*Fi*lter, *Co*rrelate, *Co*mpress: Training-Free Token Reduction for MLLM Acceleration

Yuhang Han[1]*, Xuyang Liu[2]*, Zihan Zhang[3], Pengxiang Ding[1], Junjie Chen[2], Donglin Wang[1], Honggang Chen[2], Qingsen Yang[4,5], Siteng Huang[6]†

[1] WU [2] SCU [3] JHU [4] NPU [5] SRI-NPU [6] ZJU

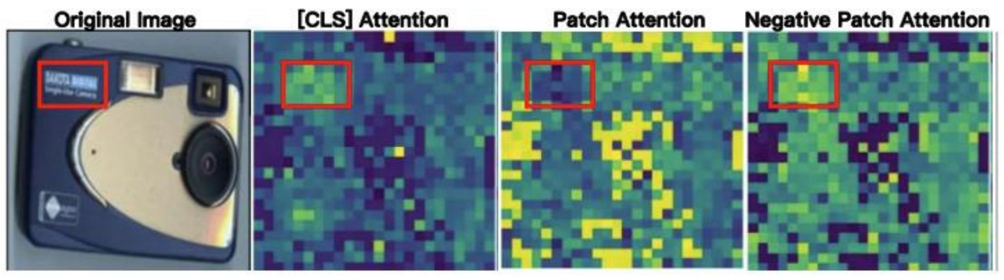
*Equal contribution. † Corresponding author: siteng.huang@gmail.com

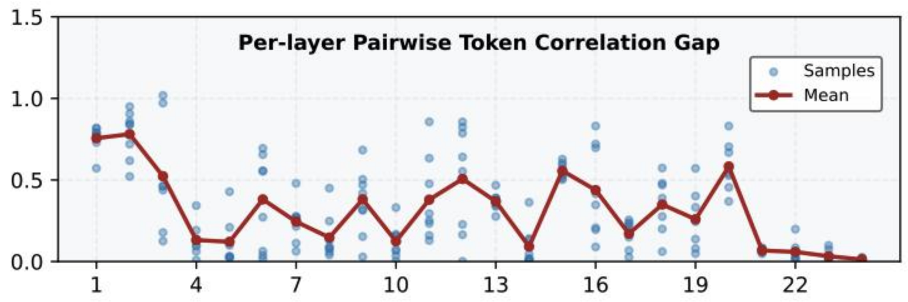
Association for the Advancement of Artificial Intelligence 2026

