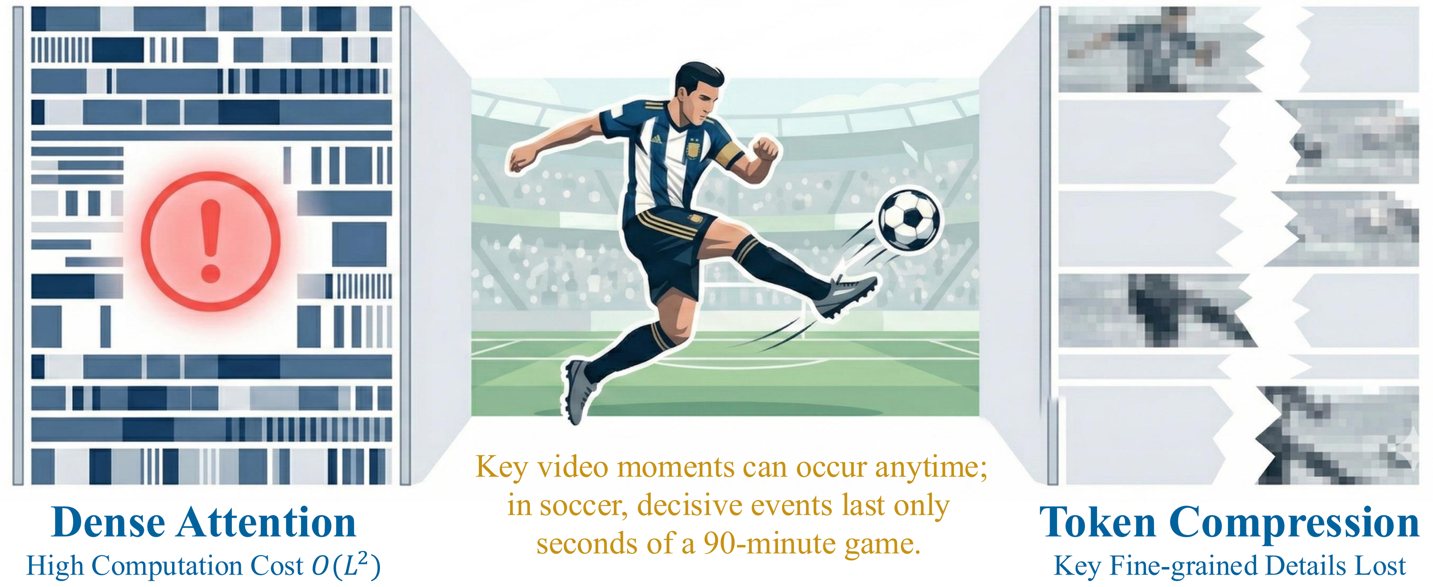
VideoNSA: Native Sparse Attention Scales Video Understanding

Enxin Song[1] Wenhao Chai[2] Shusheng Yang[3] Ethan Armand[1] Xiaojun Shan[1] Haiyang Xu[1] Jianwen Xie[4] Zhuowen Tu[1]

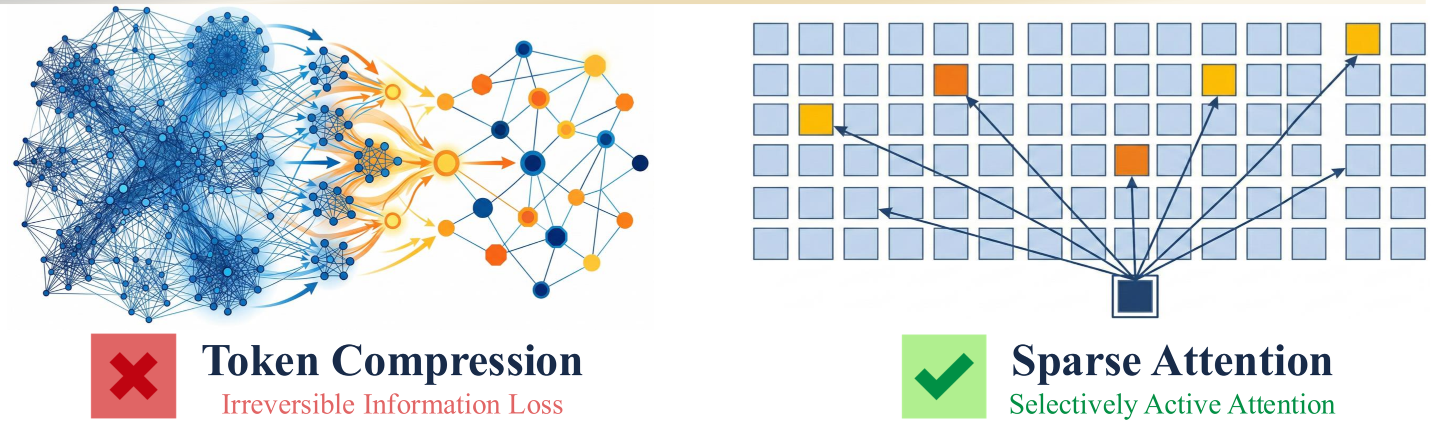
[1]University of California, San Diego [2]Princeton University [3]New York University [4]Lambda, Inc

