github repo

FlexSelect: Flexible Token Selection for Efficient Long Video Understanding

Yunzhu Zhang*, Yu Lu*, Tianyi Wang, Fengyun Rao, Yi Yang, Linchao Zhu†

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

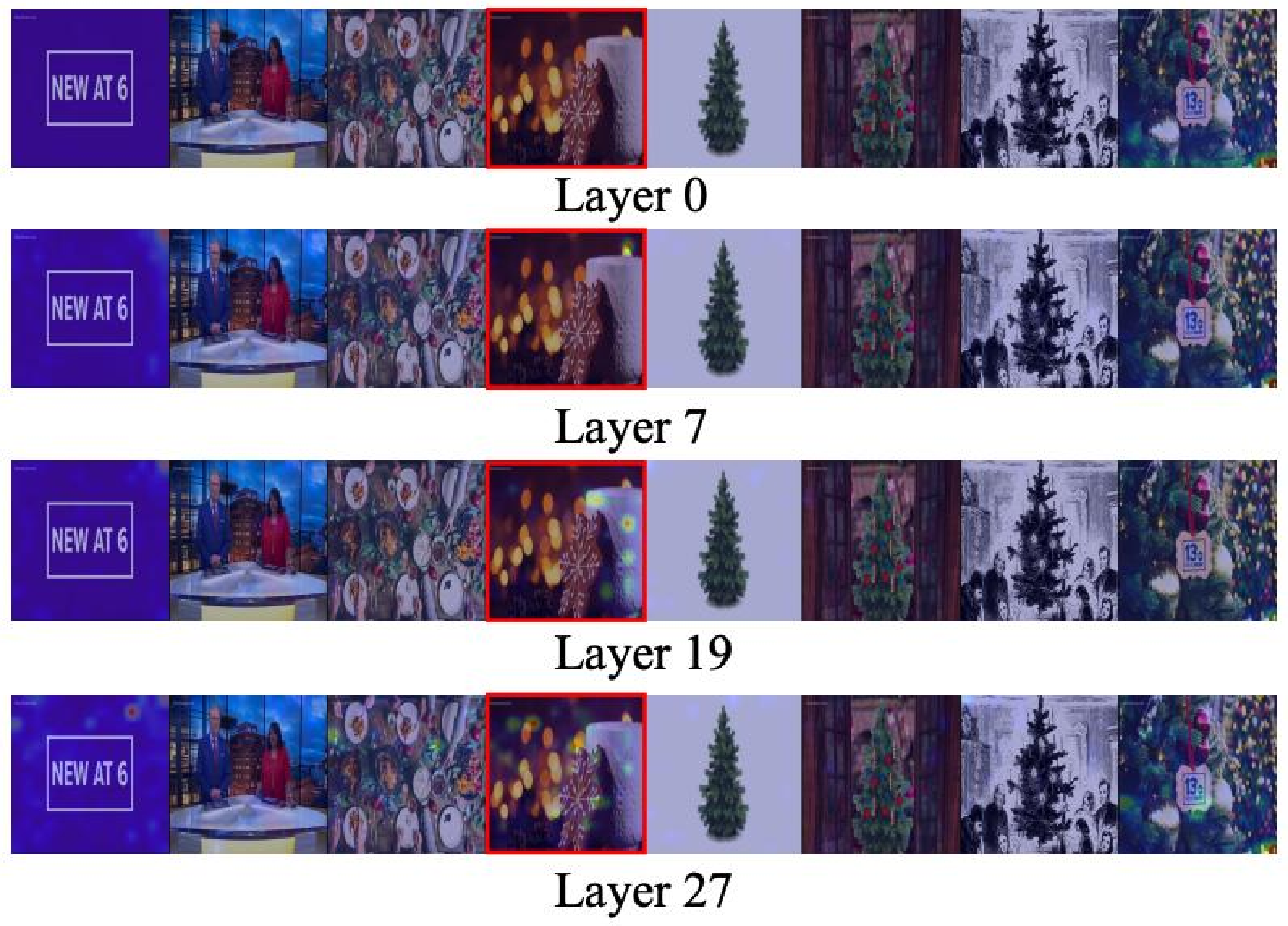
➢ Challenges of Video-LLM in Processing Long Videos

- ◆ Millions of tokens exceed the LLM's context window, leading to degraded model performance and significant computational costs.
- ◆ The user's question only relates to a few clips, while the long video consists of irrelevant and repeated content.

