Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models

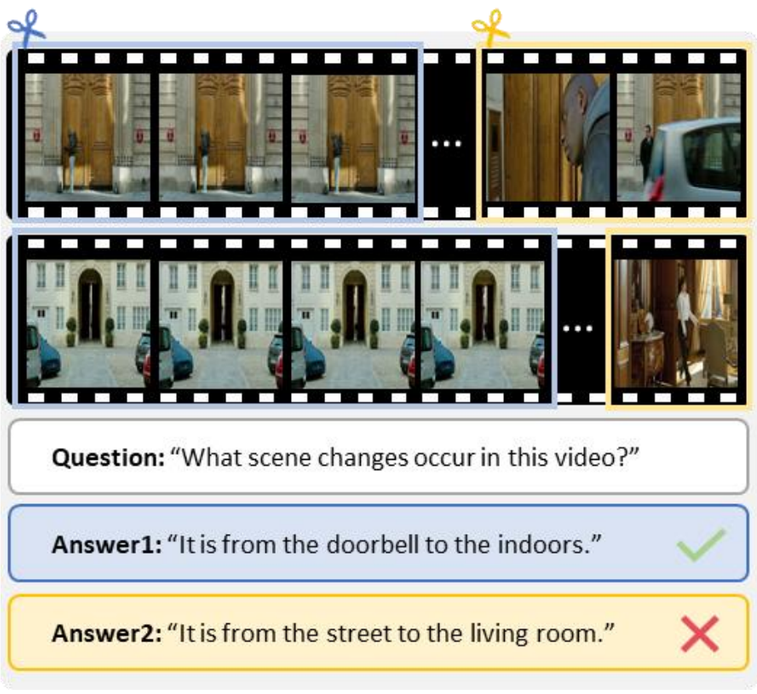
Xuyang Liu[1,2]*, Yiyu Wang[1]*, Junpeng Ma[3], Linfeng Zhang[1]✉

[1]EPIC Lab, Shanghai Jiao Tong University, [2]Sichuan University, [3]Fudan University

*Equal contribution. ✉ Corresponding author: zhanglinfeng@sjtu.edu.cn

EMNLP 2025 Suzhou, China 中国苏州

Motivation and Research Status



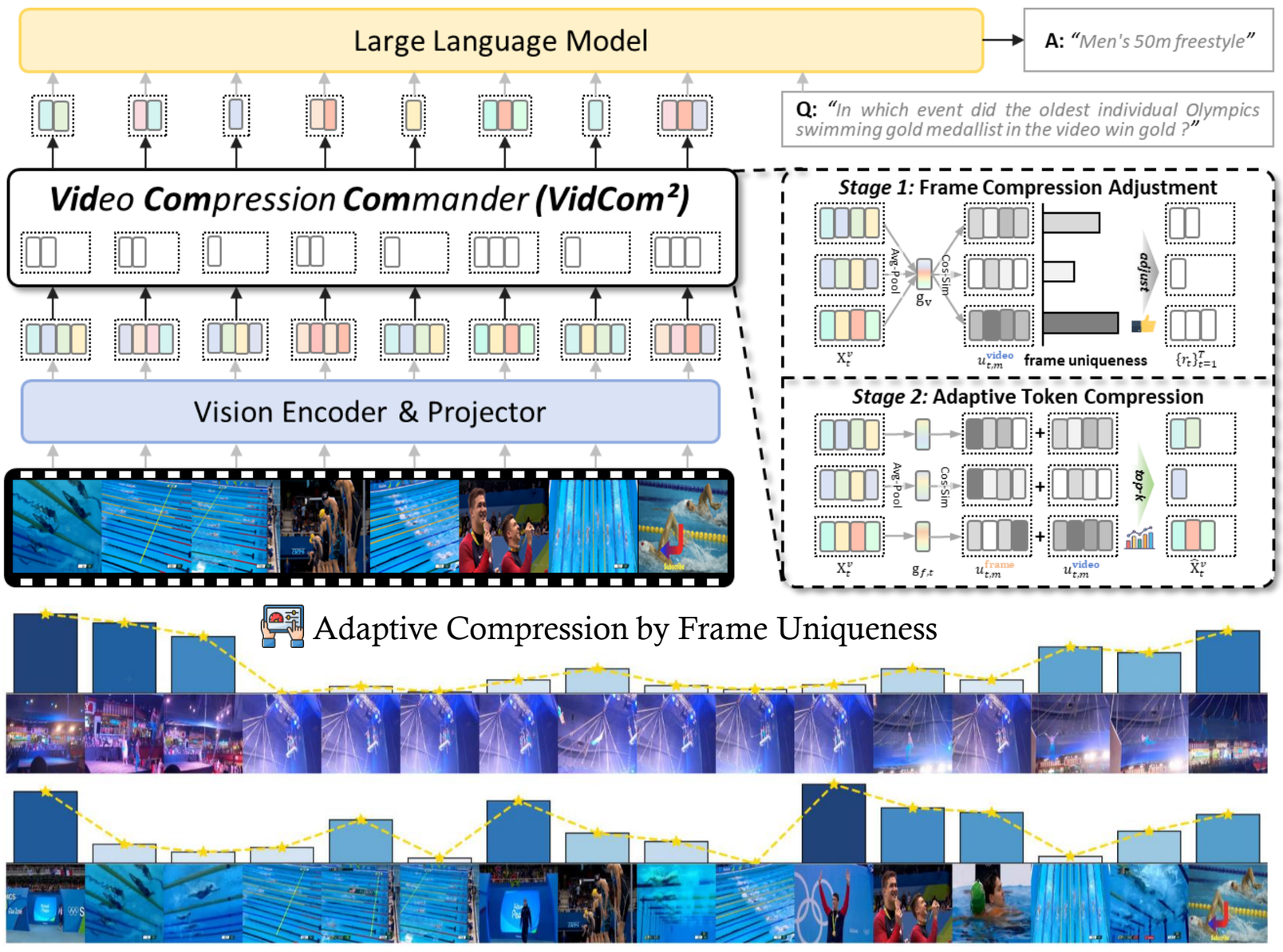
Methods	Pre-LLM	Intra-LLM	[CLS] Dependency	Video-Specific	Frame Uniqueness	Efficient Attention
FastV		✓				
PDrop		✓				
SparseVLM		✓				
MUSTDrop	✓	✓	✓			
FiCoCo	✓	✓	✓			
FasterVLM	✓		✓			✓
DyCoke	✓			✓		✓
VidCom²	✓			✓	✓	✓

