Shifting AI Efficiency From Model-Centric to Data-Centric Compression

Xuyang Liu

M.S.@SCU, RA@EPIC Lab

Talk@PolyU NLP, Jun. 10, 2025

Xuyang Liu, second-year M.S. at SCU, supervised by Prof. Honggang Chen. RA at EPIC Lab of SJTU, working with Prof. Linfeng Zhang.

- Interned Ant Security Lab / Taobao & Tmall Group.
- Work on **efficient vision-language models** (understanding and generation).
- Selected works: GlobalCom^2 for LVLMs, VidCom^2 for VideoLLMs, and ToCa for DiTs.

Outline of the Talk:

Background of AI Model Efficiency

- Motivation of shifting the efficiency research focus.
- Token overhead across various domains.
- Model efficiency from different perspectives.

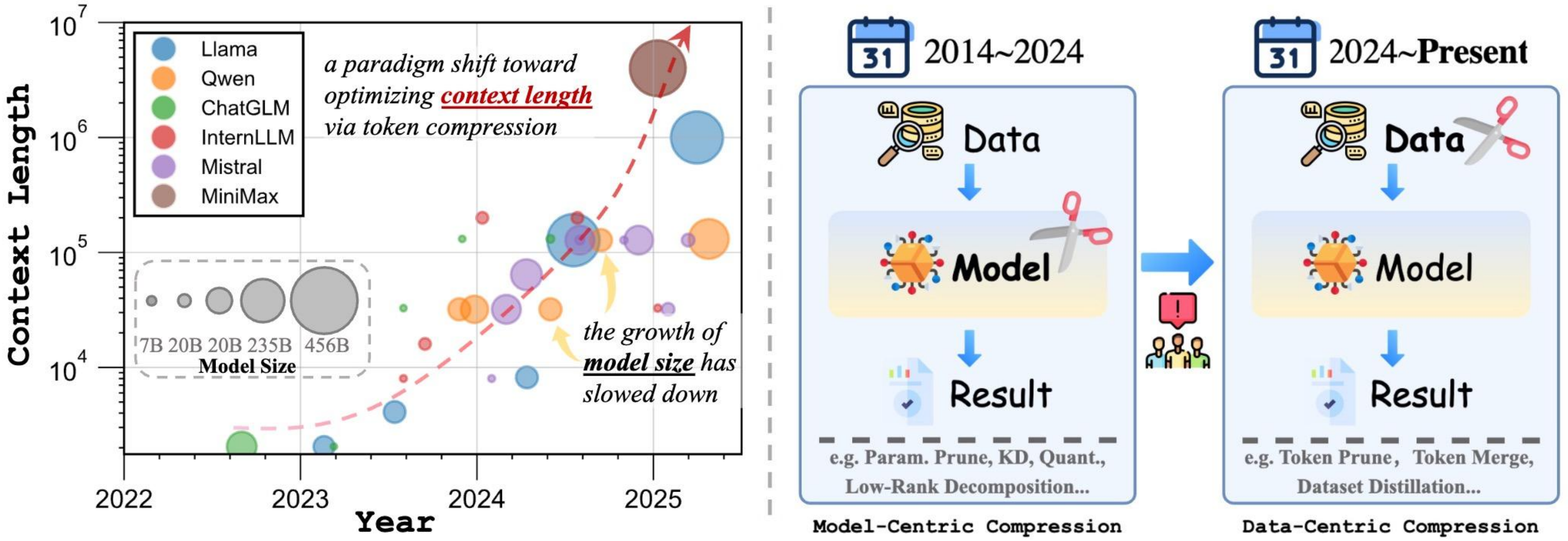
Methodology of Token Compression

- Two stage process of token compression.
- Detailed applications of token compression across domains.
- Five compiling advantages of token compression.

Challenges and Future Works

- Performance degradation by token compression.
- Fair comparison of token compression methods.

Motivation: model capabilities have shifted from scaling model size to extending context length.



Position: the AI community should shift its efficiency optimization paradigm from *model-centric* to ***data-centric*** compression

Background: token overhead across various domains.

Longer Context Length in LLMs

- From 2,048 tokens of Llama to 10M tokens in Llama 4 Scout.
- Long CoT reasoning models (e.g., Qwen3 and Deepseek-R1).
- Multi-agent collaboration systems (e.g., MetaGPT)

Higher Resolution and Longer Video for LVLMs

- From 224*224 of LLaVA to 4K high-resolution of InternVL3.
- Long video understanding (e.g., LongVILA and Video-XL).
- Multi-modal large reasoning models (e.g., QVQ).

More Complex Content for DiTs

- From 512*512 of Stable Diffusion to 4K high-resolution of PixArt-Σ.
- Long video generation (e.g., Sora and HunyuanVideo).

Background: model efficiency from different perspectives.

$$\boldsymbol{F(W;X)}$$

Efficient Architecture

Transformer
Linear Attention
SSM (RWKV, Mamba)
Mixture-of-Experts
Speculative Sampling
DDIM (Diffusion Models)
…

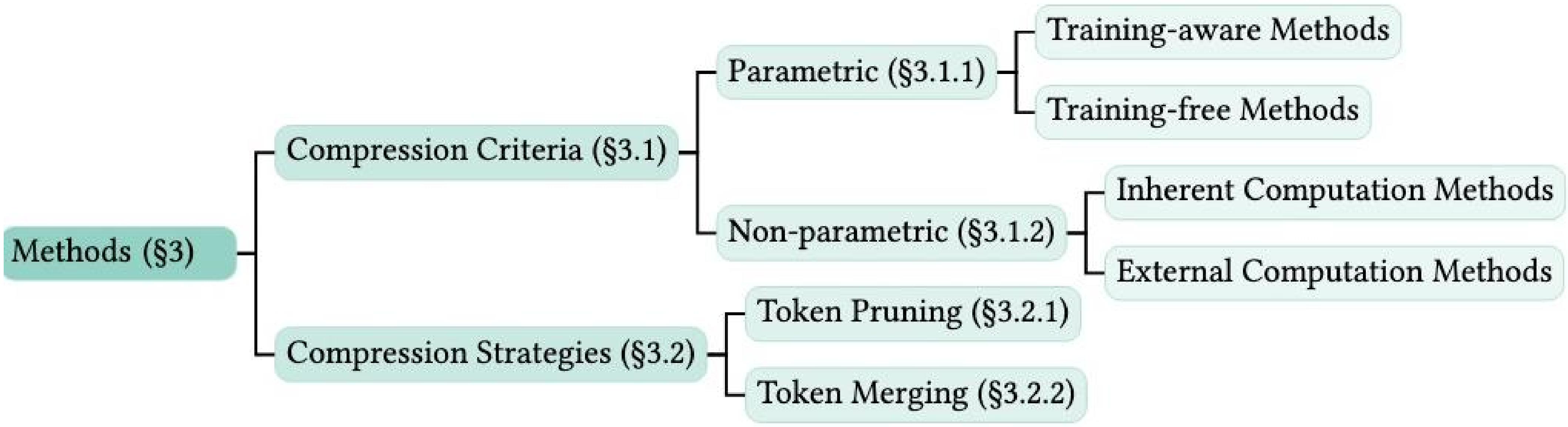
Model-centric Compression

Network Pruning
Model Quantization
Knowledge Distillation
Low Rank Decomposition
…

Data-centric Compression

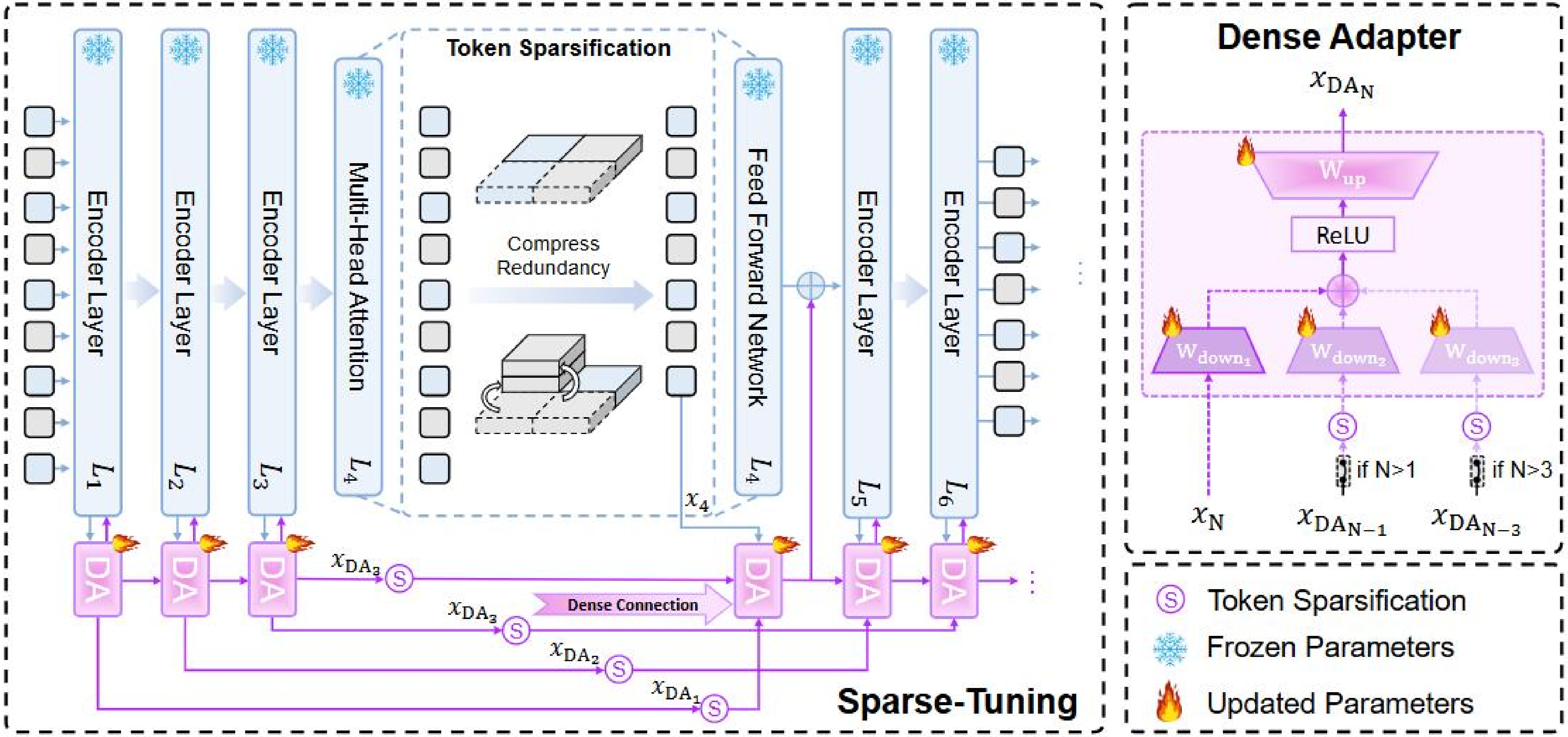
Token Compression
KV Cache Compression
Dataset Distillation
…