➢ Motivation: select the query-related visual tokens

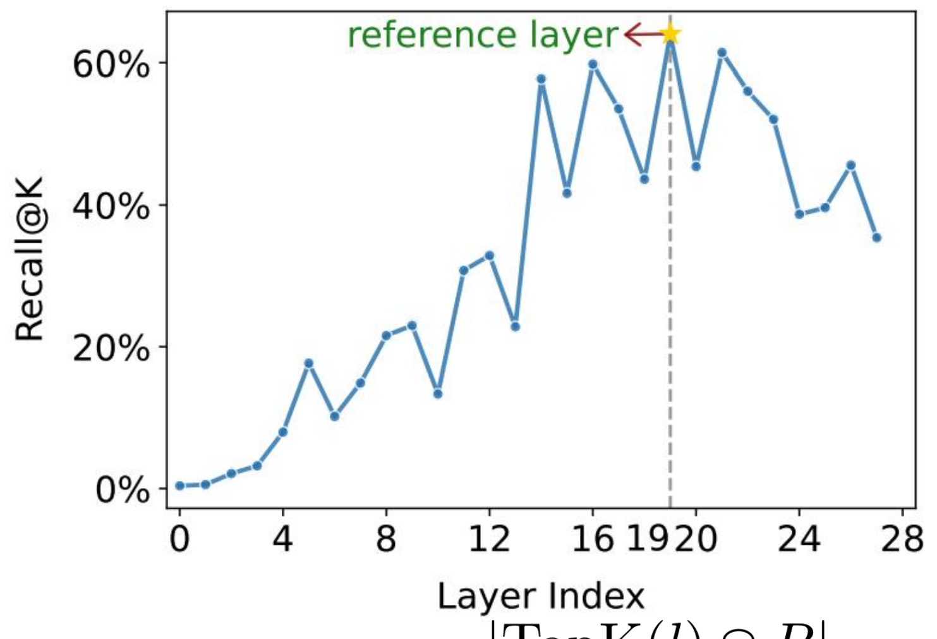
- ◆ Can the attention scores in LLM decoder's internal transformer layers identify relevant visual tokens?

Yes! But not all transformer layers.



Visualization samples of query-to-visual attention heatmaps from different layers of LLaVA-Video-7B, where the query is "what's the color of this cup". The attention scores from shallow layers fail to locate the "cup", while deeper layers can successfully identify it.

- ◆ Which layer should be chosen to select the visual tokens?



