs of the MLLM with and without DyVTE, e.g., the answer strings A.

$$\mathbf{y} = \begin{cases} 1, & A'_l = A, \\ 0, & A'_l \neq A. \end{cases}$$

Where A'_l denotes the answer predicted with visual-token exit at the l-th layer.

Besides, to make this supervision more robust, we also consider the prediction uncertainty as a regularization term, thereby , making the MLLM behaviors closer to the default inference:

$$\mathbf{y} = \begin{cases} 1, & A'_l = A \wedge \rho'_c < \tau, \\ 0, & otherwise. \end{cases}$$

Here, ρ'_c denotes the prediction uncertainty valued by cross-entropy and multiplied by a scaling factor.

With this supervision, the hyper-networks in DyVTE can be optimized by the cross-entropy loss:

$$\mathcal{L}_D = -(\mathbf{y}\cdot\log(\mathbf{p}_1) + (1-\mathbf{y})\cdot\log(\mathbf{p}_0)).$$

EXPERIMENT

Tab 1. Results of MLLMs with and without DyVTE on five MLLM benchmarks.

Method	SEED Accuracy ↑	SEED TFLOPs ↓	MME Score ↑	MME TFLOPs ↓	MMB Accuracy ↑	MMB TFLOPs ↓	POPE Accuracy ↑	POPE TFLOPs ↓	MM-Vet Accuracy ↑	MM-Vet TFLOPs ↓
EAGLE-X5 7B	73.9	47.8	1528.0	27.8	68.4	29.6	88.8	27.7	37.4	27.6
EAGLE-DyVTE 7B	73.6 (-0.4%)	43.0 (-10.0%)	1581.7 (+3.5%)	20.3 (-27.0%)	68.8 (+0.6%)	23.7 (-19.9%)	88.4 (-0.5%)	20.0 (-27.8%)	37.8 (+1.1%)	23.5 (-14.9%)
VILA 7B	61.7	9.2	1489.2	8.9	69.9	9.5	86.3	8.8	36.3	8.7
VILA-DyVTE 7B	61.8 (+0.2%)	5.9 (-35.9%)	1503.1 (+0.1%)	4.6 (-48.3%)	69.8 (-0.1%)	6.0 (-36.8%)	85.6 (-0.8%)	4.5 (-48.9%)	36.7 (+1.1%)	6.6 (-24.1%)
InternVL 7B	59.2	16.0	1525.1	15.5	64.6	16.2	86.4	15.4	31.2	15.4
InternVL-DyVTE 7B	59.1 (-0.2%)	11.9 (-25.6%)	1474.1 (-3.3%)	10.9 (-29.7%)	64.4 (-0.3%)	12.0 (-25.9%)	81.3 (-5.9%)	10.9 (-29.2%)	29.5 (-5.4%)	13.0 (-15.6%)
LLaVA-1.5 7B	58.6	9.2	1510.7	8.9	64.3	9.6	85.9	8.8	30.5	8.7
LLaVA-DyVTE 7B	58.6 (0.0%)	5.0 (-45.7%)	1491.4 (-1.3%)	4.3 (-51.7%)	64.7 (+0.6%)	5.4 (-43.8%)	81.6 (-5.0%)	4.1 (-53.4%)	31.9 (+4.6%)	6.3 (-27.6%)
LLaVA-1.5 13B	61.6	17.6	1531.3	16.9	67.7	18.3	85.9	16.8	36.1	16.7
LLaVA-DyVTE 13B	59.3 (-3.7%)	7.1 (-59.7%)	1546.4 (+1.0%)	7.2 (-57.4%)	66.0 (-2.5%)	7.8 (-57.4%)	84.8 (-1.3%)	7.6 (-54.8%)	34.8 (-3.6%)	10.6 (-36.5%)

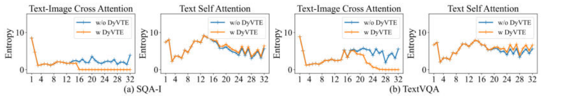
Tab 2. Results of MLLMs with and without DyVTE on four VL benchmarks.