Methodology: how token compression works?



Most token compression methods fundamentally operate through **a two-stage process**: first, identifying tokens eligible for compression within the existing token sequence $X = [x_1, x_2, ..., x\ ]$ using carefully designed ***compression criteria*** through a scoring function $\mathcal{E}: X \rightarrow \{s\ \} = 1^{T}$, and then determining the precise handling of these tokens through the specific ***compression strategies*** $\mathcal{P}(X, \{s\ \} = 1^{T}) \rightarrow X'$ that transform the original sequence into a compressed one where $|X'| < |X|$. Given that existing research primarily revolves around these two key components.

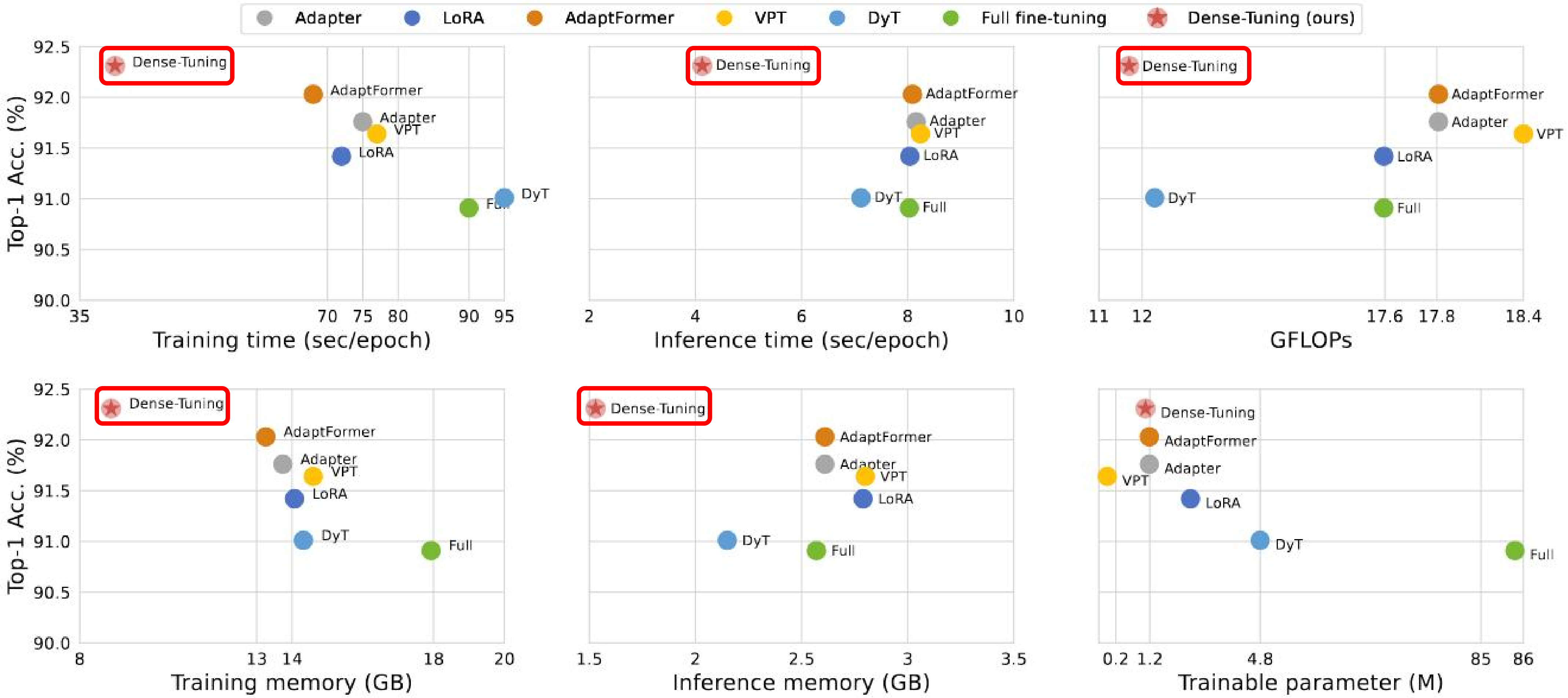
Token Compression in ViTs: Dense-Tuning.



Freezing the pre-trained ViT and update the DAs to efficiently fine-tune thepre-trained ViT.

Ting Liu*, Xuyang Liu*, Siteng Huang, Liangtao Shi, Zunnan Xu , Yi Xin, Quanjun Yin, "Dense-Tuning: Densely Adapting Vision Transformers with Efficient Fine-tuning and Inference". arXiv preprint arXiv:2405.14700.

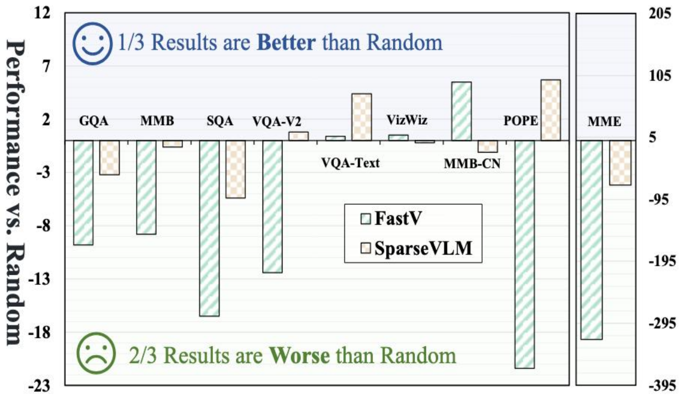
Token Compression in ViTs: Dense-Tuning.

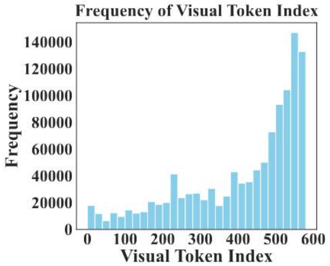


Dense-Tuning effectively adapts the pre-trained ViT with efficient fine-tuning and inference!

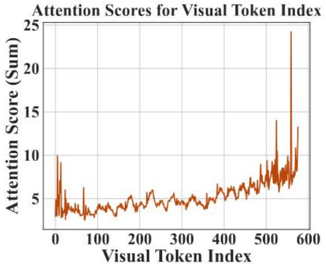
Ting Liu*, Xuyang Liu*, Siteng Huang, Liangtao Shi, Zunnan Xu , Yi Xin, Quanjun Yin, "Dense-Tuning: Densely Adapting Vision Transformers with Efficient Fine-tuning and Inference". arXiv preprint arXiv:2405.14700.

Token Compression in LVLMs: DART.

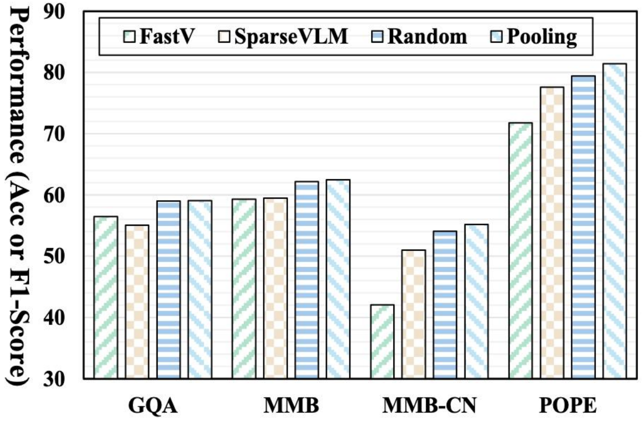




(a) Token Frequency



(b) Token Attention



Method	GQA	MMB	MMB-CN	MME	POPE	SQA	VQA^{Text}	VizWiz
			Upper Bound, 576 Tokens **(100%)**					
Vanilla	61.9	64.7	58.1	1862	85.9	69.5	58.2	50.0
			Retain 144 Tokens (↓ 75.0%)					
Random	59.0	62.2	54.1	1736	79.4	67.8	51.7	**51.9**
Pooling	59.1	**62.5**	**55.2**	**1763**	**81.4**	69.1	53.4	**51.9**
Window FastV	**59.2**	59.3	51.0	1737	80.3	66.4	50.8	50.3
Vanilla FastV	56.5	59.3	42.1	1689	71.8	65.3	53.6	51.3
Reverse FastV	49.9	36.9	26.4	1239	59.8	60.9	36.9	48.4
SparseVLM	55.1	59.5	51.0	1711	77.6	**69.3**	**54.9**	51.4