Motivation

Key Information vs. Compute Cost



Attention as Message Passing in a Graph

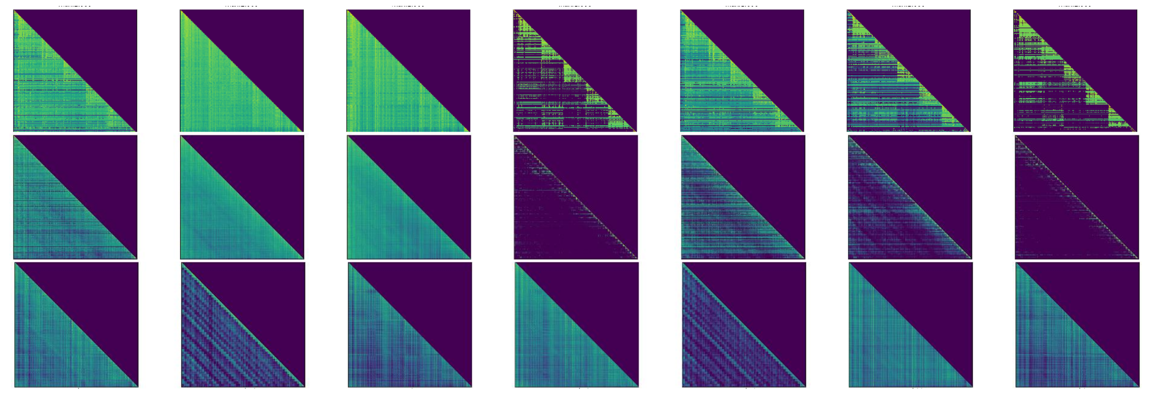


Unique Sparse Attention Pattern

300 Tokens Layer 28

3K Tokens Layer 28

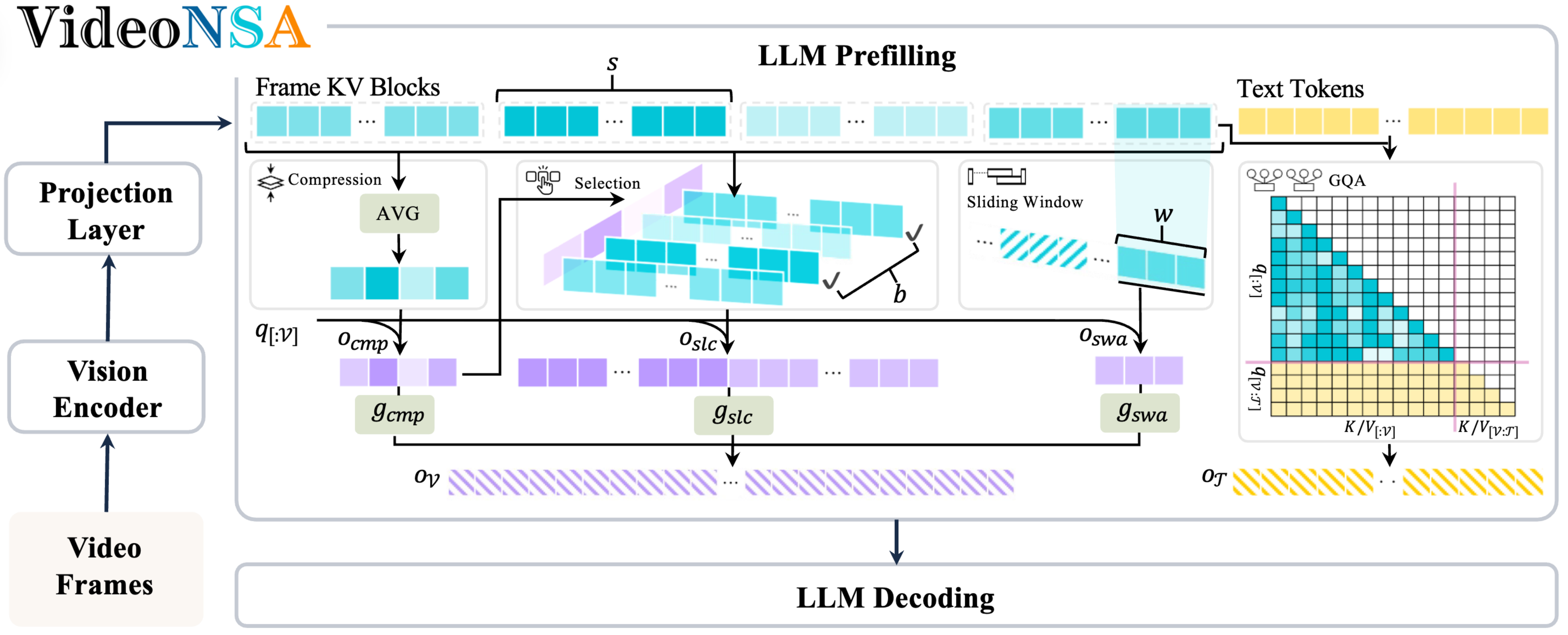
3K Tokens Layer 27



Contribution

- We propose VideoNSA, a hardware-aware native sparse attention mechanism, and systematically investigate its effectiveness for video understanding, scaling up to a 128K vision context length.
- We introduce hybrid sparse attention in VideoNSA, enabling flexible allocation of information and attention budgets to achieve optimal performance across diverse task.
- We dynamically combine global and local attention through three complementary branches, which effectively reduce attention sinks in long vision contexts.

Method



Experiments

Model	Long-form Video				Temporal	Spatial
	LVB	$MLVU_{test}$	TimeScope	LTS	Tomato	VSIBench
LLaVA-OneVision-7B (Li et al., 2024a)	56.3	–	–	–	25.5	32.4
LLaVA-Video-7B (Zhang et al., 2024c)	58.2	–	74.1	34.0	–	35.6
VideoLLaMA3-8B (Zhang et al., 2025a)	59.8	47.7	69.5	–	–	–
InternVL2.5-8B (Chen et al., 2024c)	60.0	–	55.8	–	–	–
Video-XL-2 (Qin et al., 2025b)	**61.0**	**52.2**	–	–	–	–
