EMNLP 2025
Suzhou, China | 中国苏州

EPIC-LAB
SHANGHAI JIAOTONG UNIVERSITY

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

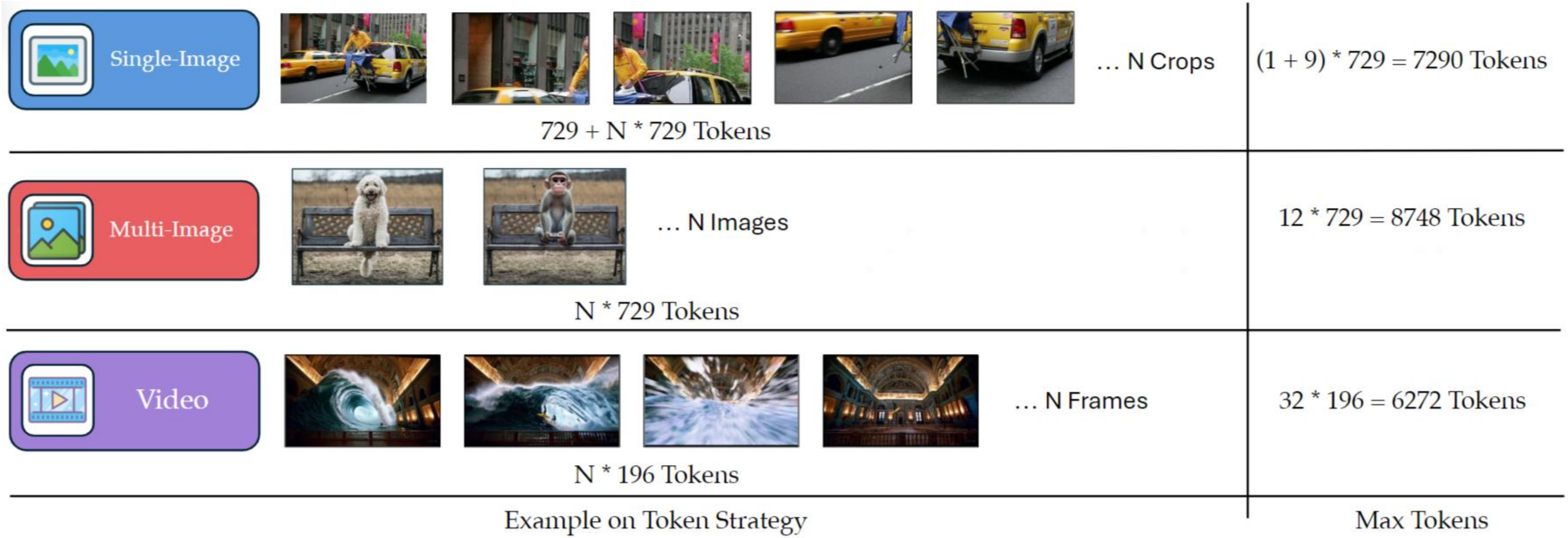
Xuyang Liu[1,2]*, Yiyu Wang[1]*, Junpeng Ma[3], Linfeng Zhang[1]✉

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

Paper: https://arxiv.org/pdf/2505.14454

Code: https://github.com/xuyang-liu16/VidCom2

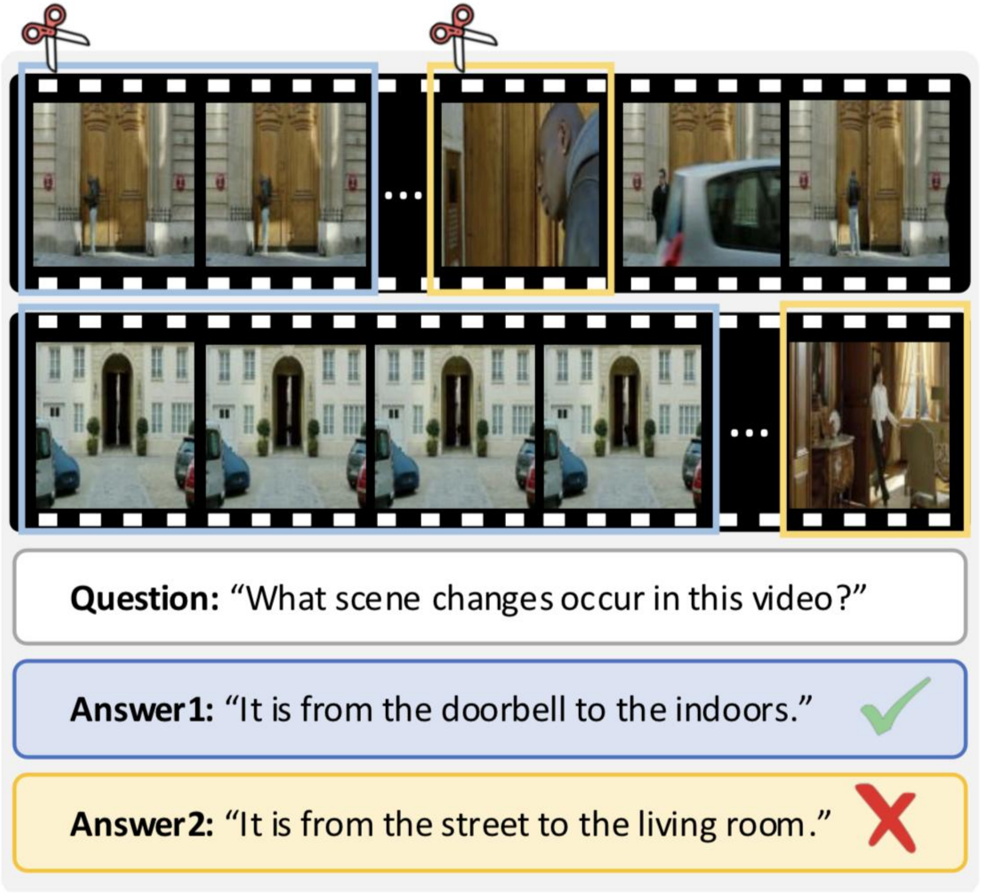
Research Background: Token Overhead in Visual Understanding



