Global Compression Commander: Plug-and-Play Inference Acceleration for High-Resolution Large Vision-Language Models

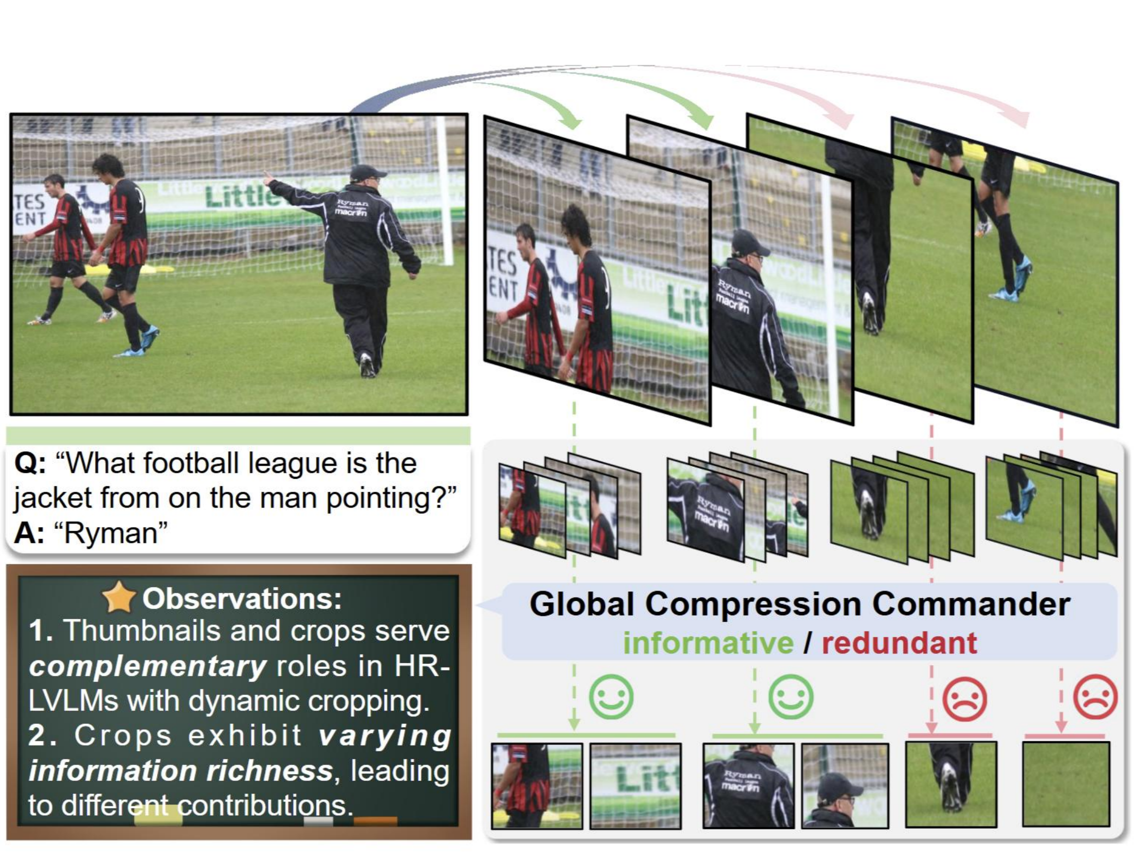
Xuyang Liu[1]*, Ziming Wang[2], Junjie Chen[1], Yuhang Han[3], Yingyao Wang[2], Jiale Yuan[2], Jun Song[2]✉, Siteng Huang[4], Honggang Chen[1]✉