$$\text{Recall@}K(l) = \frac{|\text{TopK}(l) \cap R|}{|R|},$$

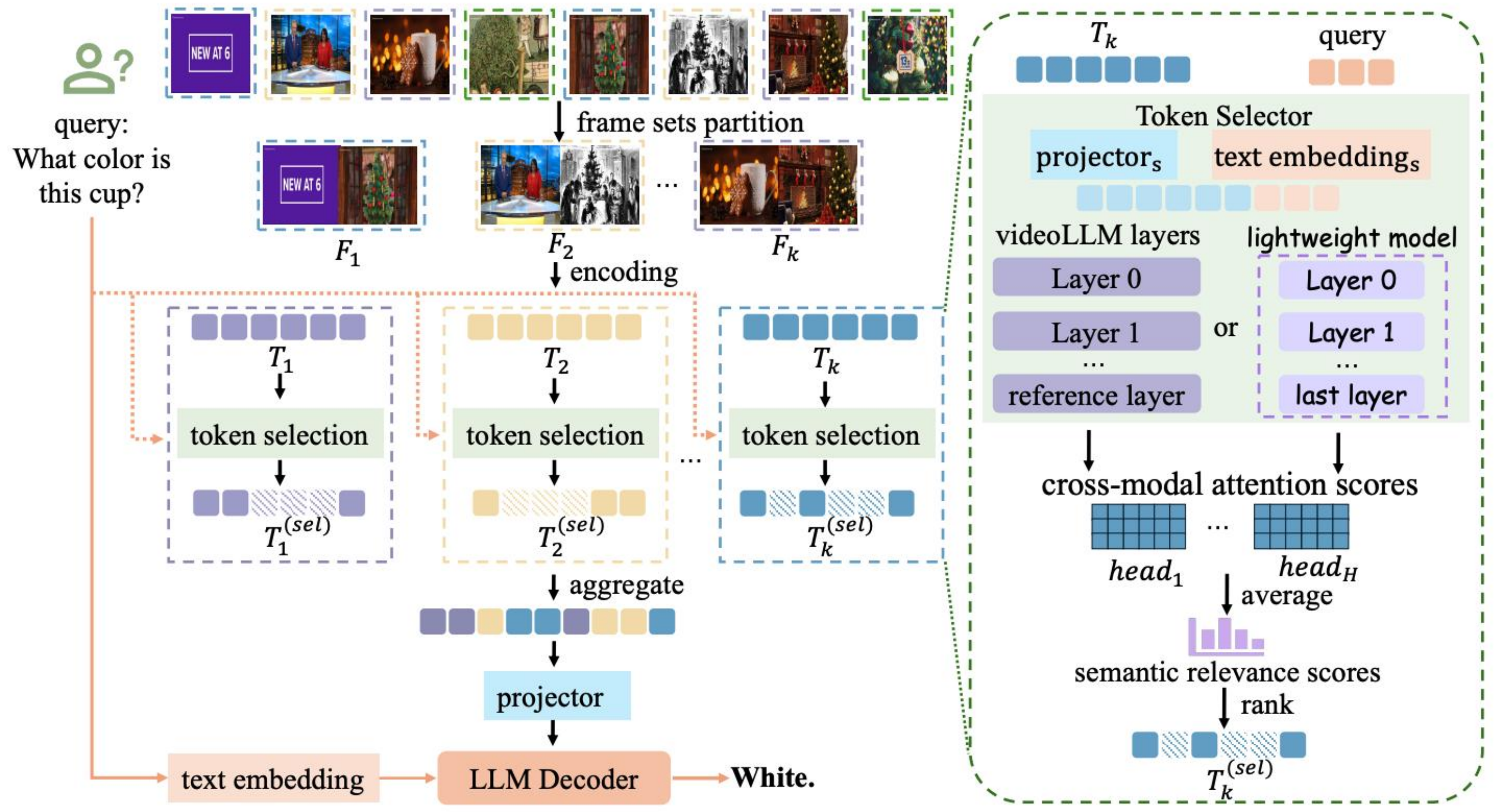
Here, TopK(l) respresent the token set with top-k attention scores at layer l ; R represents the query-related token set.

A higher Recall@K indicates the attention scores of that layer can more accurately identify the semantically related visual tokens.

