Dynamic Image Tokenization for Efficient VLMs

THE OHIO STATE UNIVERSITY

Yusen Peng[1], Tejas Naik[1], Valentin Hofmann[2], Yuki M Asano[3], Sachin Kumar[1]
[1]The Ohio State University, [2]Ludwig-Maximilians-Universität München, [3]University of Technology Nuremberg
Corresponding authors: {peng.1007, kumar.1145}@osu.edu

Problem Statement

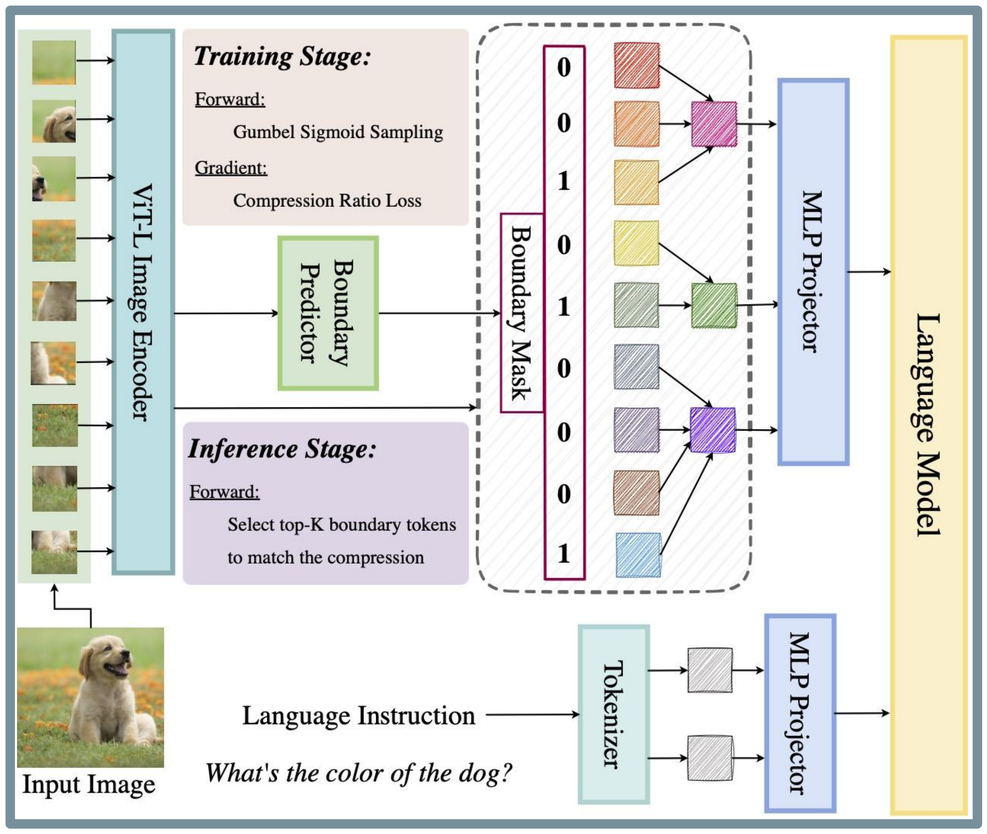
Challenges:

- The large number of image tokens makes Vision Language Models (VLMs) inefficient.
- Prior efforts relied on training-free approaches, which cannot align the LLM with compressed visual representations at inference time.

Research Questions:

- Can simple training-based methods (i.e., fixed tokenization) beat training-free approaches?
- Can dynamic tokenization further improve performance, especially on complex tasks?