Motivation and Research Status

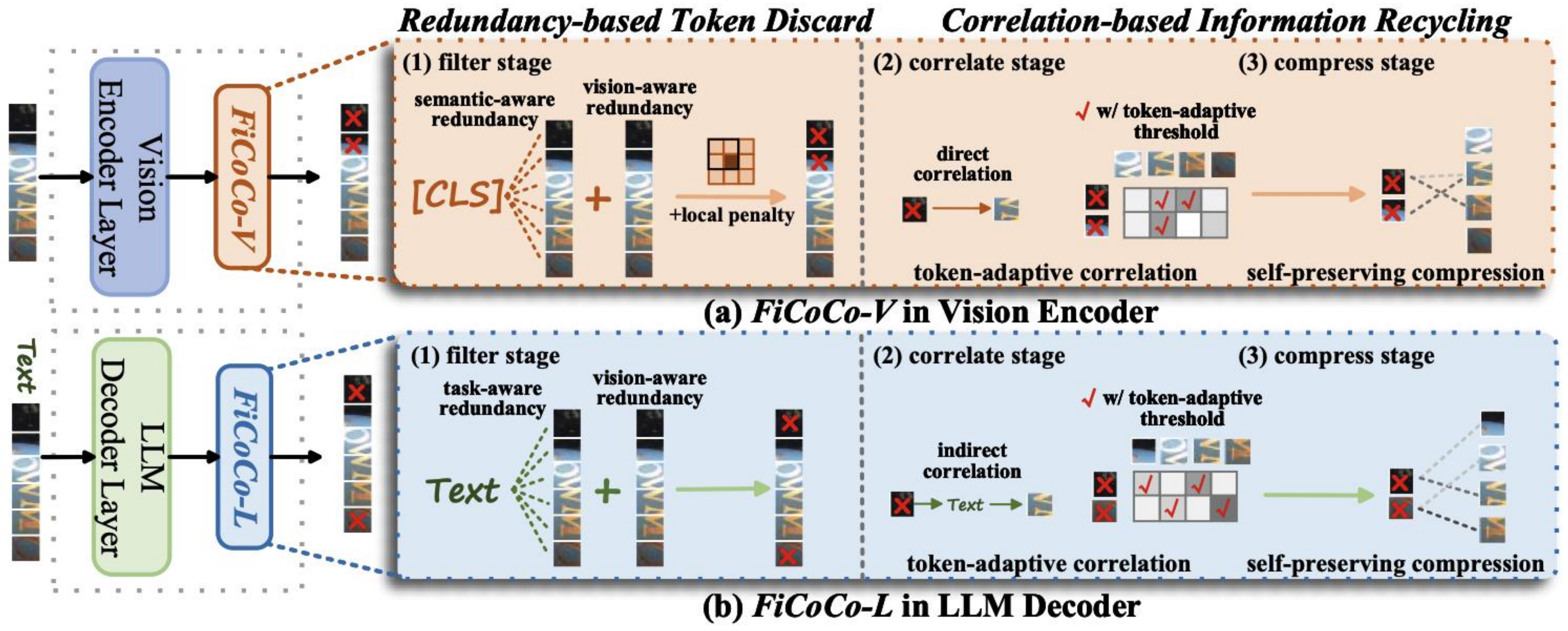
Current methods face ***three critical issues***:

- **Redundancy Myopia:** Single-metric redundancy modeling, limiting token importance estimation.
- **Coarse Token Retention:** Direct pruning or one-to-one merging of redundant tokens, losing fine-grained information.
- **Entangled Compression Pipeline:** Tight coupling of redundancy estimation and compression, limiting interpretable information flow.





Our Solution: *FiCoCo*



(a) *FiCoCo-V* in Vision Encoder

(b) *FiCoCo-L* in LLM Decoder

Different stages are designed to address different problems:

- **Filter stage:** What token should be discarded?
- **Correlate stage:** Where should discarded information be recycled?
- **Compress stage:** How to effectively recycle information?

Performance and Efficiency on LLaVA-1.5 7B/13B

- **Strong Performance:** 82.4%/47.6% TFLOPs reduction while retaining ⩾92%/95% performance (7B/13B).

