DyCoke : Dynamic Compression of Tokens for Fast Video Large Language Models

Keda Tao Can Qin Haoxuan You Yang Sui Huan Wang*

CVPR Nashville JUNE 11-15, 2025 | Paper | Code

Introduction

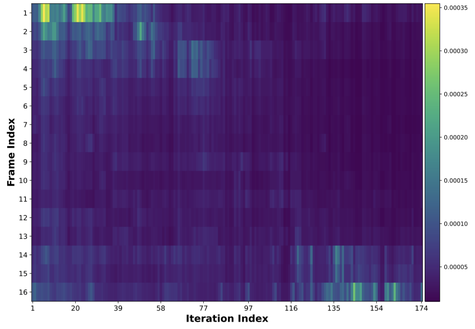
1. Motivation

- The development of multi-modal language models is limited by model efficiency.
- The **tens of thousands of** visual tokens brought by the video input greatly increase **the inference time and memory requirements.**
- There is a lot of redundancy in the visual tokens corresponding to the video input.

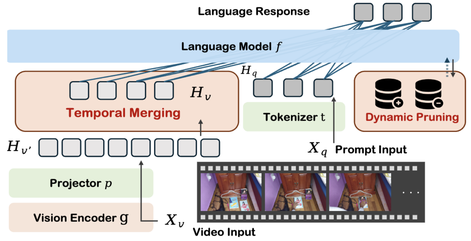
2. Why do we need dynamic pruning?

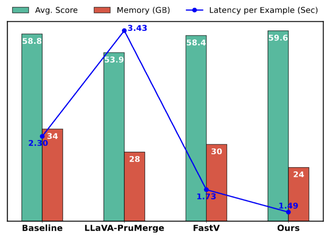
💥 Some video frames become **less important** as the decoding proceeds, while others may become **more important**.

🔥 Unlike image input, one-time pruning at the prefilling stage is prone to falsely pruning important information of the video.



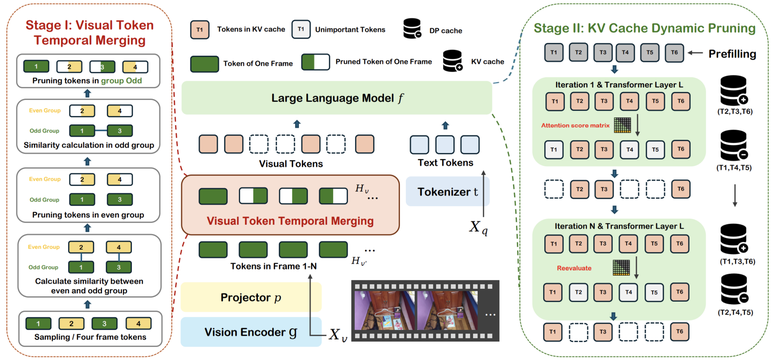
3. Key Ideas





- **Two-stage pruning**, the first stage performs preliminary temporal token merging (TTM), and the second stage performs dynamic pruning of KV cache.
- **Dynamic** pruning stores the pruned tokens for later reuse, thereby minimizing information loss.
- **Plug and play, no training required.**

Method



1. TTM Block

- Redundancy is assessed by computing the similarity of visual tokens in the same position between consecutive frames.
- The tokens are then partitioned into **two groups** for pruning.

2. KV Cache Dynamic Pruning

- In the decoding stage, a small number of tokens are selectively selected to participate in the attention calculation