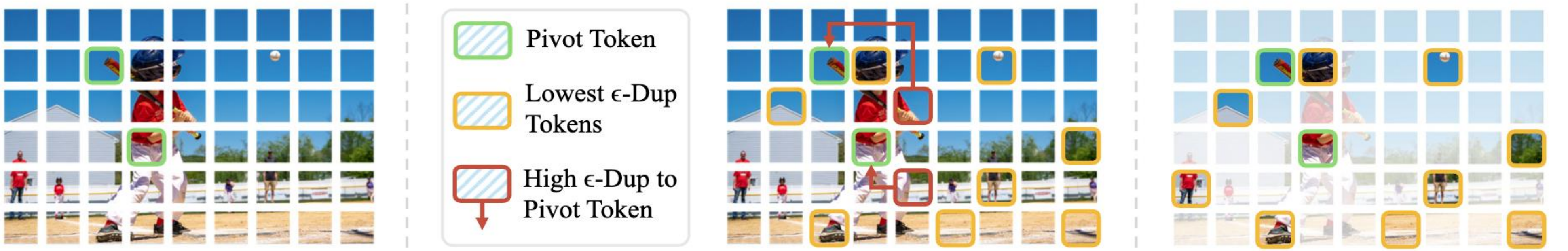
Limitations of Existing Methods:

(I) Ignoring interactions between tokens during pruning; (II) Incompatibility to efficient attention

(III) Bias in token positions; (IV) Significant accuracy drop

Zichen Wen, Yifeng Gao, Shaobo Wang, Weijia Li, Conghui He, Linfeng Zhang, et al. "Stop looking for important tokens in multimodal language models: Duplication matters more." arXiv preprint arXiv:2502.11494 (2025).
Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, Linfeng Zhang, "Token Pruning in Multimodal Large Language Models: Are We Solving the Right Problem?". In *Findings of the Association for Computational Linguistics (ACL)*, 2025.

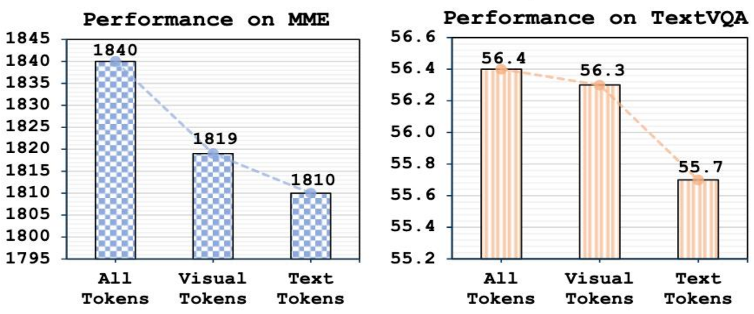
Token Compression in LVLMs: DART.



(a) Pivot Token Selection

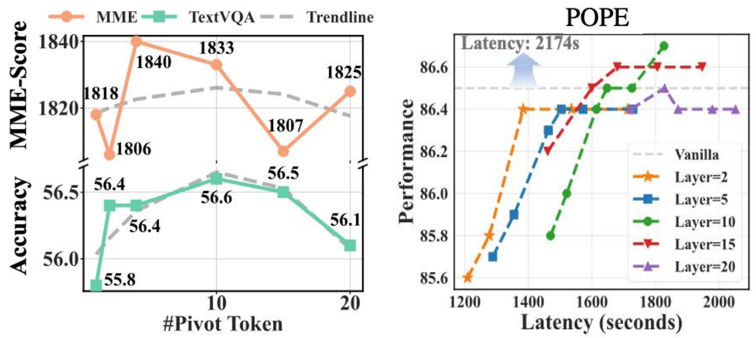
(b) Calculate ϵ-Duplicate Score Between Pivot Tokens and the Remainder

(c) Token Reduction to Keep Tokens With Least Duplication



(a) MME (b) TextVQA

Benchmark	Vanilla	Pivot Token Selection							Other Methods	
		Random	A-Score♠	A-Score♡	K-norm♠	K-norm♡	V-norm♠	V-norm♡	SparseVLM	FastV
GQA	61.9	$59.0_{\pm 0.3}$	59.2	58.4	58.7	59.1	57.3	59.4	56.0	49.6
MMB	64.7	$63.2_{\pm 0.7}$	63.1	62.9	63.2	64.0	62.5	64.3	60.0	56.1
MME	1862	$1772_{\pm 17.9}$	1826	1830	1840	1820	1760	1825	1745	1490
POPE	85.9	$80.6_{\pm 0.49}$	81.1	81.0	80.1	80.2	76.8	81.6	80.5	59.6
SQA	69.5	$69.0_{\pm 0.3}$	69.9	68.9	69.1	68.7	69.2	68.9	68.5	60.2
VQA^{V2}	78.5	$75.2_{\pm 0.2}$	75.9	76.0	75.9	75.6	75.4	76.1	73.8	61.8
VQA^{Text}	58.2	$56.0_{\pm 0.3}$	55.7	56.5	56.4	55.4	55.5	56.0	54.9	50.6
Avg.	100%	96.0%	96.9%	96.7%	96.8%	96.8%	94.9%	97.2%	93.9%	81.5%



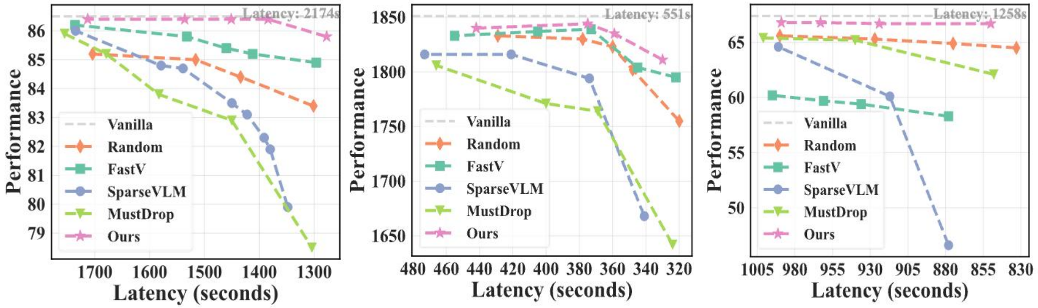
- **Beyond Token Importance**: Rethinking Reduction through Token Duplication.
- **Diverse Pivot Token Selection**: Towards More Robust Methods!

Zichen Wen, Yifeng Gao, Shaobo Wang, Weijia Li, Conghui He, Linfeng Zhang, et al. "Stop looking for important tokens in multimodal language models: Duplication matters more." arXiv preprint arXiv:2502.11494 (2025).

Token Compression in LVLMs: DART.