Qwen2.5-VL-7B (Qwen et al., 2025)	58.7	51.2	81.0	40.7	22.6	29.7
Qwen2.5-VL-7B-AWQ (Team, 2024)	59.0	46.0	–	–	–	35.0
Qwen2.5-VL-7B-SFT	57.8	51.2	76.8	40.2	21.7	30.5
Token Compression Methods						
+ FastV (Chen et al., 2024a)	57.3	41.8	46.5	35.6	21.6	32.0
+ VScan (Zhang et al., 2025b)	58.7	48.1	80.3	31.1	19.1	34.4
+ VisionZip (Yang et al., 2025c)	52.4	33.1	43.5	40.4	23.6	32.1
Sparse Attention Methods						
+ Tri-Shape (Li et al., 2024c)	59.5	49.2	82.7	28.4	22.1	34.9
+ MInference (Jiang et al., 2024)	59.2	49.2	82.7	**44.4**	23.0	36.5
+ FlexPrefill (Lai et al., 2025)	58.4	46.0	83.0	39.1	23.7	34.0
+ XAttention (Xu et al., 2025a)	59.1	50.2	83.1	41.1	21.4	**36.6**
VideoNSA	60.0	51.8	**83.7**	**44.4**	**26.5**	36.1

Branch			Long Video Understanding				Temporal Reasoning	Spatial Understanding
CMP	SLC	SWD	LVB	$MLVU_{test}$	TimeScope	LTS	Tomato	VSIBench
✓			48.1	43.9	41.5	25.1	23.3	29.2
	✓		48.4	47.7	63.7	37.1	24.0	27.6
		✓	49.1	40.2	59.3	29.8	24.0	29.8
✓	✓		49.4	42.7	57.3	32.4	23.5	29.4
✓		✓	49.3	42.4	65.2	34.4	23.0	29.1
	✓	✓	48.8	43.4	57.3	31.6	24.5	30.3
✓	✓	✓	**60.0**	**51.8**	**83.7**	**44.4**	**26.5**	**36.1**

Table 12: Results on LSDBench (Qu et al., 2025).