Main Results

Method	Pruning Settings			ActNet-QA		NextQA	PercepTest	VideoDC	Videomme	
	Retained Ratio (Final)	FLOPs (T)	FLOPs Ratio	Acc.	Sco.	mc	val	test	wo	w-subs
LLaVA-OV-0.5B										
Full Tokens	100%	3.4	100%	47.93	2.66	57.2	49.1	2.86	44.1	43.5
FastV	35%	1.4	41%	46.74	2.58	56.5	49.1	2.36	42.0	41.4
PruMerge	55%	1.5	44%	41.68	2.35	54.0	47.5	2.10	38.8	39.3
Ours (K=0.3, L=3, P=0.7)	23.25%	2.4	70%	**47.90**	**2.65**	57.2	48.9	**2.63**	**45.4**	**43.8**
Ours (K=0.5, L=3, P=0.7)	18.75%	1.8	53%	47.80	**2.65**	57.2	**49.5**	2.62	45.1	43.4
Ours (K=0.7, L=3, P=0.7)	14.25%	1.2	35%	47.70	**2.65**	**57.7**	**49.5**	2.56	45.2	43.3
LLaVA-OV-7B										
Full Tokens	100%	41.4	100%	51.93	2.86	79.4	57.1	3.30	58.5	61.3
FastV	35%	17.9	43%	50.93	2.80	78.2	56.7	3.09	57.3	60.5
PruMerge	55%	21.1	51%	50.45	2.78	76.0	54.3	2.88	52.9	57.0
Ours (K=0.3, L=3, P=0.7)	23.25%	30.8	75%	51.80	2.85	**79.1**	57.2	3.19	58.8	61.0
Ours (K=0.5, L=3, P=0.7)	18.75%	24.1	59%	**52.08**	**2.88**	78.5	**57.6**	**3.29**	**59.5**	**61.4**
Ours (K=0.7, L=3, P=0.7)	14.25%	17.9	43%	51.80	2.85	78.2	**57.6**	3.20	58.3	60.7
LLaVA-OV-72B										
Full Tokens	100%	436.1	100%	52.96	2.92	80.2	66.9	3.34	66.2	69.5
FastV	45%	202.4	46%	52.63	2.82	77.2	58.5	3.01	62.1	66.7
PruMerge	55%	229.8	53%	50.91	2.80	75.2	55.6	2.81	61.9	63.8
Ours (K=0.5, L=3, P=0.7)	18.75%	262.5	60%	**52.81**	**2.92**	**79.1**	**60.2**	**3.35**	**66.3**	**69.7**
Ours (K=0.7, L=3, P=0.7)	14.25%	195.1	44%	52.38	2.88	78.8	59.6	3.27	64.9	68.7

More Results

1. Evaluation on MVBench Dataset

Method	FR	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI	*Avg.*
LLaVA-OV-7B																						
Full Tokens	100%	70.7	71.5	84.6	45.0	79.0	57.1	81.0	38.0	24.5	46.0	91.5	44.0	46.5	72.0	51.0	51.0	68.0	36.0	71.5	51.0	58.0
PruMerge	51%	63.3	63.0	84.6	39.5	76.5	51.0	61.0	36.0	**32.5**	**49.0**	88.0	41.0	39.5	62.5	46.5	45.5	54.0	35.5	67.5	38.5	52.6
FastV	43%	**73.5**	73.5	76.9	44.0	76.5	56.1	76.0	**42.5**	19.5	40.0	92.0	43.5	43.0	68.5	48.0	**50.0**	67.0	33.5	69.0	43.5	56.1
Ours (K = 0.5)	59%	69.7	73.5	84.6	**47.5**	79.0	**57.1**	77.5	39.0	22.5	47.0	**93.0**	44.0	**45.0**	**74.0**	**50.5**	**50.0**	66.0	**36.5**	**71.5**	**51.0**	**58.0**
Ours (K = 0.7)	43%	67.6	**74.0**	**92.3**	47.0	**80.0**	55.6	**78.0**	39.0	24.0	45.0	**93.0**	**45.0**	**45.5**	71.0	50.0	49.0	**67.5**	34.5	70.0	49.0	57.5
LLaVA-OV-0.5B																						
Full Tokens	100%	53.7	59.5	30.8	37.5	60.5	46.5	68.5	35.0	21.0	32.0	87.0	44.5	29.5	54.5	37.0	45.0	46.0	29.0	70.0	39.5	47.1
PruMerge	46%	37.8	48.5	23.1	30.5	53.5	45.5	45.5	31.0	20.0	**39.5**	83.0	38.5	**30.5**	47.0	35.5	43.5	38.0	28.0	69.0	**44.0**	42.5
FastV	46%	52.1	**60.5**	**38.5**	35.0	**62.0**	46.0	63.0	35.5	**23.5**	28.5	84.5	42.0	29.0	50.5	37.0	41.0	**47.5**	28.5	67.5	**44.0**	46.1
Ours (K = 0.5)	60%	53.7	59.5	30.8	**39.0**	60.0	46.0	**68.5**	32.5	21.0	30.0	87.0	42.0	29.0	**55.5**	**39.0**	**44.0**	46.0	28.5	**70.0**	42.5	47.0
Ours (K = 0.7)	44%	**54.8**	59.5	30.8	**39.0**	61.0	**47.5**	68.0	34.0	21.5	30.0	**88.0**	**44.5**	29.0	54.5	38.0	41.0	45.5	**29.5**	69.0	41.5	**47.1**