Method	GQA	MMB	MMB-CN	MME	POPE	SQA	VQA^{V2}	VQA^{Text}	VizWiz	OCRBench	Avg.
LLaVA-1.5-7B	*Upper Bound, 576 Tokens* **(100%)**										
Vanilla	61.9	64.7	58.1	1862	85.9	69.5	78.5	58.2	50.0	297	100.0%
LLaVA-1.5-7B	*Retain 192 Tokens* (↓ 66.7%)										
ToMe (ICLR23)	54.3	60.5	-	1563	72.4	65.2	68.0	52.1	-	-	-
FastV (ECCV24)	52.7	61.2	57.0	1612	64.8	67.3	67.1	52.5	50.8	291	91.2%
HiRED (AAAI25)	58.7	62.8	54.7	1737	82.8	68.4	74.9	47.4	50.1	190	91.5%
FitPrune (AAAI25)	**60.4**	63.3	56.4	1831	83.4	67.8	-	57.4	50.9	-	-
LLaVA-PruMerge (2024.05)	54.3	59.6	52.9	1632	71.3	67.9	70.6	54.3	50.1	253	90.8%
SparseVLM (ICML25)	57.6	62.5	53.7	1721	**83.6**	69.1	75.6	56.1	50.5	292	96.3%
PDrop (CVPR25)	57.1	63.2	56.8	1766	82.3	68.8	75.1	56.1	51.1	290	96.7%
FiCoCo-V (2024.11)	58.5	62.3	55.3	1732	82.5	67.8	74.4	55.7	51.0	-	96.1%
MustDrop (2024.11)	58.2	62.3	55.8	1787	82.6	69.2	76.0	56.5	**51.4**	289	97.2%
DART (Ours)	60.0	**63.6**	**57.0**	**1856**	82.8	**69.8**	**76.7**	**57.4**	51.2	**296**	**98.8%**
DART † (Ours)	60.9	66.3	59.5	1829	85.3	70.1	78.2	56.8	51.3	304	100.4%
LLaVA-1.5-7B	*Retain 128 Tokens* (↓ 77.8%)										
ToMe (ICLR23)	52.4	53.3	-	1343	62.8	59.6	63.0	49.1	-	-	-
FastV (ECCV24)	49.6	56.1	56.4	1490	59.6	60.2	61.8	50.6	51.3	285	86.4%
HiRED (AAAI25)	57.2	61.5	53.6	1710	79.8	68.1	73.4	46.1	51.3	191	90.2%
FitPrune (AAAI25)	58.5	62.7	56.2	1776	77.9	68.0	-	55.7	51.7	-	-
LLaVA-PruMerge (2024.05)	53.3	58.1	51.7	1554	67.2	67.1	68.8	54.3	50.3	248	88.8%
SparseVLM (ICML25)	56.0	60.0	51.1	1696	80.5	67.1	73.8	54.9	51.4	280	93.8%
PDrop (CVPR25)	56.0	61.1	56.6	1644	**82.3**	68.3	72.9	55.1	51.0	287	95.1%
FiCoCo-V (2024.11)	57.6	61.1	54.3	1711	82.2	68.3	73.1	55.6	49.4	-	94.9%
MustDrop (2024.11)	56.9	61.1	55.2	1745	78.7	68.5	74.6	56.3	**52.1**	281	95.6%
DART (Ours)	**58.7**	**63.2**	**57.5**	**1840**	80.1	**69.1**	**75.9**	**56.4**	51.7	**296**	**98.0%**
DART † (Ours)	59.8	65.6	58.3	1849	84.4	70.7	77.5	56.4	52.6	299	99.9%
LLaVA-1.5-7B	*Retain 64 Tokens* (↓ 88.9%)										
ToMe (ICLR23)	48.6	43.7	-	1138	52.5	50.0	57.1	45.3	-	-	-
FastV (ECCV24)	46.1	48.0	52.7	1256	48.0	51.1	55.0	47.8	50.8	245	77.3%
HiRED (AAAI25)	54.6	60.2	51.4	1599	73.6	68.2	69.7	44.2	50.2	191	87.0%
FitPrune (AAAI25)	52.3	58.5	49.7	1556	60.9	68.0	-	51.2	51.1	-	-
LLaVA-PruMerge (2024.05)	51.9	55.3	49.1	1549	65.3	68.1	67.4	54.0	50.1	250	87.4%
SparseVLM (ICML25)	52.7	56.2	46.1	1505	75.1	62.2	68.2	51.8	50.1	180	84.6%
PDrop (CVPR25)	41.9	33.3	50.5	1092	55.9	68.6	69.2	45.9	50.7	250	78.1%
FiCoCo-V (2024.11)	52.4	60.3	53.0	1591	**76.0**	68.1	71.3	53.6	49.8	-	91.5%
MustDrop (2024.11)	53.1	60.0	53.1	1612	68.0	63.4	69.3	54.2	51.2	267	90.1%
DART (Ours)	**55.9**	**60.6**	**53.2**	**1765**	73.9	**69.8**	**72.4**	**54.4**	**51.6**	**270**	**93.7%**
DART † (Ours)	57.1	64.7	56.7	1823	79.3	71.1	74.6	54.7	52.1	286	97.2%

Methods	Tokens ↓	Total Time ↓ (Min:Sec)	Prefilling Time ↓ (Min:Sec)	FLOPs ↓	KV Cache ↓ (MB)	POPE ↑ (F1-Score)	Speedup ↑ (Total)	Speedup ↑ (Prefilling)
Vanilla LLaVA-Next-7B	2880	36:16	22:51	100%	1512.1	86.5	1.00×	1.00×
+ FastV	320	18:17	7:41	**12.8%**	168.0	78.3	1.98×	2.97×
+ SparseVLM	320	23:11	-	15.6%	168.0	82.3	1.56×	-
+ DART	320	**18:13**	**7:38**	**12.8%**	168.0	**84.1**	**1.99×**	**2.99×**



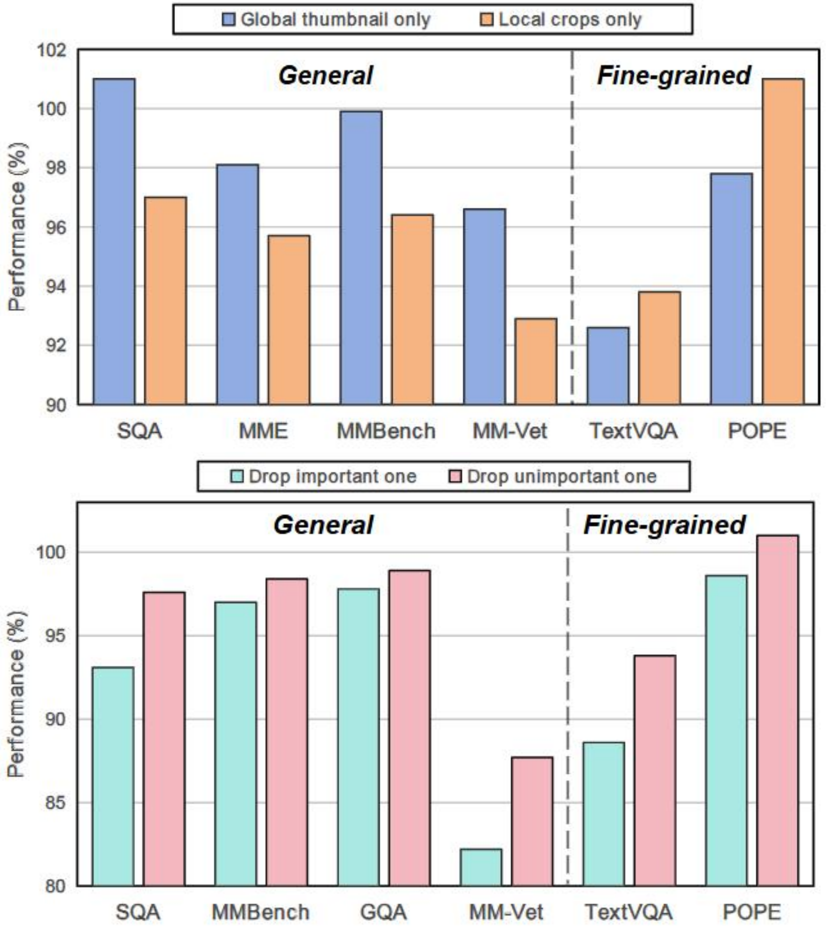
(a) POPE (b) MME (c) MMBench

Method	GQA	MMB	MMB-CN	MME	POPE	SQA	VQA^{V2}	VQA^{Text}	VizWiz	OCRBench	Avg.
LLaVA-Next-7B	*Upper Bound, 2880 Tokens* **(100%)**										
Vanilla	64.2	67.4	60.6	1851	86.5	70.1	81.8	64.9	57.6	517	100.0%
LLaVA-Next-7B	*Retain 320 Tokens* (↓ 88.9%)										
FastV (ECCV24)	55.9	61.6	51.9	1661	71.7	62.8	71.9	55.7	53.1	374	86.4%
HiRED (AAAI25)	59.3	64.2	55.9	1690	83.3	66.7	75.7	58.8	54.2	404	91.8%
LLaVA-PruMerge (2024.05)	53.6	61.3	55.3	1534	60.8	66.4	69.7	50.6	54.0	146	79.9%
SparseVLM (ICML25)	56.1	60.6	54.5	1533	82.4	66.1	71.5	58.4	52.0	270	85.9%
PDrop (CVPR25)	56.4	63.4	56.2	1663	77.6	67.5	73.5	54.4	54.1	259	86.8%
MustDrop (2024.11)	57.3	62.8	55.1	1641	82.1	68.0	73.7	**59.9**	54.0	382	90.4%
FasterVLM (2024.12)	56.9	61.6	53.5	1701	83.6	66.5	74.0	56.5	52.6	401	89.8%
GlobalCom² (2025.01)	57.1	61.8	53.4	1698	83.8	67.4	76.7	57.2	54.6	375	90.3%
DART (Ours)	**61.7**	**65.3**	**58.2**	**1710**	**84.1**	**68.4**	**79.1**	58.7	**56.1**	**406**	**93.9%**

DART achieves state-of-the-art efficiency and performance across various models and benchmarks !

Zichen Wen, Yifeng Gao, Shaobo Wang, Weijia Li, Conghui He, Linfeng Zhang, et al. "Stop looking for important tokens in multimodal language models: Duplication matters more." arXiv preprint arXiv:2502.11494 (2025).