Method	Source	TFLOPs↓	SQA	VQA^{T}	POPE	GQA	MMB	VQAv2	Avg	Avg(%)
					TFLOPs=8.5					
LLaVA-1.5-7B	*NeurIPS23*	8.5	69.5	58.2	86.4	62.5	66.1	79.1	70.3	100
					TFLOPs=3.3$_{(\downarrow 61.2\%)}$					
ToMe	*ICLR23*	3.3	65.2	52.1	72.4	54.3	60.5	68.0	62.1	88.3
FastV	*ECCV24*	3.3	67.3	52.5	64.8	52.7	61.2	67.1	60.9	86.6
SparseVLM	*ICML25*	3.3	69.1	56.1	83.6	57.6	62.5	75.6	67.4	95.9
PDrop	*CVPR25*	3.3	68.8	56.1	82.3	57.1	63.2	75.1	67.1	95.4
PruMerge	*ICCV25*	3.3	67.9	54.3	71.3	54.3	59.6	70.6	63.0	89.6
FiCoCo-V	Ours	3.3	67.8	55.7	82.5	58.5	62.3	74.4	66.9	95.2
FiCoCo-L	Ours	3.3	**69.6**	**56.6**	**84.6**	**61.1**	**64.6**	**76.8**	**68.9**	**98.0**
					TFLOPs=2.4$_{(\downarrow 71.8\%)}$					
ToMe	*ICLR23*	2.5	59.6	49.1	62.8	52.4	53.3	63.0	56.7	80.7
FastV	*ECCV24*	2.5	60.2	50.6	59.6	49.6	56.1	61.8	56.3	80.1
SparseVLM	*ICML25*	2.5	67.1	54.9	80.5	56.0	60.0	**73.8**	65.4	93.0
PDrop	*CVPR25*	2.5	68.3	55.1	82.3	56.0	61.1	72.9	65.9	93.8
PruMerge	*ICCV25*	2.5	67.1	54.3	67.2	53.3	58.1	68.8	61.5	87.5
FiCoCo-V	Ours	2.4	68.3	55.6	82.2	57.6	61.1	73.1	66.3	94.3
FiCoCo-L	Ours	2.4	**69.4**	**56.3**	**84.4**	**60.6**	**61.9**	73.4	**67.7**	**96.3**
					TFLOPs=1.5$_{(\downarrow 82.4\%)}$					
ToMe	*ICLR23*	1.6	50.0	45.3	52.5	48.6	43.7	57.1	49.5	70.4
FastV	*ECCV24*	1.6	51.1	47.8	48.0	46.1	48.0	61.8	50.5	71.8
SparseVLM	*ICML25*	1.5	62.2	51.8	75.1	52.4	56.2	68.2	61.0	86.8
PDrop	*CVPR25*	1.6	68.6	45.9	55.9	41.9	33.3	69.2	52.5	74.6
PruMerge	*ICCV25*	1.5	68.1	54.0	65.3	51.9	55.3	67.4	60.3	85.8
FiCoCo-V	Ours	1.5	68.4	55.5	79.8	**54.9**	60.2	**72.1**	65.2	92.7
FiCoCo-L	Ours	1.5	**69.5**	**55.7**	**82.1**	53.2	**61.5**	69.7	**65.3**	**92.8**
					TFLOPs=24.9					
LLaVA-1.5-13B	*NeurIPS23*	24.9	71.4	61.3	86.2	63.4	68.0	80.0	71.7	100
					TFLOPs=15.4$_{(\downarrow 47.6\%)}$					
FastV	*ECCV24*	15.4	57.0	56.0	79.3	57.7	57.9	-	61.6	85.9
SparseVLM	*ICML25*	15.4	69.9	49.9	81.1	57.9	65.8	-	64.9	90.5
FiCoCo-V	Ours	15.4	72.1	57.2	82.3	59.2	63.1	76.8	68.5	95.5
FiCoCo-L	Ours	15.4	**72.4**	**58.3**	**83.1**	**60.1**	**65.2**	**77.6**	**69.5**	**96.9**

Efficiency on LLaVA-1.5 7B/13B

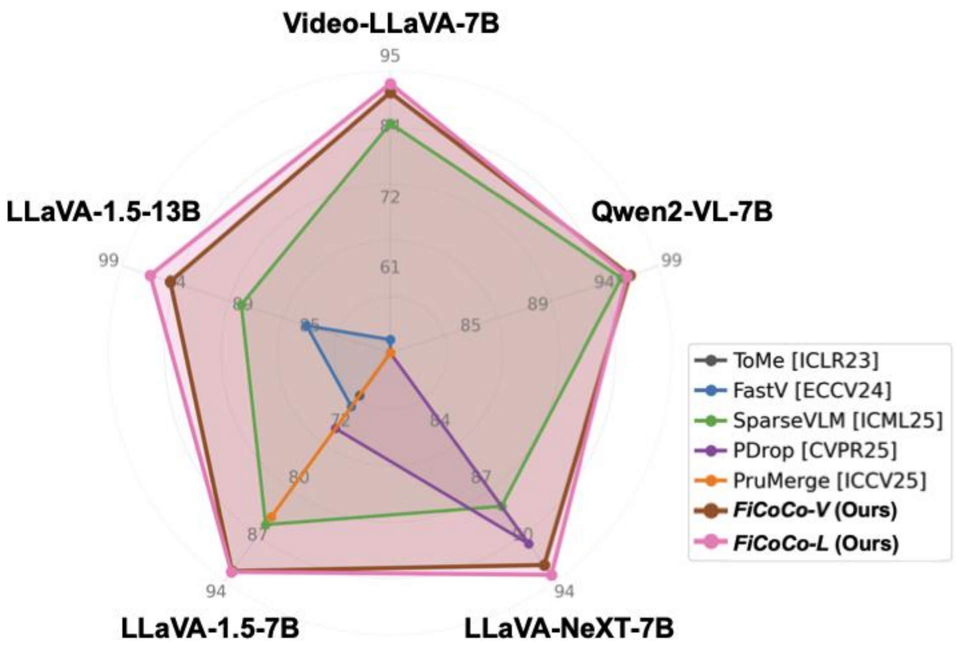
- **High Efficiency:** TFLOPs, memory footprint, and KV-cache size are all significantly reduced.