As model capabilities improve, the demand for high-resolution image and long-video understanding is growing, making the **token overhead issue increasingly pronounced** in visual understanding.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, YanweiLi, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. arXiv preprintarXiv:2408.03326.

Relevant Works: Token Compression for VideoLLMs



Methods	Pre-LLM	Intra-LLM	[CLS] Dependency	Video-Specific	Frame Uniqueness	Efficient Attention
FastV		✓				
PDrop		✓				
SparseVLM		✓				
MUSTDrop	✓	✓	✓			
FiCoCo	✓	✓	✓			
FasterVLM	✓		✓			✓
DyCoke	✓			✓		✓
VidCom^2	✓			✓	✓	✓

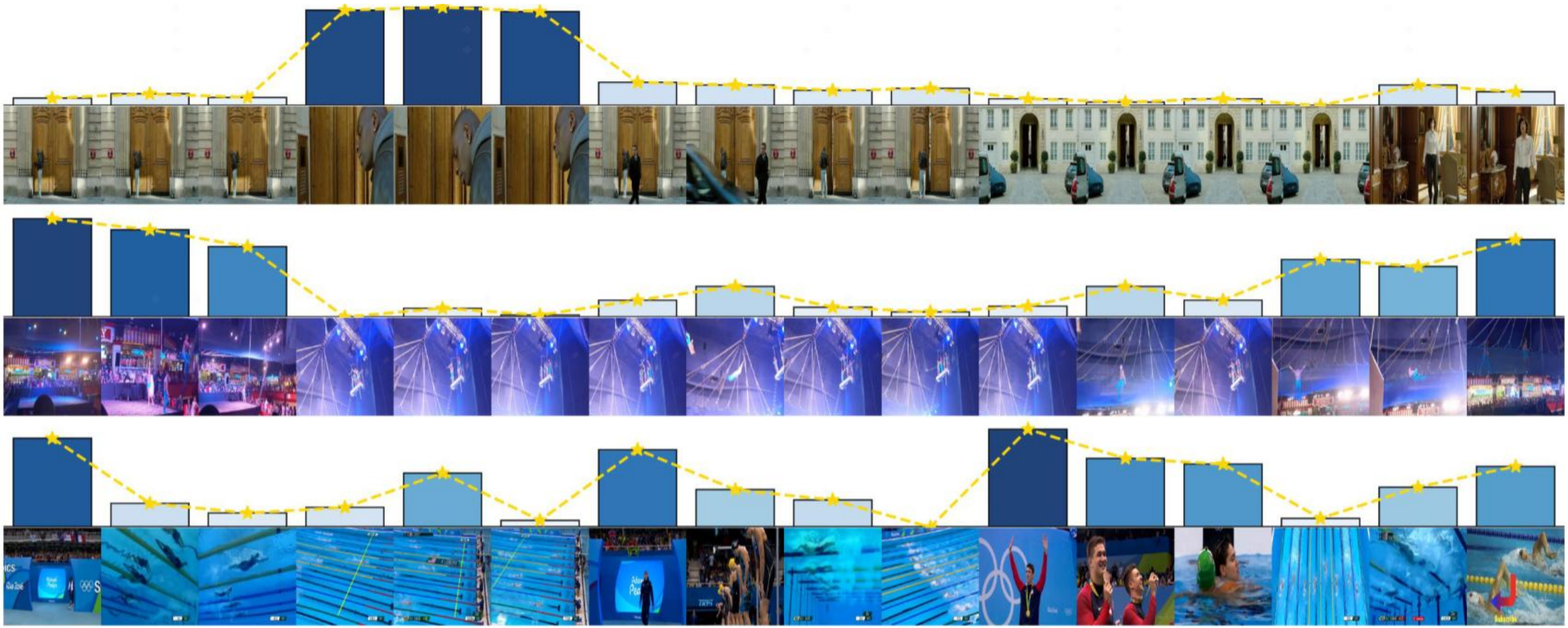
Current methods face ***two critical issues*** for VideoLLMs:

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

Xuyang Liu, Yiyu Wang, Junpeng Ma, Linfeng Zhang, "Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models". In *the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.