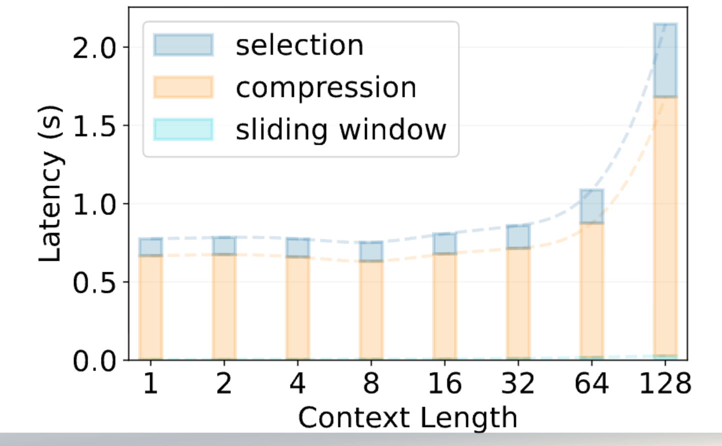
Model	Accuracy
LongVA (Zhang et al., 2024b)	32.5
LongVila (Chen et al., 2024b)	49.8
InternVL2.5 (Chen et al., 2024c)	50.1
Qwen2.5-VL-7B (Qwen et al., 2025)	52.2
Qwen2.5-VL-7B-SFT	52.5
Sparse Attention Methods	
+ Tri-Shape (Li et al., 2024c)	49.5
+ MInference (Jiang et al., 2024)	49.5
+ FlexPrefill (Lai et al., 2025)	52.3
+ XAttention (Xu et al., 2025a)	51.3
VideoNSA	**55.2**

Findings

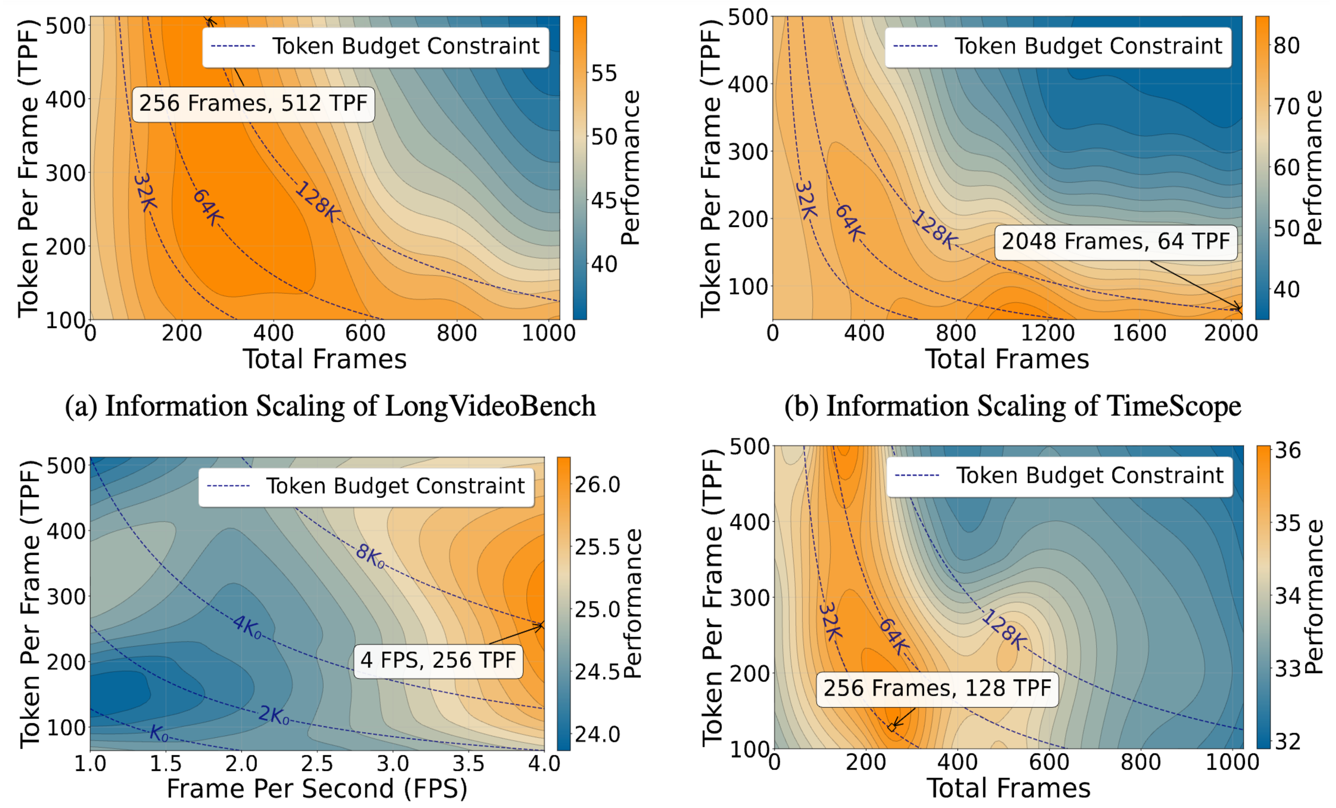
Sparse Attention Training Benefits Dense Attention Inference

