EarlyTom: Early Token Compression Completes Fast Video Understanding

1. Zhejiang University
2. Westlake University
3. Alibaba Cloud Computing

† Equal contribution
* Corresponding author

Hesong Wang[1, 2, 3†], Xin Jin[2†], Lu Lu[3*], Chenhaowen Li[3], Jian Chen[3], Qiang Liu[3], Huan Wang[2*]

paper

code

ENCODE LAB
Efficient Neural Computing and Design

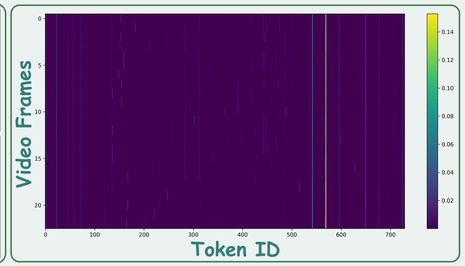
CVPR
JUNE 3-7, 2026

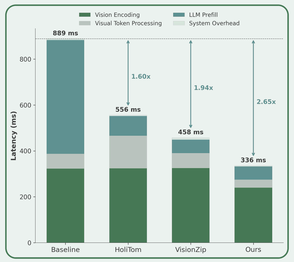
DENVER
COLORADO

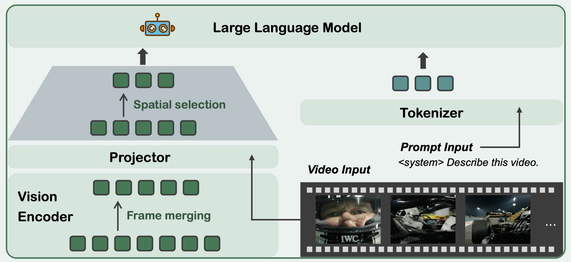
Introduction

1. Motivation

- **Massive visual tokens** make Video-LLMs inference costly and slow.
- Existing compression methods mainly reduce tokens **after vision encoding**, while the vision encoder still dominates TTFT.
- Naive attention-based Top-K selection can be biased by **video sink tokens**.
- How to design **early and robust token compression** for fast video understanding remains under-explored.



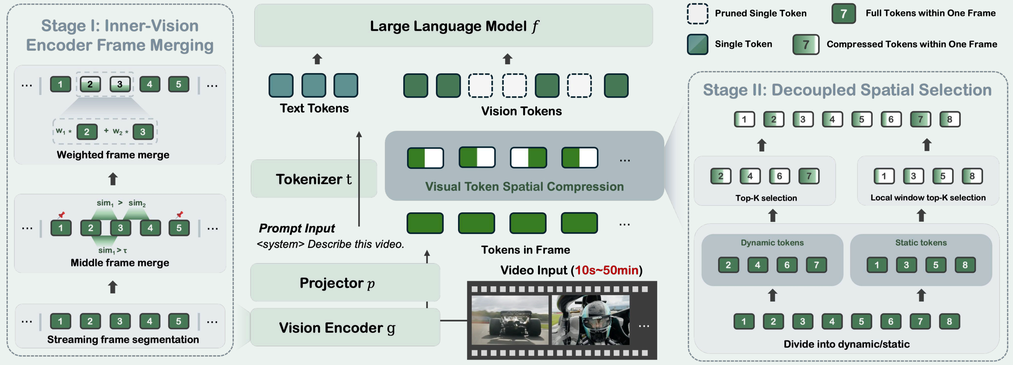




2. Key ideas

- **Early token compression**. Compress visual tokens during vision encoding rather than after it, achieving faster TTFT by shifting token compression to an earlier stage of the inference process.
- **Bias-agnostic token selection**. Use local-window top-K instead of global top-K to retain visual tokens, alleviating the bias introduced by vision sink tokens and maintaining a distribution closer to the original one.

Method



- **Inner-Encoder Frame Merging**: Segments frames by streaming cosine similarity (EMA-smoothed). Merges redundant middle frames with a local-optimal strategy and weighted fusion, compressing temporal redundancy before tokens enter the LLM.
- **Decoupled Spatial Selection**: Splits frames into dynamic (segment head/tail) and static (middle) sets. Dynamic frames → global top-K (motion-aware); Static frames → local-window top-K (distribution-preserving, sink-agnostic). CPU–GPU co-design offloads static selection for extra efficiency.