Method



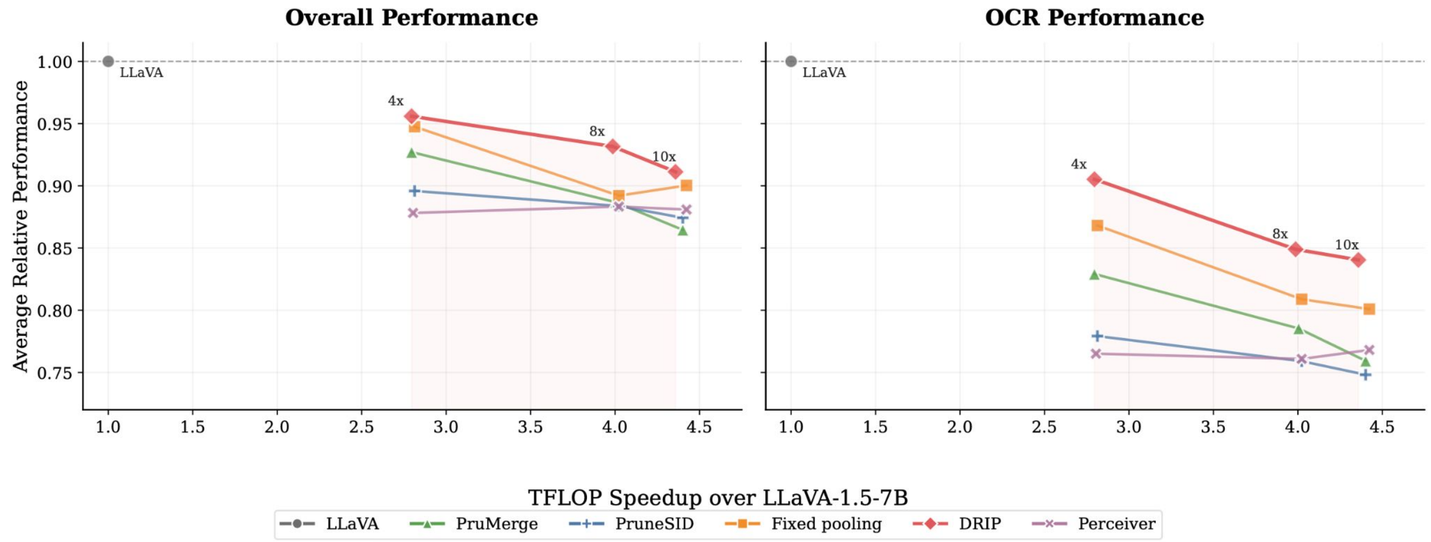
Key Takeaways

On most VQA benchmarks:
training-based > training-free
On complex OCR benchmarks:
dynamic tokenization > fixed tokenization

Results on LLaVA 1.5 7B

Model / Benchmark Type	VQAv2	SQA	MME	MM-Bench	GQA	MMMU	TextVQA	OCRBench	OCRBenchv2	DocVQA	ChartQAPro	POPE	LLaVA-Wild	MM-Vet	TFLOPs
	general				*reasoning*		*OCR*					*hallucination*	*free-response*		
upper bound (keep 576/576 tokens out of the image encoder)															
LLaVA-1.5-7B	76.75	68.42	1425.34	74.43	58.34	34.9	58.16	18.5	23.8	25.61	12.04	85.18	57.4	29.0	8.89
4x compression (keep 144/576 tokens out of image encoder)															
PruMerge(ICCV'25)	74.01	68.07	**1408.83**	73.66	55.04	34.6	54.69	14.9	20.3	16.37	10.93	81.36	61.7	27.0	3.18
PruneSID(ICLR'26)	73.44	67.33	1290.23	72.74	54.86	33.1	53.03	14.5	19.4	13.40	10.38	79.92	59.9	27.6	3.16
Perceiver(ICML'21)	71.05	67.97	1306.41	71.56	52.50	33.2	53.10	13.4	19.1	13.83	10.18	81.50	60.5	23.5	3.17
fixed pooling	74.06	67.82	1335.40	71.72	**55.67**	34.3	55.74	15.7	**20.9**	18.78	11.10	**83.75**	**64.1**	**29.9**	3.16
DRIP	**74.59**	**69.31**	1341.93	**74.05**	55.16	**34.7**	**56.08**	**16.8**	**20.9**	**21.27**	**11.38**	82.83	59.5	28.6	3.18
8x compression (keep 72/576 tokens out of the image encoder)															
PruMerge(ICCV'25)	70.67	68.07	1311.21	72.33	52.01	**34.3**	54.46	14.6	18.4	14.28	10.48	77.53	58.8	25.3	2.22
PruneSID(ICLR'26)	72.37	68.67	1280.62	**73.02**	54.10	33.8	53.14	13.3	18.6	12.78	10.62	79.33	57.6	26.7	2.21
Perceiver(ICML'21)	70.64	**69.21**	**1379.20**	72.52	52.61	33.4	52.78	13.2	19.0	14.24	9.98	79.69	61.0	24.1	2.21
fixed pooling	72.13	69.11	1280.78	71.65	53.78	34.0	54.05	14.3	19.5	16.00	**10.81**	82.15	50.7	25.8	2.21
DRIP	**72.58**	68.47	1310.01	71.90	**54.23**	32.9	**54.75**	**15.7**	**20.4**	**18.50**	10.54	**82.52**	**63.6**	**29.5**	2.23
10x compression (keep 58/576 tokens out of the image encoder)															
PruMerge(ICCV'25)	69.15	68.17	1258.86	71.40	50.90	33.6	54.01	13.7	17.9	13.76	10.09	76.40	52.5	26.7	2.02
PruneSID(ICLR'26)	**71.67**	68.67	1297.33	**73.13**	**53.17**	33.9	52.59	13.1	18.6	12.33	10.42	78.76	55.8	25.8	2.02
Perceiver(ICML'21)	70.70	68.77	1325.25	71.33	52.36	33.6	52.47	13.6	19.2	13.81	10.32	79.63	**59.6**	24.5	2.01
fixed pooling	71.49	68.02	**1359.05**	72.17	53.02	**34.3**	53.40	13.9	19.1	15.32	**11.25**	82.14	57.1	26.1	2.01
DRIP	71.59	**69.61**	1313.76	71.35	52.93	33.2	**54.83**	**15.2**	**20.4**	**17.42**	10.84	**82.59**	55.8	**26.8**	2.04

Table 1: **LLaVA-1.5-7B full-parameter finetuning results using the last layer's image features**. The best results are highlighted among training-free methods and training-based methods.



Qualitative Examples from TextVQA