Model	Long Video Understanding				Temporal Reasoning	Spatial Understanding
	LongVideoBench	$MLVU_{Test}$	TimeScope	LongTimeScope	Tomato	VSIBench
Qwen2.5-VL-7B	58.7	51.2	81.0	40.7	22.6	29.7
Dense-SFT	57.8 (-1.5%)	51.2 (+0.0%)	76.8 (-5.2%)	40.2 (-1.2%)	21.7 (-4.0%)	30.6 (+2.1%)
Dense-NSA	56.1 (-4.4%)	51.6 (+0.8%)	**83.0 (+2.5%)**	40.9 (+0.5%)	23.4 (+3.5%)	33.1 (+10.7%)
VideoNSA	**59.4 (+1.1%)**	**51.8 (+1.2%)**	82.7 (+2.1%)	**44.4 (+9.1%)**	**26.2 (+15.9%)**	**36.1 (+20.3%)**

Latency



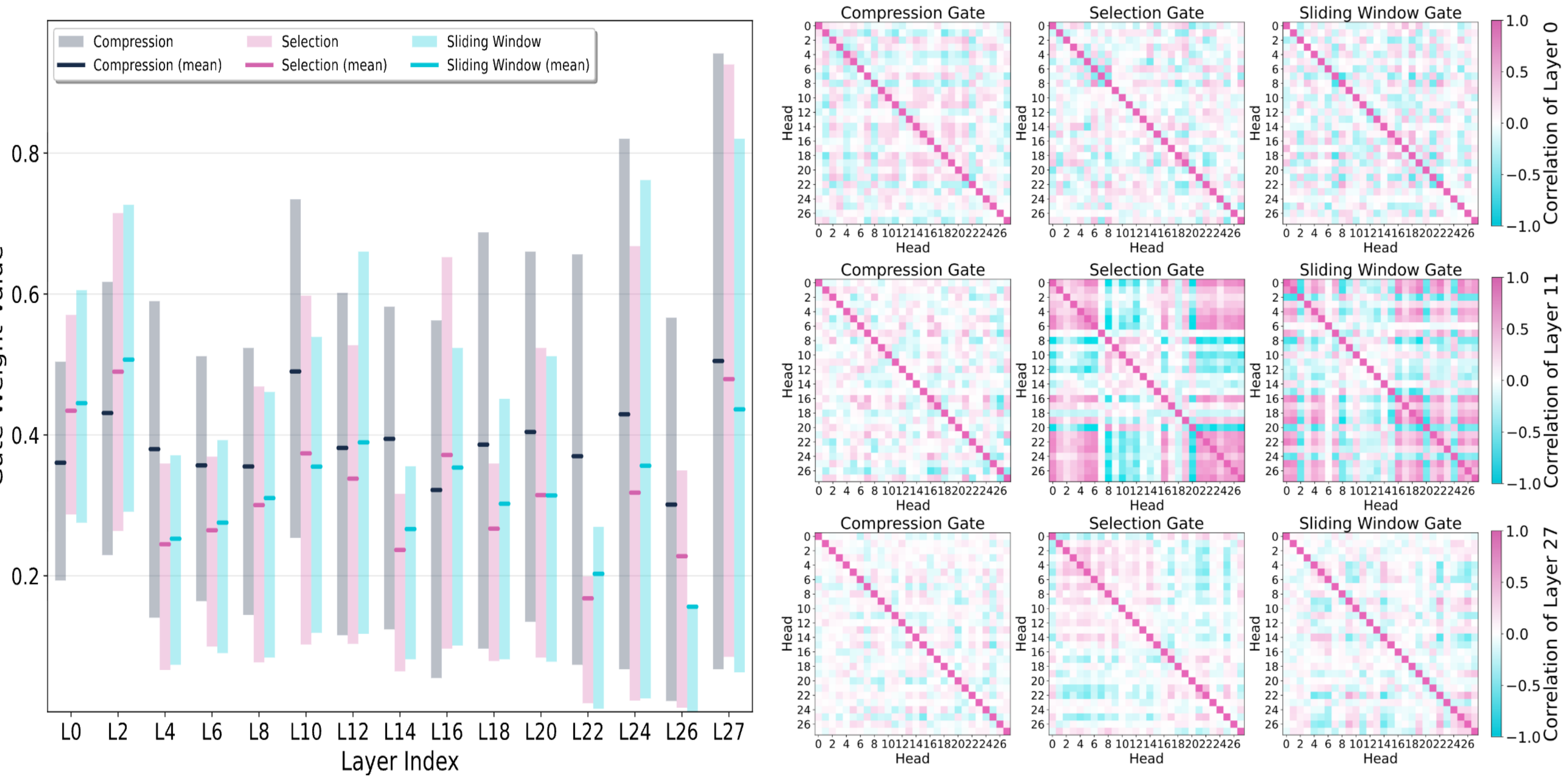
Stable Context Length Scaling



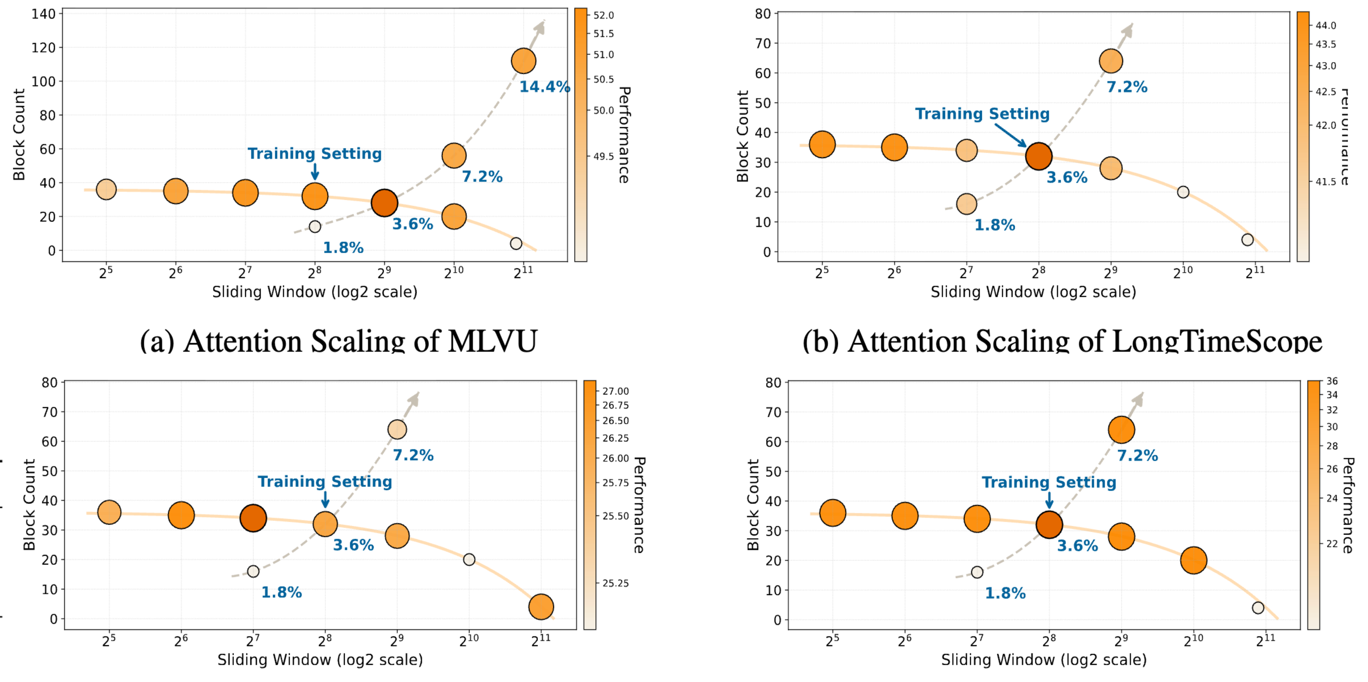
(a) Information Scaling of LongVideoBench (b) Information Scaling of TimeScope