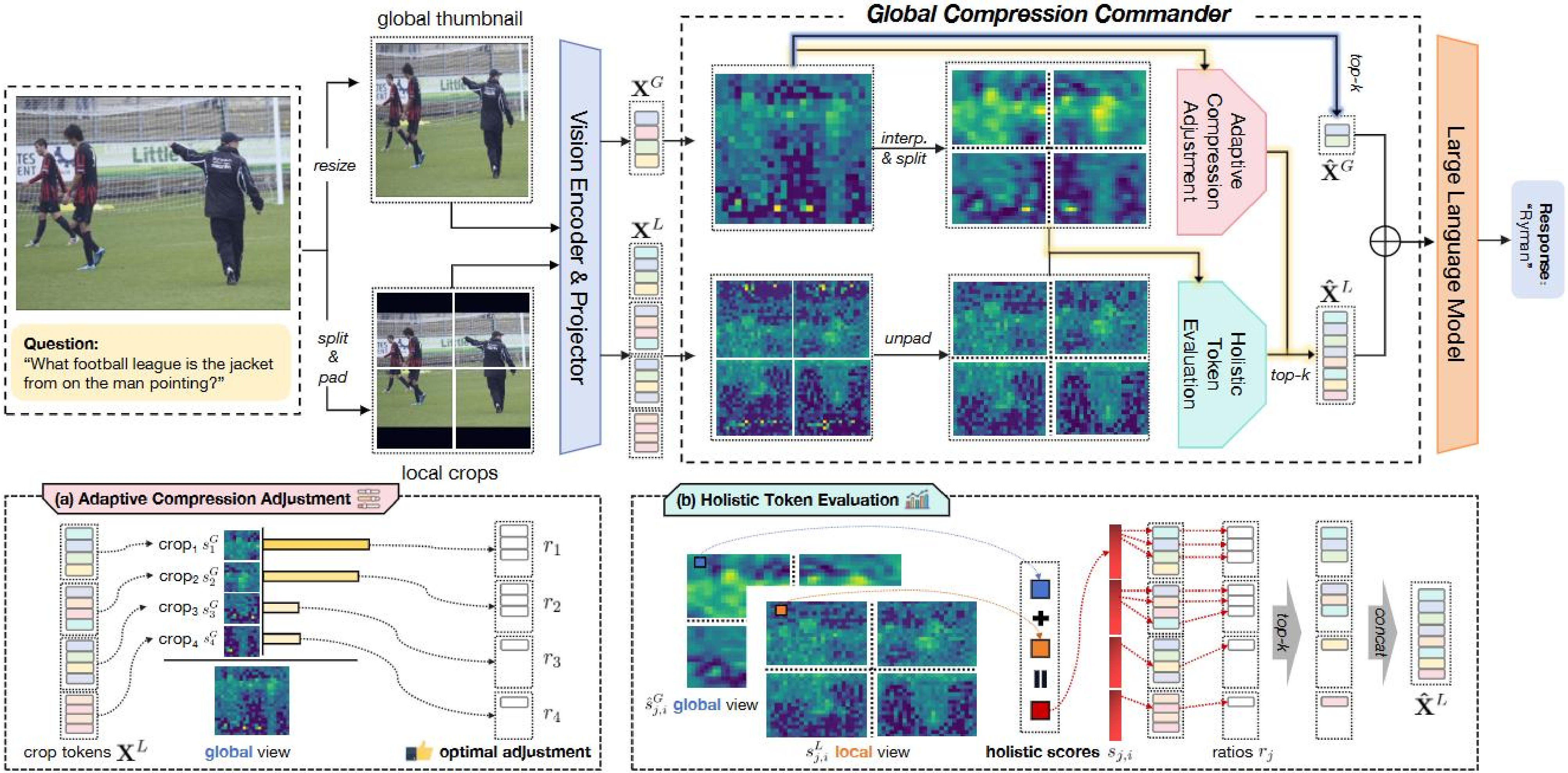
Token Compression in HR-LVLMs: GlobalCom2.



The complementary roles between thumbnails and tiles and the inherent characteristics among tiles.

Xuyang Liu, Ziming Wang, Yuhang Han, Yingyao Wang, et al., "Global Compression Commander: Plug-and-Play Inference Acceleration for High-Resolution Large Vision-Language Models". arXiv preprint arXiv:2501.05179.

Token Compression in HR-LVLMs: GlobalCom^2.



GlobalCom^2 adaptively compresses local tile tokens based on its semantic richness in global thumbnail.

Xuyang Liu, Ziming Wang, Yuhang Han, Yingyao Wang, et al., "Global Compression Commander: Plug-and-Play Inference Acceleration for High-Resolution Large Vision-Language Models". arXiv preprint arXiv:2501.05179.

Token Compression in HR-LVLMs: GlobalCom^2.

GlobalCom^2 adaptively compresses local tile tokens based on its semantic richness in global thumbnail.

Xuyang Liu, Ziming Wang, Yuhang Han, Yingyao Wang, et al., "Global Compression Commander: Plug-and-Play Inference Acceleration for High-Resolution Large Vision-Language Models". arXiv preprint arXiv:2501.05179.

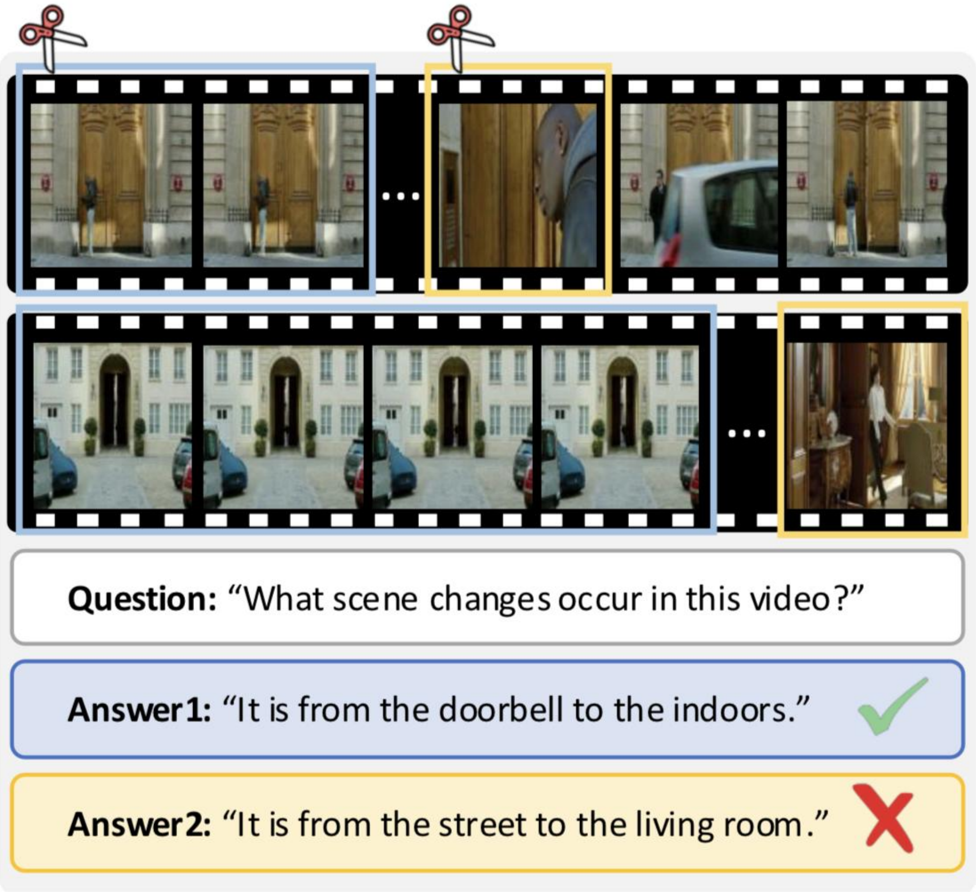
Token Compression in HR-LVLMs: GlobalCom2.

Method	VQAv2	GQA	VizWiz	SQA	VQA^T	POPE	MME	MMB	MMBCN	MM-Vet	Average
Upper Bound, 2880 Tokens											
LLaVA-NeXT-7B [29]	81.8	64.2	57.6	70.1	64.9	86.5	1519.0	67.4	60.6	43.9	100.0%
Ratio=75%, Retain up to 2160 Tokens											
FastV [8]	81.1	62.5	55.1	**69.3**	59.7	86.3	1506.3	67.6	59.0	41.7	97.5%
SparseVLM [53]	81.1	62.6	55.2	68.5	60.3	73.2	1507.8	66.1	58.6	**41.9**	95.7%
FasterVLM [52]	81.1	63.7	**56.5**	68.4	59.1	87.5	1533.4	67.5	60.2	38.5	97.4%
GlobalCom^2	**81.3**	**63.8**	**56.5**	68.7	**62.5**	**87.8**	**1548.4**	**68.0**	**60.6**	40.6	**98.8%**
Ratio=50%, Retain up to 1440 Tokens											
FastV [8]	80.7	61.8	54.9	**69.1**	59.6	85.5	1490.3	67.4	58.5	**41.2**	96.8%
SparseVLM [53]	**80.9**	62.0	55.7	68.1	60.0	73.4	1484.9	65.7	58.9	39.9	95.0%
FasterVLM [52]	80.6	63.4	56.4	**69.1**	58.9	87.7	1533.3	67.4	**60.4**	39.6	97.7%
GlobalCom^2	80.6	**63.9**	**56.5**	68.5	**62.3**	**88.1**	**1552.9**	**67.6**	60.5	40.4	**98.6%**
Ratio=25%, Retain up to 720 Tokens											
FastV [8]	78.9	60.4	54.2	**69.8**	58.4	83.1	1477.3	65.6	57.0	**41.1**	95.3%
SparseVLM [53]	78.9	60.9	55.6	67.5	58.1	71.0	1446.1	63.8	57.0	38.0	92.6%
FasterVLM [52]	78.3	61.3	55.4	67.1	58.8	87.2	1454.6	**66.0**	**58.4**	37.8	95.1%
GlobalCom^2	**79.4**	**61.4**	**55.7**	68.1	**60.9**	**87.6**	**1493.5**	65.9	58.0	40.7	**96.6%**
Ratio=10%, Retain up to 288 Tokens											
FastV [8]	71.9	55.9	53.1	**69.3**	55.7	71.7	1282.9	61.6	51.9	33.7	87.3%
SparseVLM [53]	71.6	56.1	53.2	68.6	52.0	63.2	1332.2	54.5	50.7	24.7	82.7%
FasterVLM [52]	74.0	56.9	52.6	66.5	56.5	83.6	1359.2	61.6	**53.5**	35.0	89.8%
GlobalCom^2	**76.7**	**57.1**	**54.6**	68.7	**58.4**	**83.8**	**1365.5**	**61.8**	53.4	**36.4**	**91.5%**

GlobalCom2 achieves SoTA performance! With only 10% tokens, it achieves 91.5% performance!

Xuyang Liu, Ziming Wang, Yuhang Han, Yingyao Wang, et al., "Global Compression Commander: Plug-and-Play Inference Acceleration for High-Resolution Large Vision-Language Models". arXiv preprint arXiv:2501.05179.