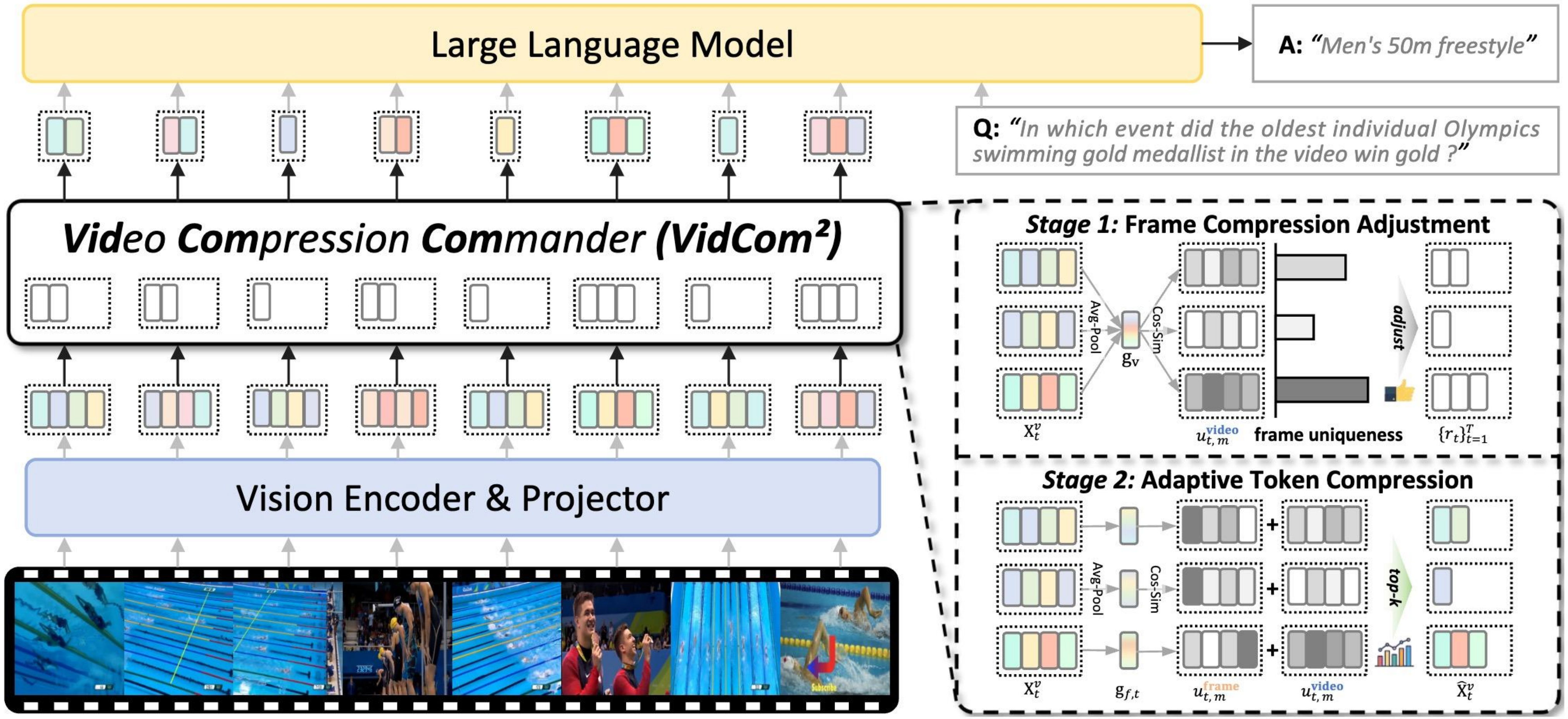
Key Highlights: Adaptive Compression by Frame Uniqueness



Core Argument: We suggest that visually distinctive frames throughout the video should retain more visual information, i.e., be allocated a larger visual token budget.

Xuyang Liu, Yiyu Wang, Junpeng Ma, Linfeng Zhang, "Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models". In *the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.