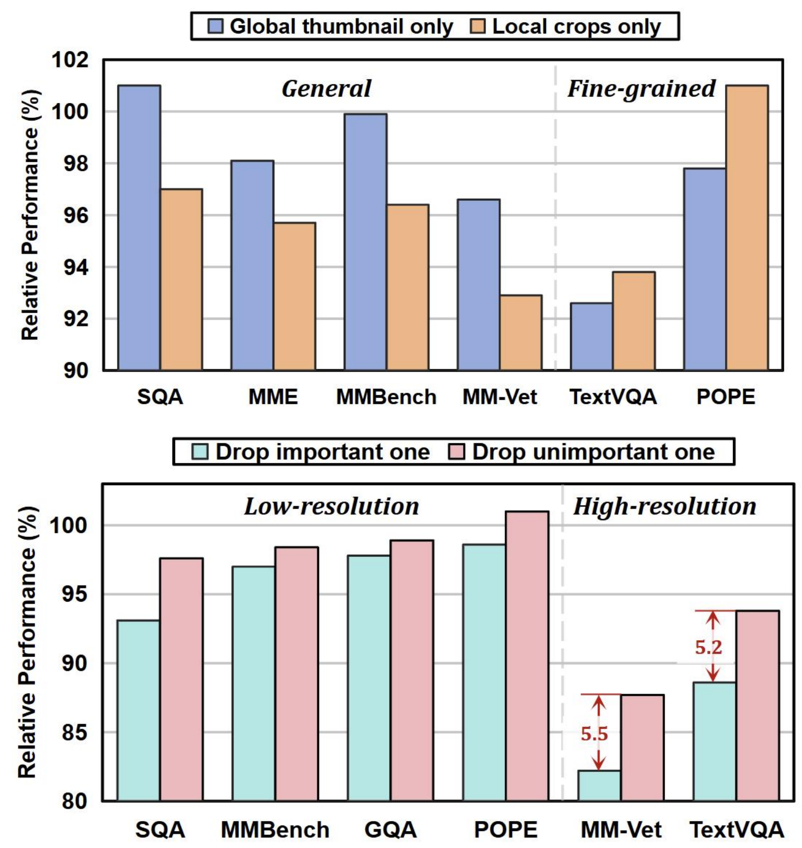
[1]Sichuan University, [2]Taobao & Tmall Group of Alibaba, [3]Westlake University, [4]Zhejiang University

*This work was done during an internship at Alibaba. ✉ Corresponding author: honggang_chen@scu.edu.cn

AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20–27, 2026 | SINGAPORE

Motivation and Research Status: The Characteristics of HR-LVLMs and the Limitations of Existing Methods