Main Results

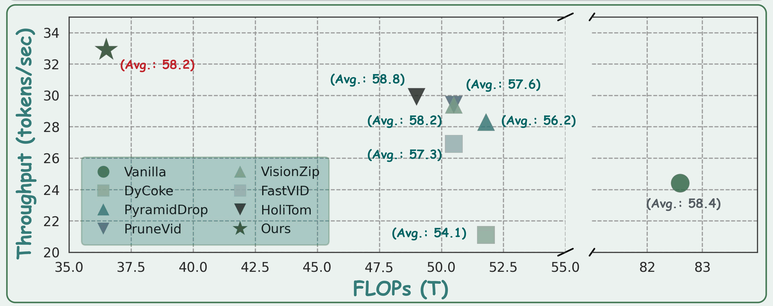
LLaVA-OneVision-7B

Method		Before LLM Retained Ratio	Prefilling FLOPs (T) ↓	FLOPs Ratio ↓	TTFT (ms) ↓	Throughput (tokens/s) ↑	MVBench ↑	EgoSchema ↑	LongVideo Bench ↑	VideoMME ↑	Avg. ↑ Score	Avg. ↑ %
LLaVA-OV-7B		100%	82.6	100%	889.9	24.4	58.3	60.4	56.4	58.6	58.4	100
FastV	*ECCV'24*	100%	51.1	61.9%	820.0	28.4	55.9	57.5	56.7	56.1	56.5	96.7
PyramidDrop	*CVPR'25*	100%	51.8	62.7%	813.4	28.3	56.1	58.0	54.1	56.4	56.2	96.2
DyCoke	*CVPR'25*	25%	50.5	61.1%	905.6	21.1	53.1	59.5	49.5	54.3	54.1	92.6
VisionZip	*CVPR'25*	25%	50.5	61.1%	516.6	29.4	57.9	60.3	56.5	58.2	58.2	99.7
PruneVid	*ACL'25*	25%	50.5	61.1%	703.6	29.4	57.4	59.9	55.7	57.4	57.6	98.6
FastVID	*NeurIPS'25*	25%	50.5	61.1%	581.6	26.9	56.5	58.2	56.3	58.0	57.3	98.1
HoliTom	*NeurIPS'25*	25%	49.0	59.3%	661.3	29.9	58.4	61.2	56.7	58.9	58.8	100.7
EarlyTom		25%	**36.5**	**44.2%**	**426.3**	**32.9**	57.4	60.5	56.3	58.5	58.2	99.7
VisionZip	*CVPR'25*	20%	48.7	58.9%	495.0	29.8	57.7	59.8	55.2	57.9	57.7	98.8
PruneVid	*ACL'25*	20%	49.0	59.3%	662.1	29.5	57.2	59.7	54.7	56.9	57.1	97.8
FastVID	*NeurIPS'25*	20%	48.7	58.9%	546.6	27.6	56.3	57.9	57.1	57.9	57.3	98.1
HoliTom	*NeurIPS'25*	20%	47.5	57.5%	622.3	30.0	58.7	61.0	57.1	58.6	58.8	100.7
EarlyTom		20%	**35.1**	**42.4%**	**415.3**	**33.4**	57.8	60.6	55.6	58.0	58.1	99.3
VisionZip	*CVPR'25*	15%	46.9	56.8%	475.9	**32.1**	56.5	59.8	54.4	56.1	56.7	97.1
PruneVid	*ACL'25*	15%	47.5	57.5%	574.1	27.1	56.8	59.7	55.4	56.6	57.1	97.8
FastVID	*NeurIPS'25*	15%	46.9	56.8%	530.8	28.7	56.0	57.4	56.2	57.7	56.8	97.3
HoliTom	*NeurIPS'25*	15%	46.0	55.7%	572.7	27.5	58.1	61.2	56.4	57.3	58.2	99.7
EarlyTom		15%	**33.6**	**40.7%**	**390.6**	30.4	57.5	60.2	54.4	56.9	57.3	98.1
VisionZip	*CVPR'25*	10%	45.2	54.7%	458.5	28.5	53.5	58.0	49.3	53.4	53.5	91.6
PruneVid	*ACL'25*	10%	45.9	55.6%	592.2	28.6	56.2	59.8	54.5	56.0	56.6	96.9
FastVID	*NeurIPS'25*	10%	45.2	54.7%	502.1	28.3	55.9	56.5	56.3	57.3	56.5	96.7
HoliTom	*NeurIPS'25*	10%	44.6	54.0%	556.6	29.0	57.3	61.2	56.3	56.8	57.9	99.1
EarlyTom		10%	**32.2**	**39.0%**	**336.2**	**31.6**	56.5	60.1	52.4	55.8	56.2	96.2

LLaVA-Video-7B

Model	Method		Before LLM Retained Ratio	Prefilling FLOPs (T) ↓	FLOPs Ratio ↓	TTFT (ms) ↓	Throughput (tokens/s) ↑	MVBench ↑	EgoSchema ↑	LongVideo Bench ↑	VideoMME ↑	Avg. ↑ Score	Avg. ↑ %
LLaVA-Video-7B	Vanilla		100%	246.2	100%	6429.3	8.1	60.4	57.2	58.9	64.3	60.2	100
	FastV	*ECCV'24*	100%	158.2	64.3%	3494.3	10.0	54.3	54.1	55.0	58.8	55.6	92.4
	PyramidDrop	*CVPR'25*	100%	159.4	64.7%	3494.8	10.1	55.9	54.3	54.7	61.9	56.7	94.2
	VisionZip	*CVPR'25*	15%	159.4	64.7%	3241.4	14.2	56.7	54.7	54.7	60.7	56.7	94.2
	HoliTom	*NeurIPS'25*	15%	156.6	63.6%	1669.5	**17.1**	57.7	54.8	56.2	62.1	57.7	95.8
	EarlyTom		15%	**86.4**	**35.1%**	**947.4**	16.7	55.8	54.7	53.9	61.3	56.4	93.7

Throughput vs. FLOPs



- 2.3× to 2.8× FLOPs reduction
- 1.2× to 2.0× throughput improvement
- 2.1× to 2.6× TTFT reduction
- Retain 96% of original accuracy

```
@inproceedings{
  wang2025earlytom,
  title = {EarlyTom: Early Token Compression Completes Fast Video Understanding},
  author = {Wang, Hesong and Jin, Xin and Lu, Lu and Chenhaowen Li, Jian Chen and Qiang Liu and Wang, Huan},
  year = {2026},
  booktitle={CVPR},
}
```