Method	GQA Accuracy ↑	GQA TFLOPs ↓	VQA Score ↑	VQA TFLOPs ↓	TextVQA Accuracy ↑	TextVQA TFLOPs ↓	SQA-I Accuracy ↑	SQA-I TFLOPs ↓	Average Accuracy ↑	Average TFLOPs ↓
EAGLE-X5 7B	64.9	27.8	83.4	27.8	71.2	29.5	69.8	29.2	72.3	28.6
EAGLE-DyVTE 7B	62.4 (-3.9%)	21.7 (-21.9%)	82.6 (-1.0%)	21.6 (-22.3%)	70.2 (-1.4%)	24.5 (-16.8%)	71.7 (+2.7%)	23.5 (-24.3%)	71.7 (-0.8%)	22.8 (-20.3%)
VILA 7B	63.1	8.8	80.3	8.8	62.6	9.5	69.5	9.8	68.8	9.2
VILA-DyVTE 7B	61.9 (-1.9%)	5.5 (-37.5%)	79.2 (-1.4%)	5.4 (-38.6%)	61.2 (-2.2%)	7.2 (-24.2%)	69.5 (0.0%)	6.1 (-37.8%)	67.9 (-1.3%)	6.0 (-34.8%)
InternVL 7B	62.9	15.4	79.3	15.4	57.0	16.1	66.2	16.4	66.4	15.8
InternVL-DyVTE 7B	61.3 (-2.5%)	11.8 (-23.4%)	77.6 (-2.1%)	11.7 (-24.0%)	55.8 (-2.1%)	13.5 (-16.1%)	66.2 (0.0%)	12.1 (-26.2%)	65.2 (-1.8%)	12.3 (-22.2%)
LLaVA-1.5 7B	62.0	8.8	78.5	8.8	58.2	9.5	69.4	9.8	67.0	9.2
LLaVA-DyVTE 7B	60.0 (-3.2%)	5.3 (-39.8%)	76.6 (-2.4%)	5.1 (-42.0%)	56.6 (-2.7%)	6.7 (-29.5%)	69.6 (+0.3%)	5.5 (-43.9%)	65.7 (-1.9%)	5.6 (-39.1%)
LLaVA-1.5 13B	63.3	16.8	80.0	16.8	61.3	18.1	72.9	18.6	69.4	17.6
LLaVA-DyVTE 13B	62.3 (-1.6%)	9.0 (-46.4%)	78.8 (-1.5%)	8.9 (-47.0%)	58.9 (-3.9%)	10.8 (-40.3%)	72.3 (-0.8%)	8.2 (-55.9%)	68.1 (-1.9%)	9.2 (-47.7%)

Tab 4. Ablation study of different token statuses for DyVTE.

State: Mean Visual	State: Last Visual	State: Mean Text	State: Last Text	GQA Acc. ↑	GQA Exit Layer ↓	TextVQA Acc. ↑	TextVQA Exit Layer ↓	MM-Vet Acc. ↑	MM-Vet Exit Layer ↓	SQA-I Acc. ↑	SQA-I Exit Layer ↓	Average Acc. ↑	Average Exit Layer ↓
✓				**61.4**	21.6	57.3	21.0	**31.5**	21.4	69.7	21.7	**55.0**	21.4
	✓			61.2	21.3	57.1	21.1	30.0	**21.0**	69.7	21.4	54.5	21.2
		✓		60.1	17.4	**57.6**	22.1	31.1	22.1	69.7	20.4	54.6	20.5
			✓	59.8	**16.3**	56.3	**19.1**	29.9	**21.0**	**69.8**	**13.8**	54.0	**17.6**
✓	✓			**61.1**	21.2	57.3	21.3	31.6	21.3	**69.7**	21.0	**54.9**	21.2
✓		✓		58.7	16.5	**57.6**	22.2	**32.7**	22.5	**69.7**	21.3	54.7	20.6
✓			✓	59.4	**16.3**	56.5	19.9	30.9	21.8	**69.7**	14.4	54.1	18.1
	✓	✓		59.4	16.7	57.4	21.8	31.3	22.0	69.6	20.3	54.4	20.2
	✓		✓	59.3	16.4	56.5	19.9	31.2	21.6	69.6	14.4	54.1	18.1
		✓	✓	60.0	16.4	56.6	**19.7**	31.9	**21.1**	69.6	**13.4**	54.5	**17.6**

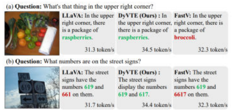
Fig 5. The entropy of two attention distributions.