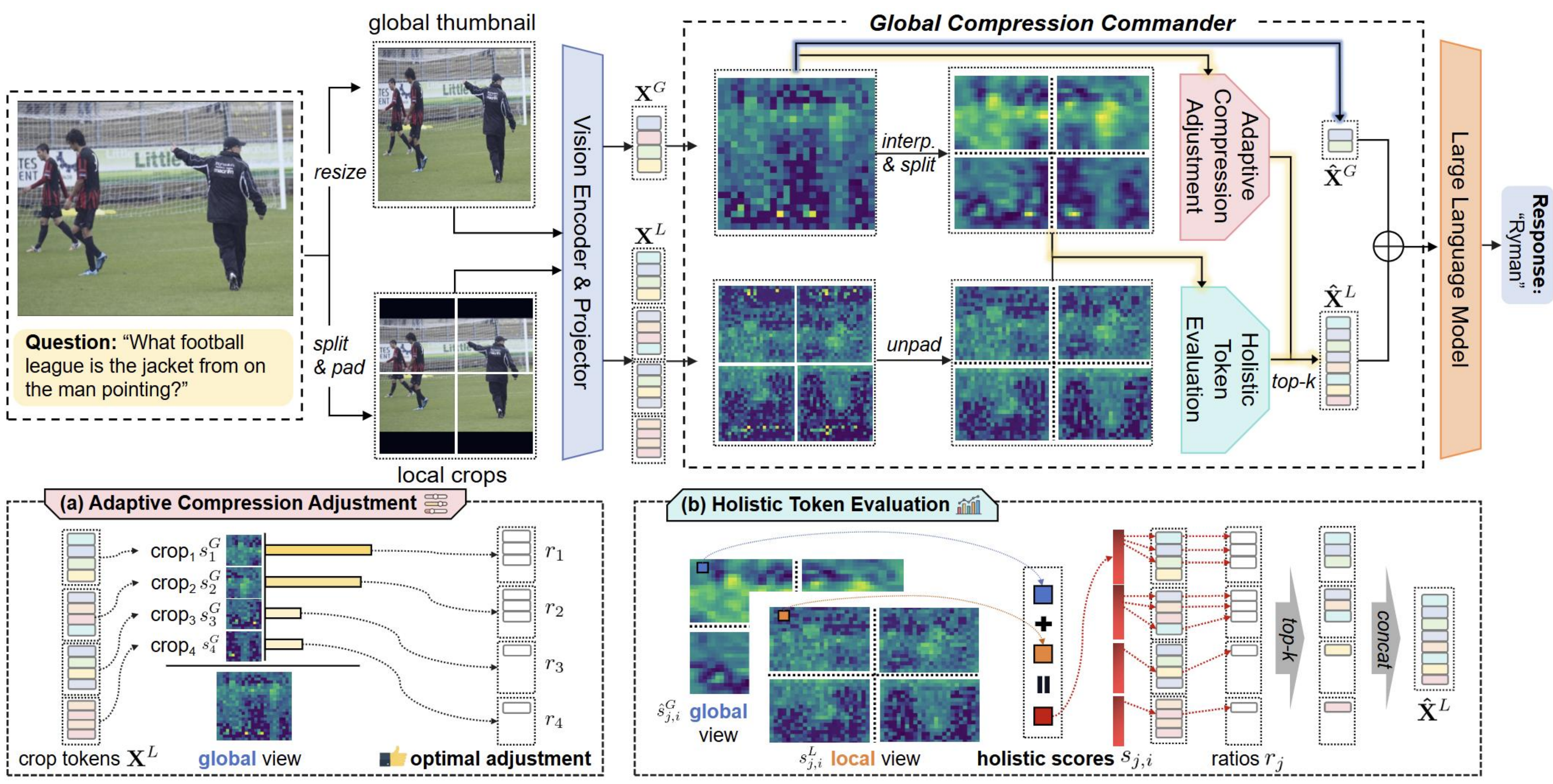


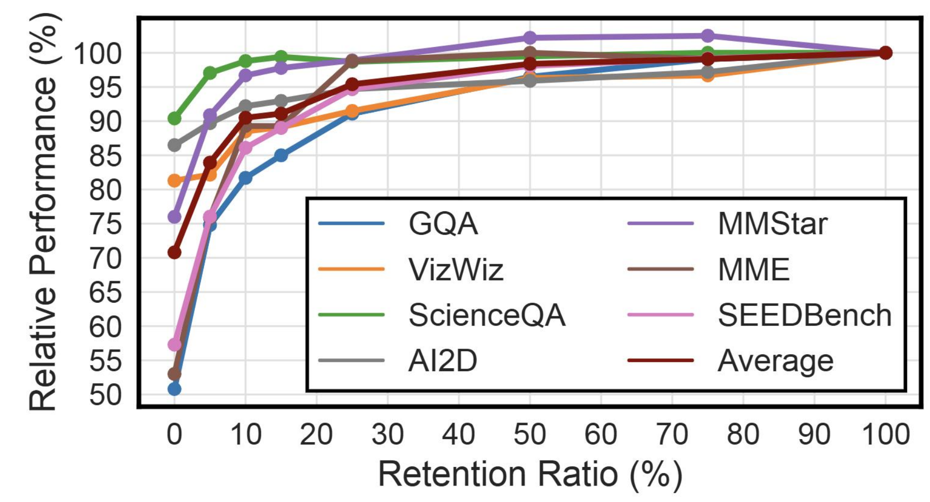


Takeaways: We derive *two obervations* for LVLMs with dynamic cropping: **(i)** Thumbnails and crops servecomplementary roles in HR-LVLMs with dynamic cropping. **(ii)** Crops exhibitvary information richness, leading to different contributions.