Metheods

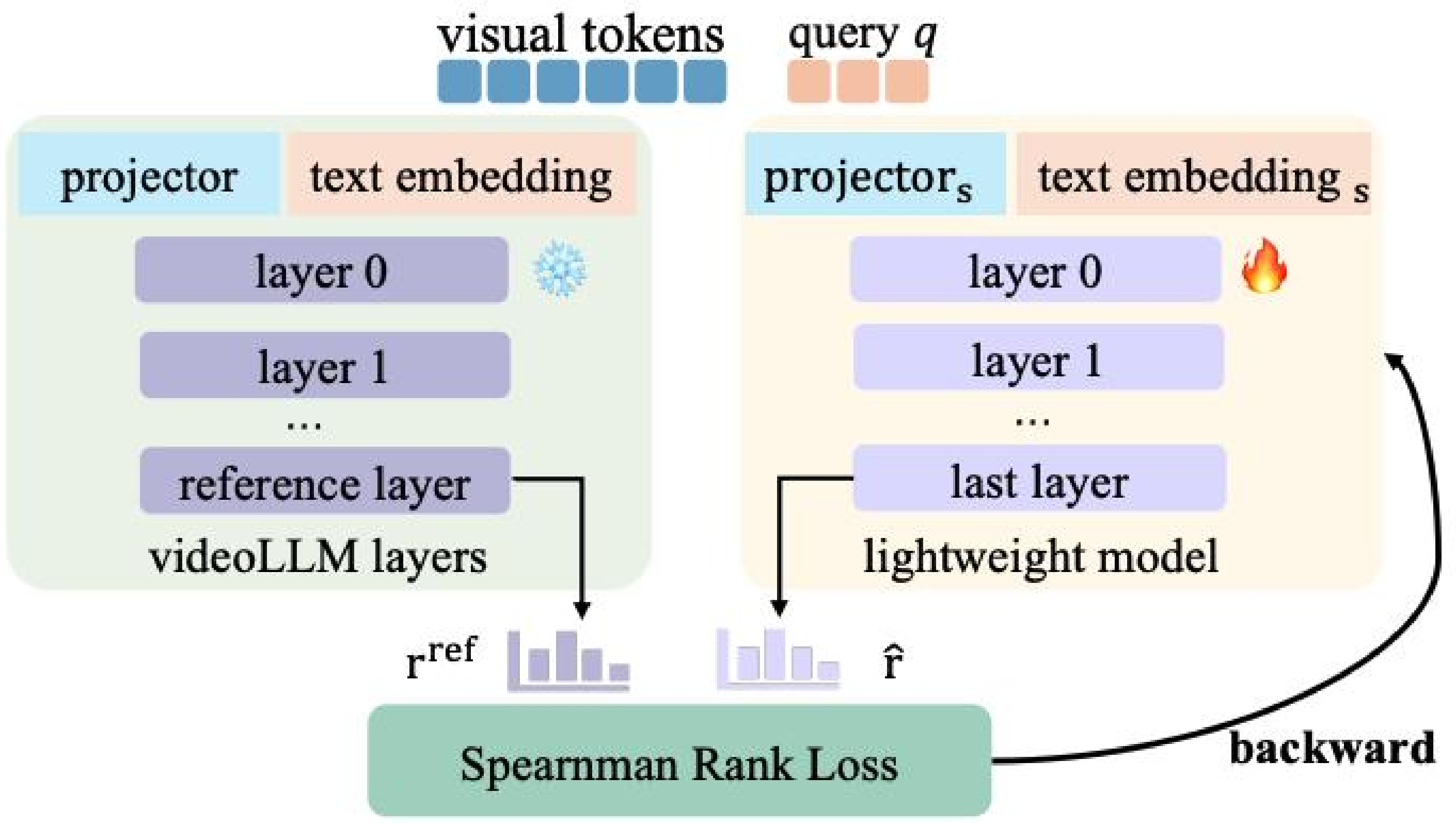
➢ FlexSelect Pipeline

- ◆ Split the sampled frames into different frame sets
- ◆ Select topk tokens at reference layer per set
- ◆ Aggregate the selected visual tokens and then forward to get final answer



➢ FlexSelect-Lite: a more efficient alternative

- ◆ Using the model's own reference layer to select tokens remains costly. What about selecting visual tokens using a small network?



We align a small model's (0.5B) cross-attn scores with the reference layer's scores by optimizing the spearman rank correlation coefficient between them. Once trained, we use the small model to rank the visual tokens, further improving the efficiency.