😢 Current methods face ***two critical issues***:

- **Design Myopia:** ignoring each frame uniqueness, leading to over-compression of distinctive video information.
- **Implementation Constraints:** limited to the specific model 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.

Our Solution: Video Compression Commander (VidCom²)



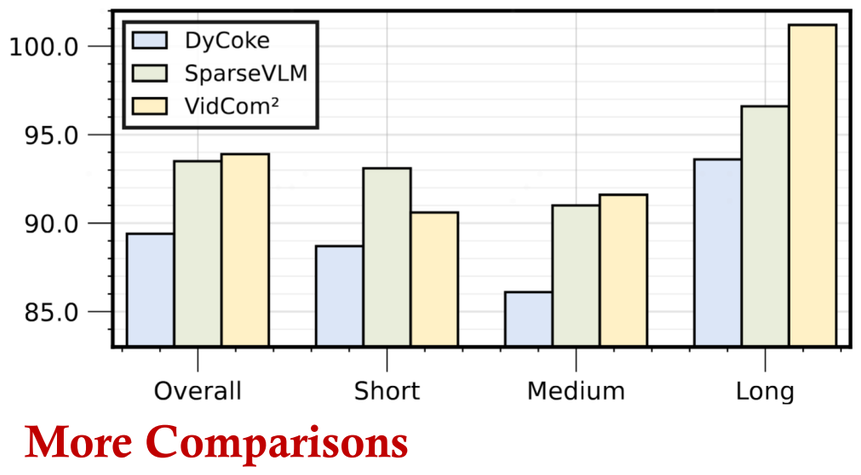
Performance and Efficiency on LLaVA-OneVision

- 💪 **Strong Performance:** Uses only 25% of tokens while maintaining **99.6%**.
- 🚀 **High Efficiency:** Cuts generation time by **70.8%** and overall latency by **43.0%**.

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
VidCom²	**57.2**	**54.9**	**62.5**	**58.6**	**69.8**	**56.4**	**49.4**	**99.6**
Retention Ratio=15%								
FastV[ECCV'24]	51.6	48.3	55.0	48.1	51.4	49.4	43.3	85.0
PDrop[CVPR'25]	53.2	47.6	54.7	50.1	58.7	48.7	45.0	87.4
SparseVLM[ICML'25]	52.9	49.7	57.4	53.4	61.0	52.1	47.0	91.2
VidCom²	**54.3**	**52.0**	**58.9**	**56.2**	**65.8**	**54.8**	**48.1**	**95.1**