Tab 3. Comparison of real inference budget between token pruning methods and vanilla decoding.

Method	SQA-I Latency ↓	SQA-I Acc. ↑	MMB Latency ↓	MMB Acc. ↑
LLaVA 13B	0.237s	72.9	0.236s	67.7
FastV	0.171s (-27.7%)	73.1 (+0.3%)	0.174s (-26.2%)	68.6 (+1.3%)
ToMe	0.175s (-26.2%)	73.2 (+0.4%)	0.179s (-24.2%)	67.4 (-0.4%)
DyVTE	0.161s (-32.1%)	72.3 (-0.8%)	0.163s (-30.9%)	66.0 (-2.5%)
DyVTE+FastV	0.154s (-34.9%)	73.0 (+0.2%)	0.155s (-34.4%)	68.5 (+1.2%)

Fig 6. Examples on LLaVA-1.5 with DyVTE.



Tab 5. Comparison between DyVTE and token pruning methods.

Method	SQA-I Accuracy ↑	SQA-I TFLOPs ↓	MM-Vet Score ↑	MM-Vet TFLOPs ↓	SEED Accuracy ↑	SEED TFLOPs ↓	MMB Accuracy ↑	MMB TFLOPs ↓	Average Accuracy ↑	Average TFLOPs ↓
LLaVA 7B	69.4	9.8	30.5	8.7	58.6	9.2	64.3	9.6	55.7	9.3
ToMe [5]	**69.6** (+0.3%)	5.9 (-39.8%)	30.6 (+0.3%)	4.9 (-43.7%)	57.8 (-1.4%)	5.5 (-40.2%)	63.7 (-0.9%)	5.7 (-40.6%)	55.4 (-0.5%)	5.5 (-40.9%)
FastV [7]	69.0 (-0.6%)	6.2 (-36.7%)	31.3 (+2.6%)	5.2 (-35.8%)	58.2 (-0.7%)	5.8 (-37.0%)	64.4 (+0.2%)	6.0 (-37.5%)	55.7 (0.0%)	5.8 (-37.6%)
DyVTE	**69.6** (+0.3%)	5.5 (-43.9%)	**31.9** (+4.6%)	6.3 (-27.6%)	**58.6** (0.0%)	5.0 (-45.7%)	**64.7** (+0.6%)	5.4 (-43.8%)	**56.2** (+0.9%)	5.5 (-40.9%)
DyVTE+FastV	68.9 (-0.7%)	**4.8** (-51.0%)	29.8 (-2.3%)	**4.0** (-54.0%)	58.2 (-0.7%)	**4.6** (-50.0%)	64.4 (+0.2%)	**4.6** (-52.1%)	55.3 (-0.7%)	**4.5** (-51.5%)

Tab 6. Comparison between DyVTE and token pruning methods.

MLLM	Trained	GQA Acc. ↑	GQA Exit Layer ↓	VQAv2 Acc. ↑	VQAv2 Exit Layer ↓	SEED Acc. ↑	SEED Exit Layer ↓	MMB Acc. ↑	MMB Exit Layer ↓	MME Acc. ↑	MME Exit Layer ↓
VILA 7B	-	63.1	-	80.3	-	61.7	-	69.9	-	1489.2	-
	VILA 7B	61.9	17.4	79.2	16.7	61.8	17.6	69.8	16.2	1503.1	13.1
	LLaVA 7B	61.3	17.1	79.1	17.1	61.8	15.8	69.8	16.5	1504.9	14.6
InternVL 7B	-	62.9	-	79.3	-	59.2	-	64.6	-	1525.1	-
	InternVL 7B	61.3	16.1	77.6	15.8	59.1	13.9	64.4	13.6	1474.1	12.1
	LLaVA 7B	61.3	17.0	77.7	16.8	59.2	18.1	64.9	17.8	1516.9	14.1