Token Compression in VideoLLMs: VidCom²



Methods	Pre-LLM	Intra-LLM	[CLS] Dependency	Video-Specific	Frame Uniqueness	Efficient Attention
FastV		✓				
PDrop		✓				
SparseVLM		✓				
MUSTDrop	✓	✓	✓			
FiCoCo	✓	✓	✓			
FasterVLM	✓		✓			✓
DyCoke	✓			✓		✓
VidCom²	✓			✓	✓	✓

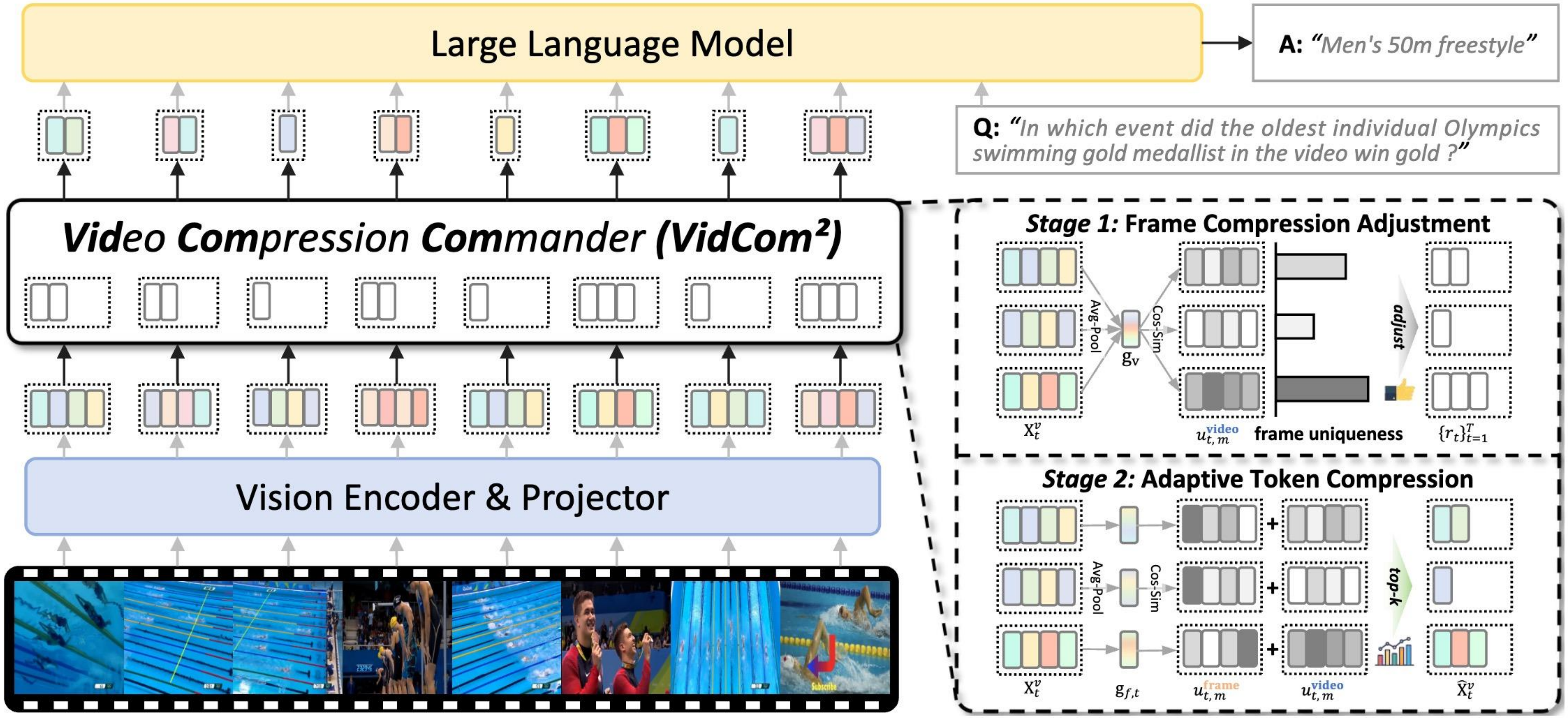
Current methods face ***two critical issues*** for VideoLLMs:

- **Design Myopia:** ignoring frame uniqueness, leading to over-compression of distinctive video information.
- **Implementation Constraints:** limited to the specific architectures or incompatible with Flash Attention.

Takeaway: We derive ***three key principles*** for effective token compression of VideoLLMs: (i) model adaptability, (ii) frame uniqueness, and (iii) operator compatibility.

Xuyang Liu, Yiyu Wang, Junpeng Ma, Linfeng Zhang, "Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models". arXiv preprint arXiv:2505.14454.

Token Compression in VideoLLMs: VidCom^2.

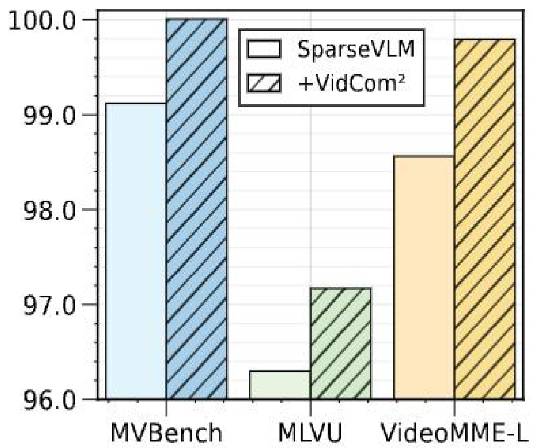


VidCom^2 dynamically compresses video tokens based on frame uniqueness.

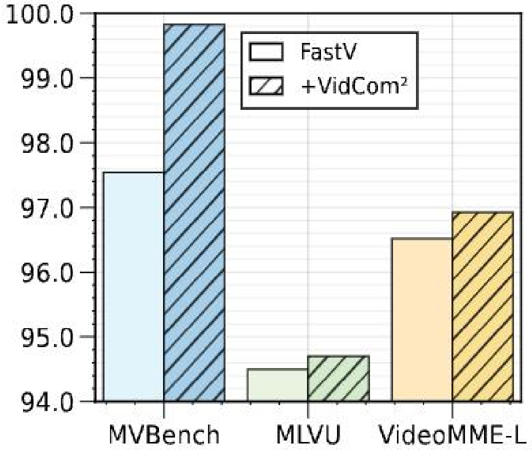
Xuyang Liu*, Yiyu Wang*, Junpeng Ma, Linfeng Zhang, "Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models". arXiv preprint arXiv:2505.14454.

Token Compression in VideoLLMs: VidCom2.

Methods	MVBench	LongVideoBench	MLVU	VideoMME Overall	VideoMME Short	VideoMME Medium	VideoMME Long	Average (%)
Upper Bound								
LLaVA-OV-7B	56.9	56.4	63.0	58.6	70.3	56.6	48.8	100.0
Retention Ratio=30%								
DyCoke [CVPR'25]	56.6	54.7	60.3	56.1	67.1	54.6	46.6	96.5
Retention Ratio=25%								
Random	54.2	52.7	59.7	55.6	65.4	53.0	48.3	94.8
FastV [ECCV'24]	55.5	53.3	59.6	55.3	65.0	53.8	47.0	94.9
PDrop [CVPR'25]	55.3	51.3	57.1	55.5	64.7	53.1	48.7	94.1
SparseVLM [ICML'25]	56.4	53.9	60.7	57.3	68.4	55.2	48.1	97.5
DyCoke [CVPR'25]	49.5	48.1	55.8	51.0	61.1	48.6	43.2	87.0
VidCom^2	**57.2**	**54.9**	**62.5**	**58.6**	**69.8**	**56.4**	**49.4**	**99.6**



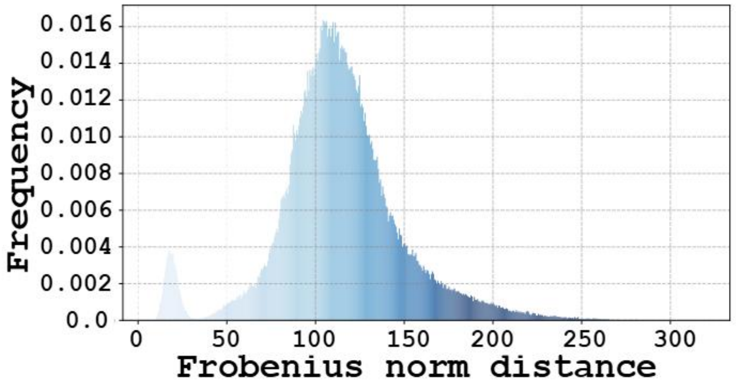
Methods	LLM Generation↓ Latency (s)	Model Generation↓ Latency (s)	Total↓ Latency (min:sec)	GPU Peak↓ Memory (GB)	Throughput↑ (samples/s)	Performance↑
LLaVA-OV-7B	618.0	1008.4	26:03	17.7	0.64	56.9
Retention Ratio=25%						
Random	178.2 (↓71.2%)	566.0 (↓43.9%)	18:44 (↓28.1%)	16.0 (↓9.6%)	0.89 (1.39×)	54.6 (↓2.3)
FastV [ECCV'24]	260.9 (↓57.8%)	648.6 (↓35.7%)	20:07 (↓22.8%)	24.7 (↑39.5%)	0.83 (1.30×)	55.5 (↓1.4)
PDrop [CVPR'25]	205.6 (↓66.7%)	592.6 (↓41.2%)	18:50 (↓27.7%)	24.5 (↑38.4%)	0.88 (1.38×)	55.3 (↓1.6)
SparseVLM [ICML'25]	410.6 (↓33.6%)	807.7 (↓19.9%)	25:03 (↓3.8%)	27.1 (↑53.1%)	0.67 (1.05×)	56.4 (↓0.5)
DyCoke [CVPR'25]	205.2 (↓66.8%)	598.0 (↓40.7%)	18:56 (↓27.4%)	16.1 (↓9.0%)	0.88 (1.38×)	49.5 (↓7.4)
VidCom^2	**180.7** (↓70.8%)	**574.7** (↓43.0%)	**18:46** (↓28.0%)	**16.0** (↓9.6%)	**0.88** (1.38×)	**57.2** (↑0.3)