Method	Quant	TFLOPs↓	Memory (GB)↓	KV-Cache (MB)↓
LLaVA-1.5	FP16	8.5	22.4	333
FiCoCo-V	FP16	1.5 (↓82%)	14.4 (↓36%)	65.0 (↓80%)
FiCoCo-L	FP16	1.5 (↓82%)	14.3 (↓36%)	64.2 (↓81%)
LLaVA-1.5	INT8	4.3	11.2	167
FiCoCo-V	INT8	0.8 (↓81%)	7.8 (↓30%)	32.5 (↓81%)
FiCoCo-L	INT8	0.8 (↓81%)	7.2 (↓36%)	32.1 (↓81%)
LLaVA-1.5	INT4	2.1	6.2	83.4
FiCoCo-V	INT4	0.4 (↓81%)	4.4 (↓29%)	16.3 (↓81%)
FiCoCo-L	INT4	0.4 (↓81%)	3.3 (↓47%)	16.1 (↓81%)

Method	Quant	TFLOPs↓	Memory (GB)↓	KV-Cache (MB)↓
LLaVA-1.5	FP16	28.6	56.1	891
FiCoCo-V	FP16	15.4 (↓46%)	38.6 (↓31%)	488 (↓43%)
FiCoCo-L	FP16	15.4 (↓46%)	38.4 (↓32%)	485 (↓46%)
LLaVA-1.5	INT8	14.3	28	446
FiCoCo-V	INT8	7.7 (↓46%)	19.3 (↓31%)	244 (↓45%)
FiCoCo-L	INT8	7.7 (↓46%)	19.2 (↓31%)	242 (↓46%)
LLaVA-1.5	INT4	7.6	14	223
FiCoCo-V	INT4	3.9 (↓46%)	9.6 (↓32%)	122 (↓49%)
FiCoCo-L	INT4	3.9 (↓49%)	9.5 (↓32%)	121 (↓46%)

Paper Code WeChat

Comparison to existing methods



Ablation Study and Analysis

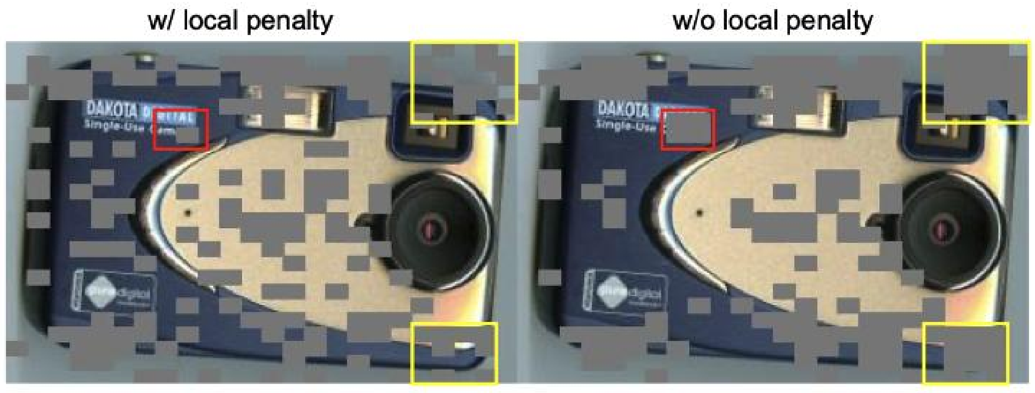
Stage	Method *FiCoCo-V*	SQA **68.37**	TextVQA **55.46**
Filter	w/o vision-aware redundancy	67.81	52.51
	w/o semantic-aware redundancy	64.67	48.74
	w/o local penalty	68.12	53.24
Correlate	fixed K=0	67.82	53.56
	fixed K=1	67.43	46.97
	fixed K=2	67.21	51.36
	convergent correlation	67.60	54.38
Compress	average compression	67.92	53.34