Contributions: (i) analyze dynamic-cropping HR-LVLMs, revealing global-context neglect, crop informativeness disparity, and content-agnostic positional bias; **(ii)** propose GlobalCom², a training-free plug-and-play global-to-local compressor; **(iii)** retain >90% performance while pruning 90% visual tokens.

Our Solution: Global Compression Commander (GlobalCom²) - "Global-to-Local" Guided Compression Philosophy





Performance: GlobalCom² maintains over **90%** performance while compressing **90%** visual tokens across multiple vision-language understanding benchmarks.

Efficiency: GlobalCom² cuts FLOPs to **9.1%** and peak GPU memory to **60%**, delivering a **1.8×** throughput gain.

Experimental Results: Optimal Performance-Efficiency Tradeoffs

Method	GQA	VizWiz	SQA	MMB	POPE	VQA^T	MME	MM-Vet	Average
Upper Bound, 2880 Tokens									
LLaVA-NeXT-7B	64.2	57.6	70.1	67.4	86.5	64.9	1519.0	43.9	100.0%
Ratio=50%, Retain up to 1440 Tokens									
FastV (ECCV24)	61.8	54.9	69.0	67.4	85.5	59.6	1490.3	37.6	95.5%
PDrop (CVPR25)	63.7	**57.9**	**69.2**	**67.7**	87.9	61.6	1499.6	37.5	97.4%
SparseVLM (ICML25)	63.7	57.2	68.3	67.6	87.9	60.5	1507.2	36.8	96.8%
FasterVLM (2024.12)	63.4	56.4	69.1	67.4	87.7	58.9	1533.3	39.6	97.3%
GlobalCom²	**63.9**	56.5	68.5	67.6	**88.1**	**62.3**	**1552.9**	**40.4**	**98.5%**
Ratio=25%, Retain up to 720 Tokens									
FastV (ECCV24)	60.4	54.2	**68.8**	65.6	83.1	58.4	1477.3	35.4	93.4%
PDrop (CVPR25)	60.3	**56.8**	68.5	65.6	85.5	59.8	1473.7	31.1	93.3%
SparseVLM (ICML25)	59.9	56.0	67.5	65.6	85.0	58.3	1465.9	38.5	94.6%
FasterVLM (2024.12)	61.3	55.4	67.1	**66.0**	87.2	58.8	1454.6	37.8	94.8%
GlobalCom²	**61.5**	55.7	68.1	65.9	**87.6**	**60.9**	**1493.5**	**40.7**	**96.7%**
Ratio=10%, Retain up to 288 Tokens									
FastV (ECCV24)	55.9	53.1	68.1	61.6	71.7	55.7	1282.9	27.2	85.4%
PDrop (CVPR25)	54.5	54.4	67.7	59.0	77.6	54.4	1262.1	24.0	84.3%
SparseVLM (ICML25)	56.3	52.1	68.5	60.0	80.1	53.9	1334.2	26.5	86.1%
PruMerge (ICCV25)	53.6	54.0	66.4	61.3	60.8	50.6	1149.3	25.5	80.6%
FasterVLM (2024.12)	56.9	52.6	66.5	61.6	83.6	56.5	1359.2	35.0	89.9%
GlobalCom²	**57.1**	**54.6**	**68.7**	**61.8**	**83.8**	**58.4**	**1365.5**	**36.4**	**91.6%**

Ablation Studies

Method	SQA	POPE	VQA^T	MME	MM-Vet	Avg.
Upper Bound, 2880 Tokens						
Vanilla	70.1	86.5	64.9	1519.0	43.9	100.0%
Ratio=25%, Retain up to 720 Tokens						
Uniform	67.1	87.2	60.1	1454.6	37.8	94.2%
$n_{\text{top-}k}$	67.4	87.3	59.8	1471.5	35.7	94.5%
Softmax (max)	67.3	87.2	60.3	1462.6	38.4	94.7%
Softmax (sum)	**67.6**	**87.4**	**60.6**	**1473.3**	**39.6**	**95.6%**