2. Showcases of DyCoke

Challenging Video Understanding

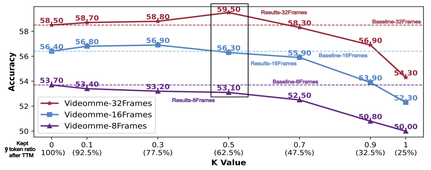
User: *What color is the object that is stationary? | What direction is the yellow sqhere moving in?*

LLaVA-OV: The object that is stationary is **gray.** ☹ | From the left to the right side of the frame. ☹
LLaVA-OV w/ FastV: The object that is stationary is blue. ☹ | Towards the left side of the frame. ☹
LLaVA-OV w/ DyCoke: The object that is stationary is **gray.** ☹ | From the left to the right side of the frame. ☹

User: *The person uses multiple similar objects to play an occlusion game. Where is the hidden object at the end of the game from the person's point of view?*

LLaVA-OV: The hidden object is under the middle purple cup. ☹
LLaVA-OV w/ FastV: The hidden object is under the middle purple cup. ☹
LLaVA-OV w/ DyCoke: The hidden object is under the **third purple cup from the left.** ☹

3. Additional Experiments



- There is almost **no performance loss.**
- **1.5×** inference speedup, **1.4×** memory reduction.

Method	Total Latency ↓	GPU Mem. ↓	Accuracy ↑	Latency per Example ↓
LLaVA-OV-7B				
Full Tokens	1:19:27	27G	57.58 (±0.11)	1.19s (1.00×)
PruMerge	2:02:17	20G	52.59 (±0.08)	1.83s (0.65×)
FastV	1:03:50	24G	56.05 (±0.13)	0.96s (1.24×)
Ours (K = 0.5)	59:02	21G	**57.88** (±0.10)	0.88s (1.35×)
Ours (K = 0.7)	**57:13**	**19G**	57.45 (±0.18)	**0.85s (1.40×)**
LLaVA-OV-0.5B				
Full Tokens	34:05	19G	47.09 (±0.09)	0.56s (1.00×)
PruMerge	1:35:08	8.5G	42.53 (±0.30)	1.42s (0.41×)
FastV	32:30	10G	46.14 (±0.03)	0.49s (1.14×)
Ours (K = 0.5)	32:17	8.9G	46.76 (±0.13)	0.48s (1.16×)
Ours (K = 0.7)	**31:45**	**7.4G**	**47.09** (±0.11)	**0.47s (1.19×)**

(a) The number of video input frames is 16

Method	Total Latency ↓	GPU Mem. ↓	Accuracy ↑	Latency per Example ↓
LLaVA-OV-7B				
Full Tokens	2:33:30	34G	58.81 (±0.14)	2.30s (1.00×)
PruMerge	3:48:59	28G	53.93 (±0.06)	3.43s (0.64×)
FastV	1:55:20	30G	58.36 (±0.09)	1.73s (1.32×)
Ours (K = 0.5)	1:53:17	28G	59.09 (±0.23)	1.69s (1.36×)
Ours (K = 0.7)	**1:39:49**	**24G**	**59.56** (±0.19)	**1.49s (1.54×)**
LLaVA-OV-0.5B				
Full Tokens	1:19:06	21G	48.23 (±0.15)	1.18s (1.00×)
PruMerge	3:06:06	13G	42.55 (±0.18)	2.79s (0.42×)
FastV	1:13:25	15G	47.04 (±0.05)	1.10s (1.07×)
Ours (K = 0.5)	1:10:32	12G	**48.13** (±0.09)	1.06s (1.11×)
Ours (K = 0.7)	**1:02:15**	**10G**	47.77 (±0.10)	**0.93s (1.27×)**

(b) The number of video input frames is 32