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)
VidCom²	**180.7** (↓70.8%)	**574.7** (↓43.0%)	**18:46** (↓28.0%)	**16.0** (↓9.6%)	**0.88** (1.38×)	**57.2** (↑0.3)

Performance on Qwen2-VL



More Comparisons

Methods	EgoSchema	PerceptionTest
Upper Bound		
LLaVA-OV-7B	60.4 (100%)	57.1 (100%)
Retention Ratio=25%		
FastV[ECCV'24]	57.5 (95.2%)	55.4 (97.0%)
PDrop[CVPR'25]	58.0 (96.0%)	55.6 (97.4%)
DyCoke[CVPR'25]	59.5 (98.5%)	56.4 (98.8%)
VidCom²	**59.7 (98.8%)**	**56.7 (99.3%)**

Ablation Study and Analysis

Metrics	MLVU	VideoMME Overall	VideoMME Short	VideoMME Medium	VideoMME Long	Avg.
Vanilla	63.0	58.6	70.3	56.6	48.8	100.0
$s_{t,m}^{\text{frame}}$	59.5	54.0	62.2	54.2	45.3	94.1
$-s_{t,m}^{\text{frame}}$	61.9	57.9	68.8	56.9	48.1	98.8
$s_{t,m}^{\text{video}}$	58.9	53.3	61.7	52.1	46.1	93.2
$-s_{t,m}^{\text{video}}$	61.4	58.3	69.3	56.1	49.3	99.3
$u_{t,m}^{\text{frame}} + u_{t,m}^{\text{video}}$	**62.1**	**58.5**	**69.6**	**56.3**	**49.3**	**99.7**

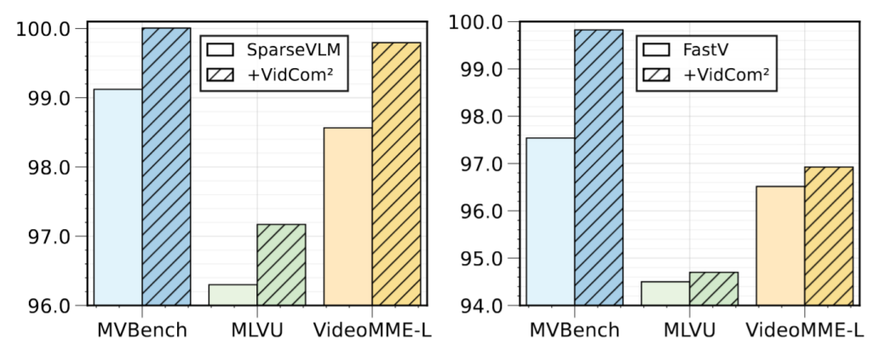
Metrics	MLVU	VideoMME Overall	VideoMME Short	VideoMME Medium	VideoMME Long	Avg.
Vanilla	63.0	58.6	70.3	56.6	48.8	100.0
Uniform	61.9	57.9	68.8	**56.9**	48.1	98.8
Frame Compression Adjustment						
$\max u_{t,m}^{\text{video}}$	62.1	58.1	68.4	56.7	49.3	99.4
$\overline{u_{t,m}^{\text{video}}}$	**62.3**	**58.2**	**69.1**	55.9	**49.6**	**99.6**

Size	MVBench	VideoMME Overall	VideoMME Short	VideoMME Medium	VideoMME Long	Avg.
Vanilla	56.9	58.6	70.3	56.6	48.8	100.0
4	56.8	57.9	69.6	55.6	48.7	99.1
8	56.8	58.3	69.8	56.4	48.6	99.6
16	57.2	58.5	**70.0**	**56.7**	48.9	100.1
32	**57.2**	**58.6**	69.8	56.4	**49.4**	**100.1**

Metrics	MVBench	VideoMME Overall	VideoMME Short	VideoMME Medium	VideoMME Long	Avg.
Vanilla	56.9	58.6	70.3	56.6	48.8	100.0
$u_{t,m}^{\text{frame}}$	56.8	57.9	68.8	56.9	48.1	98.8
$u_{t,m}^{\text{video}}$	56.8	58.3	69.3	56.1	49.3	99.3
Combination						
$u_{t,m}^{\text{frame}} + u_{t,m}^{\text{video}}$	57.2	58.6	69.8	56.4	49.4	100.3
$u_{t,m}^{\text{frame}} + 2u_{t,m}^{\text{video}}$	56.1	58.4	69.7	56.4	49.0	99.5
$2u_{t,m}^{\text{frame}} + u_{t,m}^{\text{video}}$	56.9	58.6	69.7	56.8	49.3	100.0

Broader Applicability of VidCom²

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



More Information

Paper | Code | WeChat