Stage	Method *FiCoCo-L*	SQA **69.46**	TextVQA **55.72**
Filter	w/o vision-aware redundancy	69.16	55.43
	w/o task-aware redundancy	68.22	55.64
	w/ local penalty	68.79	55.38
Correlate	w/o indirect correlation	68.89	54.78
	w/o direct correlation	68.45	55.45
	fixed K=0	68.96	50.33
	fixed K=1	68.57	50.11
	fixed K=2	68.32	50.18
	convergent correlation	67.80	54.89
Compress	average compression	68.32	54.66

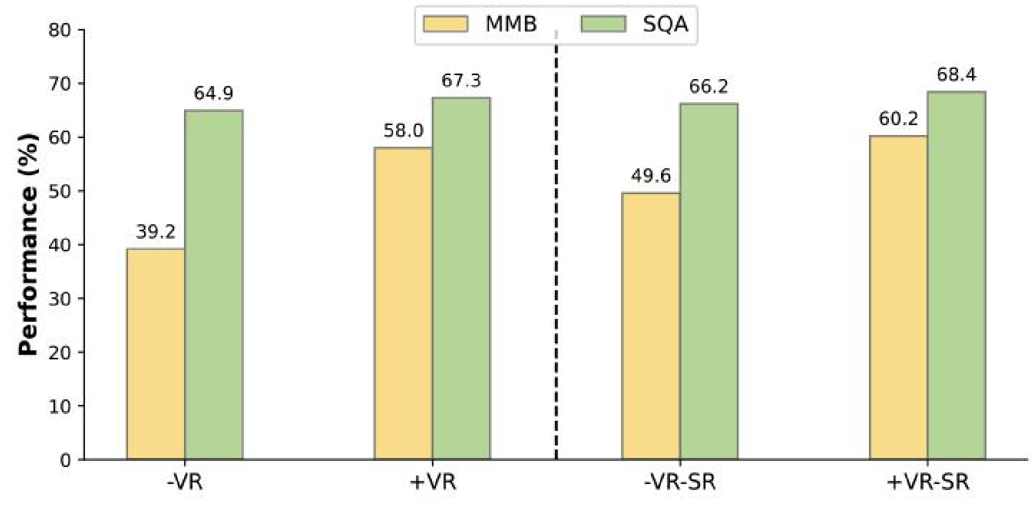
w/o [CLS] token Disscuss

Method	VQA^{T}	MMB	POPE	MM-Vet	Vizwiz	Avg (%)
			TFLOPs=8.5			
LLaVA-1.5	58.2	66.1	86.4	31.6	50.0	58.46
			TFLOPs=1.5$_{(\downarrow 82.4\%)}$			
$\mathbf{a}_i^{\mathrm{CLS}}$	55.5	60.2	79.8	26.8	52.4	54.94
$\mathbf{a}_i^{\mathrm{H}}$	54.2	59.6	81.4	25.9	49.8	54.18
$\mathbf{a}_i^{\mathrm{Eq}}(Quary)$	52.0	57.8	79.6	25.1	49.9	52.89
$\mathbf{a}_i^{\mathrm{Eq}}(Value)$	54.3	61.4	81.0	25.4	50.8	54.59
$\mathbf{a}_i^{\mathrm{Eq}}(Key)$	54.8	60.3	81.4	26.5	50.9	54.78

Visualization of local penalty



Vision-aware redundancy comparison


MMB
SQA
Performance (%)
-VR
+VR
-VR-SR
+VR-SR