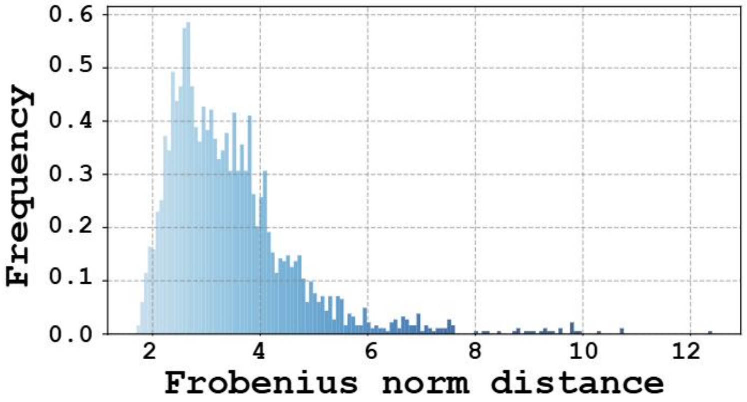
VidCom2 achieves state-of-the-art efficiency and performance across models and benchmarks.

Xuyang Liu, Yiyu Wang, Junpeng Ma, Linfeng Zhang, "Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models". arXiv preprint arXiv:2505.14454.

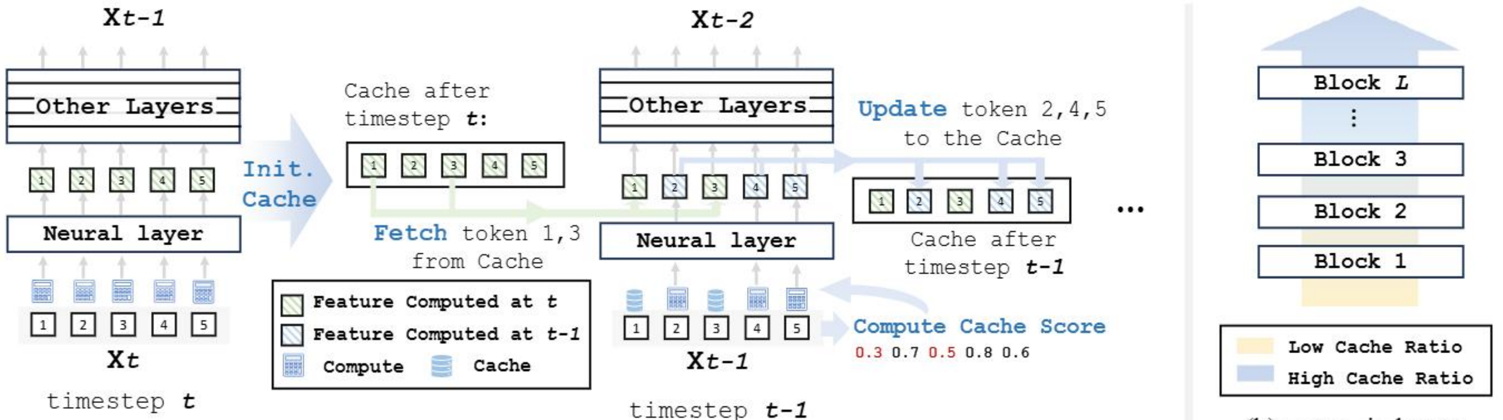
Token Compression in DiTs: ToCa.



Difference in Temporal Redundancy

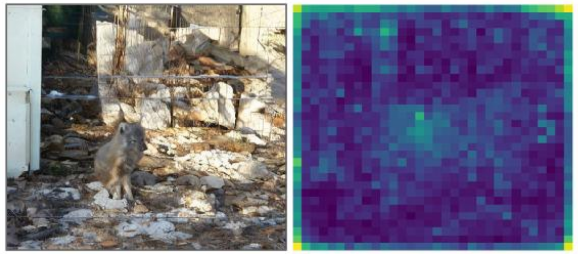


Difference in Error Propagation



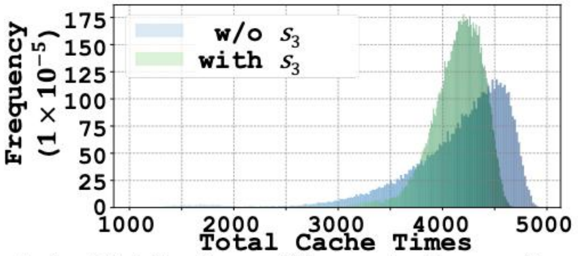
(a) Overview of ToCa in timestep t and $t-1$

(b) ToCa in layers at different depth



(a) ToCa w/o Cache Frequency Score s_3

(b) ToCa with Cache Frequency Score s_3



(c) Distributions of the cache times of each token (with and w/o s_3)

ToCa achieves fine-grained token-wise feature caching for DiT-based generation models.

Chang Zou*, Xuyang Liu*, Ting Liu, Siteng Huang, Linfeng Zhang, "Accelerating Diffusion Transformers with Token-wise Feature Caching". In *International Conference on Learning Representations (ICLR)*, 2025.

Token Compression in DiTs: ToCa.

Method	Latency(s) ↓	FLOPs ↓	Speed ↑	MS-COCO2017 FID-30k ↓	MS-COCO2017 CLIP ↑	PartiPrompts CLIP ↑
PixArt-α (Chen et al., 2024a)	0.682	11.18	1.00×	28.09	16.32	16.70
50% **steps**	0.391	5.59	2.00×	37.46	15.85	16.37
FORA($\mathcal{N}=2$) (Selvaraju et al., 2024)	0.416	5.66	1.98×	29.67	16.40	17.19
FORA($\mathcal{N}=3$) (Selvaraju et al., 2024)	0.342	4.01	2.79×	29.88	16.42	17.15
ToCa ($\mathcal{N}=3, R=60\%$)	0.410	6.33	1.77×	**28.02**	**16.45**	17.15
ToCa ($\mathcal{N}=3, R=70\%$)	0.390	5.78	1.93×	28.33	16.44	17.75
ToCa ($\mathcal{N}=3, R=80\%$)	0.370	5.05	2.21×	28.82	16.44	**17.83**
ToCa ($\mathcal{N}=3, R=90\%$)	0.347	4.26	2.62×	29.73	16.45	17.82

Method	Latency(s) ↓	FLOPs(T) ↓	Speed ↑	VBench(%) ↑
OpenSora (Zheng et al., 2024)	81.18	3283.20	1.00×	79.13
Δ-DiT* (Chen et al., 2024b)	79.14	3166.47	1.04×	78.21
T-GATE* (Zhang et al., 2024b)	67.98	2818.40	1.16×	77.61
PAB[1]* (Zhao et al., 2024)	60.78	2657.70	1.24×	78.51
PAB[2]* (Zhao et al., 2024)	59.16	2615.15	1.26×	77.64
PAB[3]* (Zhao et al., 2024)	56.64	2558.25	1.28×	76.95
50% **steps**	42.72	1641.60	2.00×	76.78
FORA(Selvaraju et al., 2024)	49.26	1751.32	1.87×	76.91
ToCa($R=80\%$)	43.52	1439.70	2.28×	**78.59**
ToCa($R=85\%$)	43.08	1394.03	**2.36**×	78.34

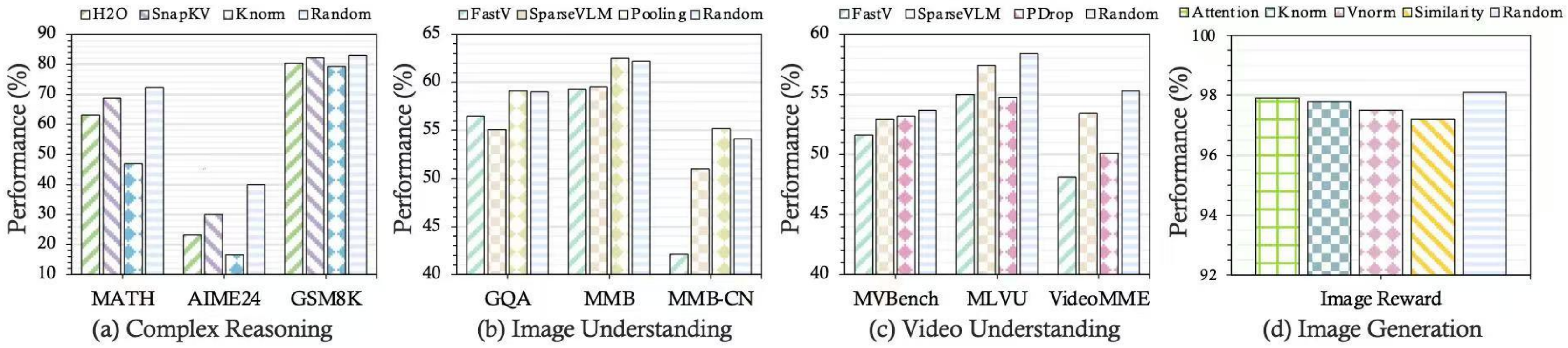
ToCa achieves nearly lossless speedups of 1.51 × , 1.93 × , and 2.36 × , and 2.75 × on FLUX, PixArt-α, OpenSora, and DiT-XL models respectively, while maintaining generation quality.