Our Solution: Video Compression Commander (VidCom²)



1.38× throughput
99.6% performance

optimal compression intensity for each frame

token compression for efficient inference

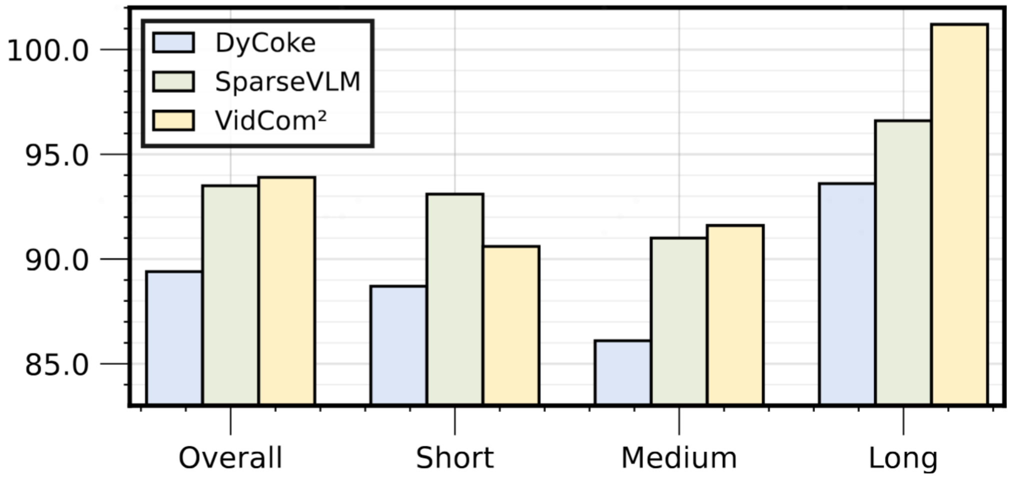
We present VidCom², a plug-and-play framework that dynamically compresses video tokens based on frame uniqueness, achieving state-of-the-art efficiency and performance across various VideoLLMs and benchmarks.

Xuyang Liu, Yiyu Wang, Junpeng Ma, Linfeng Zhang, "Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models". In *the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.

Experimental Results: State-of-The-Art Performance

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**
Upper Bound								
LLaVA-Video-7B	60.4	59.6	70.3	64.3	77.2	62.1	53.4	100.0
Retention Ratio=30%								
DyCoke [CVPR'25]	57.5	55.5	60.6	61.3	73.4	59.3	51.2	93.8
Retention Ratio=25%								
FastV [ECCV'24]	53.8	51.2	57.8	59.3	67.1	60.0	**50.8**	89.7
SparseVLM [ICML'25]	55.4	54.2	58.9	60.1	71.1	59.1	50.1	91.6
DyCoke [CVPR'25]	50.8	53.0	56.9	56.1	65.8	53.6	48.9	86.3
VidCom²	**57.0**	**55.5**	**59.0**	**61.7**	**73.0**	**61.7**	50.0	**93.6**
Retention Ratio=15%								
FastV [ECCV'24]	44.0	44.6	53.8	51.3	56.4	51.1	46.2	78.0
SparseVLM [ICML'25]	53.1	**52.7**	56.2	55.7	65.0	53.9	48.3	86.3
VidCom²	**53.3**	51.5	**56.8**	**58.3**	**68.0**	**57.3**	**49.7**	**88.5**

Performance on VideoMME with Qwen2-VL



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%)**

VidCom² achieves state-of-the-art 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". In *the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.

Experimental Results: State-of-The-Art Efficiency

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)
Retention Ratio=15%						
Random	130.3 (↓78.9%)	532.5 (↓47.2%)	18:02 (↓30.8%)	15.8 (↓10.7%)	0.92 (1.44×)	53.1 (↓3.8)
FastV [ECCV'24]	172.4 (↓72.1%)	599.3 (↓40.6%)	18:19 (↓29.7%)	24.6 (↑39.0%)	0.91 (1.42×)	51.6 (↓5.3)
PDrop [CVPR'25]	165.3 (↓73.3%)	552.6 (↓45.2%)	18:32 (↓28.9%)	24.5 (↑38.4%)	0.90 (1.41×)	53.2 (↓3.7)
SparseVLM [ICML'25]	370.4 (↓40.1%)	764.8 (↓24.2%)	24:09 (↓7.3%)	27.1 (↑53.1%)	0.69 (1.08×)	52.9 (↓4.0)
VidCom^2	**129.2** (↓79.1%)	**533.0** (↓47.1%)	**18:11** (↓30.2%)	**15.8** (↓10.7%)	**0.92** (1.44×)	**54.3** (↓2.6)

VidCom^2 achieves outstanding efficiency with a 70.8% reduction in LLM generation latency and 1.38× higher throughput, while remaining compatible with efficient attention operators.

Xuyang Liu, Yiyu Wang, Junpeng Ma, Linfeng Zhang, "Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models". In *the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.

Thanks!

Q & A