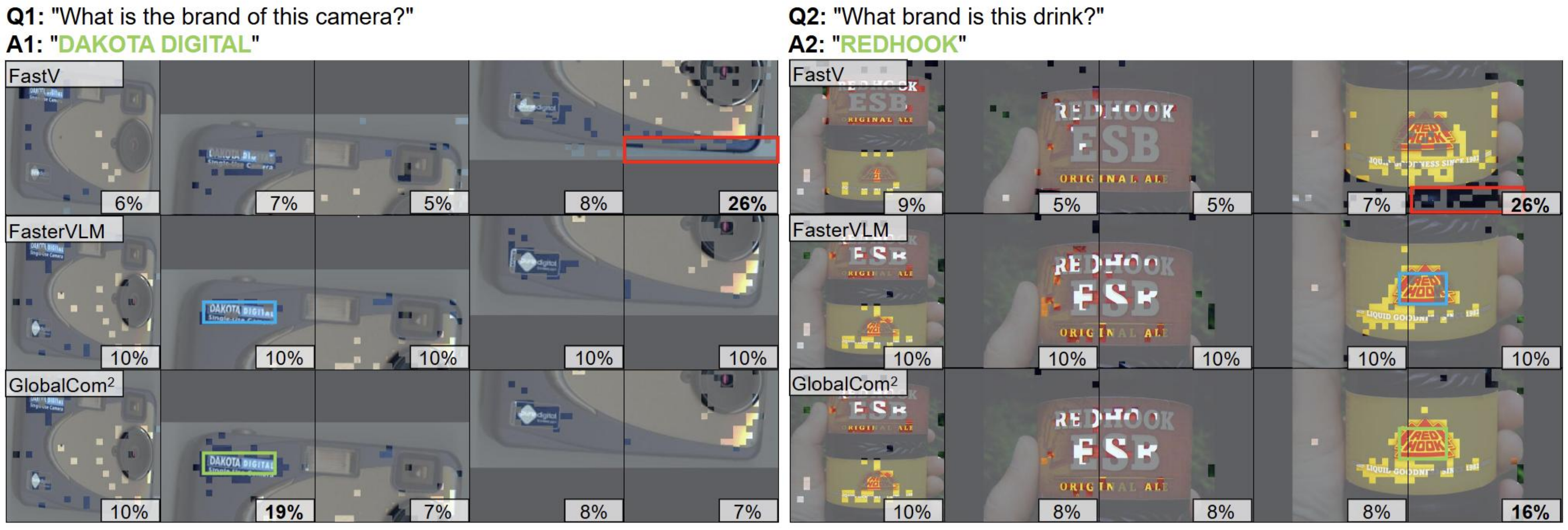
Method	SQA	POPE	VQA^T	MME	MM-Vet	Avg.
Upper Bound, 2880 Tokens						
Vanilla	70.1	86.5	64.9	1519.0	43.9	100.0%
Ratio=25%, Retain up to 720 Tokens						
Local only	67.6	87.4	60.6	1473.3	39.6	95.6%
Global only	67.9	86.4	60.2	1488.5	37.8	94.7%
Global and Local	**68.1**	**87.6**	**60.9**	**1493.5**	**40.7**	**96.7%**

Method	TFLOPs↓	Memory↓	Throughput↑	Performance↑
Upper Bound, 2880 Tokens				
Vanilla	41.7	23.0	3.8	100%
Ratio=10%, Retain up to 288 Tokens				
SparseVLM	5.4 (↓87.1%)	24.2 (↑5.2%)	5.9 (1.6×)	85.7%
FasterVLM	**3.8** (↓90.9%)	**13.6** (↓40.1%)	**6.7** (1.8×)	89.5%
GlobalCom²	**3.8** (↓90.9%)	13.9 (↓40.0%)	**6.7** (1.8×)	**90.8%**

Token Compression Visualizations: Better Information Preservation

Q1: "What is the brand of this camera?"
A1: "DAKOTA DIGITAL"

Q2: "What brand is this drink?"
A2: "REDHOOK"



Broader Applicability of GlobalCom²

Compatible with other methods: Plug-in boost for SparseVLM and FastV.