Chang Zou*, Xuyang Liu*, Ting Liu, Siteng Huang, Linfeng Zhang, "Accelerating Diffusion Transformers with Token-wise Feature Caching". In *International Conference on Learning Representations (ICLR)*, 2025.

Compelling Advantages of Token Compression:

- **Universal Applicability:** The redundancy of tokens exists consistently across modalities and tasks, making token compression possible in all kinds of settings.
- **Dual-phase Efficiency:** Token compression is capable of accelerating both model training and inference phases with minimal accuracy loss.
- **Architectural Compatibility:** Token compression is orthogonal to existing model compression methods, making it possible to be integrated seamlessly with them. Besides, it is friendly to hardware and computer systems.
- **Low Implementation Costs:** Modern neural networks, such as transformers, are able to process tokens of different lengths. As a result, token compression can be done without introducing any training costs or changes to data utilization.
- **Quadratic Gains:** The $O(n^2)$ computation complexity of widely used self-attention indicates that token compression can bring significant benefits in computation efficiency.

Current Challenges: performance degradation.



Method	RefCOCO TestA	RefCOCO TestB	RefCOCO Val	RefCOCO+ TestA	RefCOCO+ TestB	RefCOCO+ Val	RefCOCOg Test	RefCOCOg Val	Avg.
Vanilla	72.3	51.5	73.3	66.3	29.7	68.3	49.5	51.5	100%
SparseVLMs	4.0	6.9	4.0	2.9	5.9	0.9	7.9	5.9	4.8% (↓ 95.2%)
Vanilla FastV	20.8	13.9	27.7	17.8	8.9	26.7	19.8	14.9	18.8% (↓ 81.2%)
Window FastV	22.8	22.8	25.7	18.8	7.9	18.8	20.8	23.8	20.2% (↓ 79.8%)
Random	22.8	27.7	32.7	17.8	13.9	32.7	20.8	16.8	23.2% (↓ 76.8%)
Pooling	34.7	17.8	26.7	23.8	17.8	24.8	14.9	20.8	22.7% (↓ 77.3%)

Performance degradation is largely affected by methodological bottlenecks and inherent limitations.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, Linfeng Zhang, "Token Pruning in Multimodal Large Language Models: Are We Solving the Right Problem?". In *Findings of the Association for Computational Linguistics (ACL)*, 2025.

Current Challenges: fair comparison.

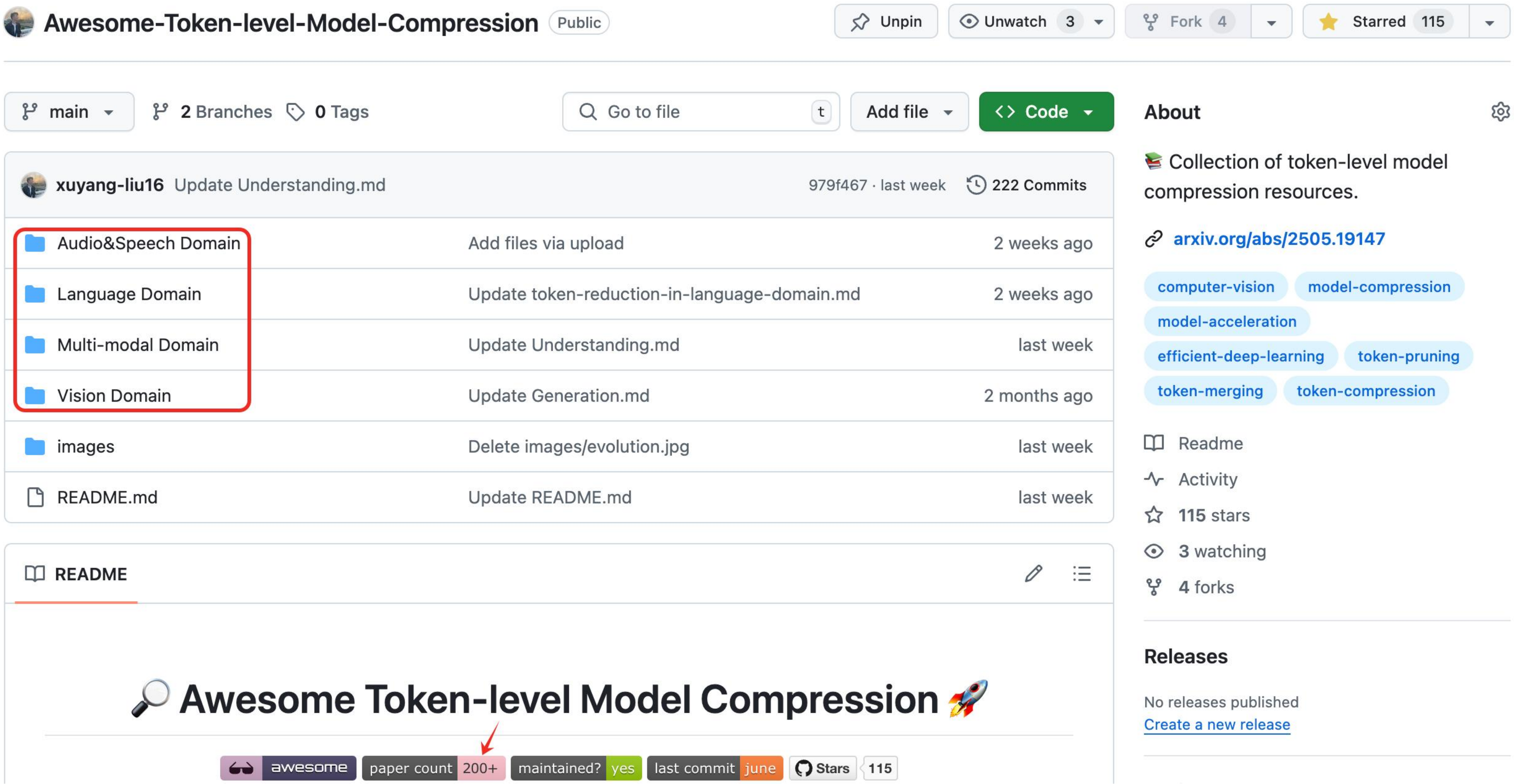
Methods	LLM Generation↓ Latency (s)	Model Generation↓ Latency (s)	Total↓ Latency (min:sec)	GPU Peak↓ Memory (GB)	Throughput↑ (samples/s)	Performance↑
LLaVA-OV-7B	618.0	1008.4	26:03	17.7	0.64	56.9
Retention Ratio=25%						
Random	178.2 (↓71.2%)	566.0 (↓43.9%)	18:44 (↓28.1%)	16.0 (↓9.6%)	0.89 (1.39×)	54.6 (↓2.3)
FastV [ECCV'24]	260.9 (↓57.8%)	648.6 (↓35.7%)	20:07 (↓22.8%)	24.7 (↑39.5%)	0.83 (1.30×)	55.5 (↓1.4)
PDrop [CVPR'25]	205.6 (↓66.7%)	592.6 (↓41.2%)	18:50 (↓27.7%)	24.5 (↑38.4%)	0.88 (1.38×)	55.3 (↓1.6)
SparseVLM [ICML'25]	410.6 (↓33.6%)	807.7 (↓19.9%)	25:03 (↓3.8%)	27.1 (↑53.1%)	0.67 (1.05×)	56.4 (↓0.5)
DyCoke [CVPR'25]	205.2 (↓66.8%)	598.0 (↓40.7%)	18:56 (↓27.4%)	16.1 (↓9.0%)	0.88 (1.38×)	49.5 (↓7.4)
VidCom^2	**180.7** (↓70.8%)	**574.7** (↓43.0%)	**18:46** (↓28.0%)	**16.0** (↓9.6%)	**0.88** (1.38×)	**57.2** (↑0.3)

Method	TFLOPs↓	Peak Memory (GB)↓	KV-Cache (MB)↓	Prefill Time (ms)↓	Throughput (samples/s)↑	Performance↑
Upper Bound, 2880 Tokens						
LLaVA-NeXT-7B	41.7	23.8	1536.0	170.7	2.5	1519.0
Ratio=75%, Retain up to 2160 Tokens						
GlobalCom^2	30.4 (↓27%)	17.8 (↓25%)	1126.4 (↓27%)	119.9 (↓30%)	3.2 (1.3×)	1548.4
Ratio=50%, Retain up to 1440 Tokens						
GlobalCom^2	19.7 (↓53%)	16.2 (↓32%)	755.0 (↓51%)	74.5 (↓57%)	4.2 (1.7×)	1552.9
Ratio=25%, Retain up to 720 Tokens						
GlobalCom^2	9.6 (↓77%)	14.8 (↓39%)	377.0 (↓76%)	34.6 (↓80%)	5.3 (2.1×)	1493.5

We advocate for more efficiency metrics to present comprehensive efficiency analysis.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, Linfeng Zhang, "Token Pruning in Multimodal Large Language Models: Are We Solving the Right Problem?". In *Findings of the Association for Computational Linguistics (ACL)*, 2025.

Paper Lists: 200+ token compression papers (maintained).



Thanks!

Q & A