Evaluation Results

➢ Main Results

- ◆ ***FlexSelect siginificantly improve performance on 3 different types of Video-LLMs, including 7B and 72B, across 4 long video benchmarks.***
- ◆ ***FlexSelect-Lite trades a slight performance drop for significantly higher efficiency, while also surpassing the baseline.***

Model	Size	VideoMME		MLVU	LongVB	LVBench
		Long	Overall	M-Avg	Val	Test
Proprietary Models						
GPT-4o [24]	-	65.3	71.9	64.6	66.7	34.7
Gemini-1.5-Pro [30]	-	**67.4**	**75.0**	-	64.0	33.1
Open-Source VideoLLMs						
mPLUG-Owl3 [39]	7B	50.1	59.3	63.7	52.1	43.5
Qwen2-VL [32]	7B	53.8	63.3	66.9	55.6	42.4
NVILA [21]	8B	54.8	64.2	70.1	57.7	-
VideoLLaMA3 [40]	7B	-	66.2	73.0	59.8	45.3
Aria [15]	8x3.5B	58.8	67.6	70.6	65.3	-
Oryx-1.5 [22]	34B	59.3	67.3	72.3	62.0	30.8
Video-XL-Pro [20]	3B	-	60.0	70.6	56.7	-
SF-LLaVA-1.5 [38]	7B	-	63.9	71.5	62.5	45.3
TPO [17]	7B	55.4	65.6	71.1	60.1	-
Quato [23]	7B	55.7	65.9	71.9	59.0	-
ViLAMP [7]	7B	57.8	67.5	72.6	61.2	45.2
LLaVA-Video [47]	7B	52.9	64.4	68.6	58.2	43.1
+ FlexSelect	7B	59.8 ↑6.9	68.9 ↑4.5	73.2 ↑4.6	61.9 ↑3.7	52.9 ↑9.8
+ FlexSelect-Lite	7B	58.3 ↑5.4	68.3 ↑3.9	71.8 ↑3.2	60.7 ↑2.5	52.2 ↑9.1
InternVL2.5 [6]	8B	52.8	64.2	68.9	59.5	43.4
+ FlexSelect	8B	58.1 ↑5.3	67.0 ↑2.8	71.9 ↑3.0	60.1 ↑0.6	49.7 ↑6.3
+ FlexSelect-Lite	8B	57.9 ↑5.1	67.2 ↑3.0	71.9 ↑3.0	61.2 ↑1.7	49.9 ↑6.5
Qwen2.5-VL [1]	7B	55.6	65.4	70.2	59.5	45.3
+ FlexSelect	7B	59.3 ↑3.7	68.2 ↑2.8	72.5 ↑2.3	62.4 ↑2.9	51.2 ↑5.9
+ FlexSelect-Lite	7B	58.6 ↑3.0	67.4 ↑2.0	70.3 ↑0.1	61.9 ↑2.4	50.0 ↑4.7
LLaVA-Video [47]	72B	61.9	70.0	71.2	62.4	45.5
+ FlexSelect	72B	66.1 ↑4.2	73.1 ↑3.1	76.0 ↑4.8	**66.9** ↑4.5	55.5 ↑10.0
Qwen2.5 VL [1]	72B	63.9	73.4	76.3	66.2	47.3
+ FlexSelect	72B	66.9 ↑3.0	74.4 ↑1.0	**76.6** ↑0.3	66.4 ↑0.2	**56.6** ↑9.3

➢ Compare to other token reduction methods

- ◆ ***FlexSelect surpasses other attention-based token pruning methods like FastV, Dycoke and others under same retain ratio. More results can be seen in paper***

Method	Retain Ratio (%)	VideoMME wo Sub	VideoMME w Sub
LLaVA-OV-7B	100	58.5	61.3
FastV	35	57.3	60.5
DyCoke	14.25	58.3	60.7
FlexSelect	14.25	**60.4**	**62.2**
DyCoke	18.75	59.5	61.4
FlexSelect	18.75	**61.0**	**62.4**
DyCoke	23.25	58.8	61.0
FlexSelect	23.25	**60.4**	**63.0**