(c) Information Scaling of Tomato (d) Information Scaling of VSIBench

Gate Weight Differs from Text-Only



Sensitive Attention Budget Scaling



(a) Attention Scaling of MLVU (b) Attention Scaling of LongTimeScope

(c) Attention Scaling of Tomato (d) Attention Scaling of VSIBench

$$K_{attn} = Block\ Size \times Block\ Count + Window\ Size$$

With context length L, the fraction of attention used $\gamma = \frac{L(cs+w)}{\frac{L(L-1)}{2}} = \frac{2(cs+w)}{L-1}$

NSA Reduces Attention Sinks

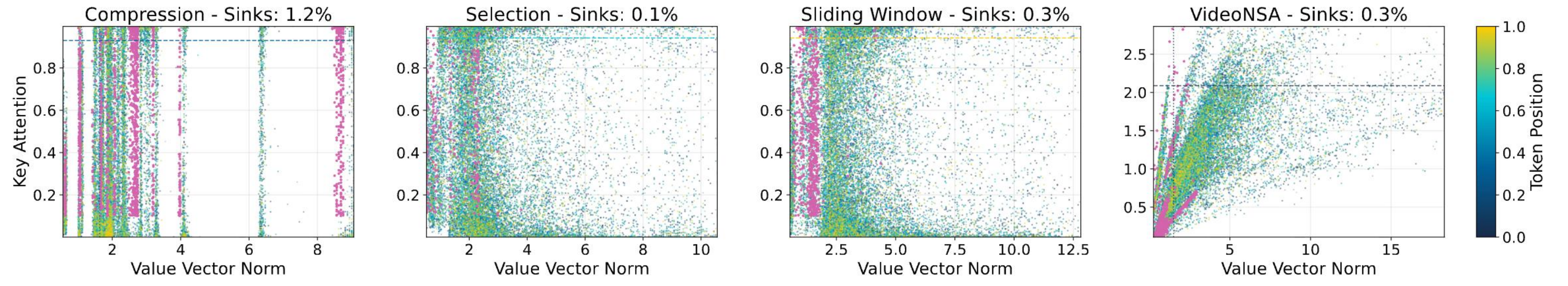


Figure 7: Attention sinks distribution of different branches. VideoNSA maintains a low overall sink ratio, with pink points indicating identified sinks.

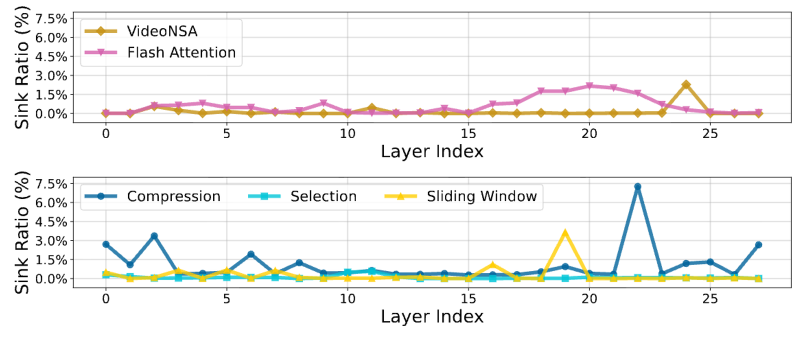


Figure 8: Layer-wise attention sink ratio distribution in different branches and Flash Attention.

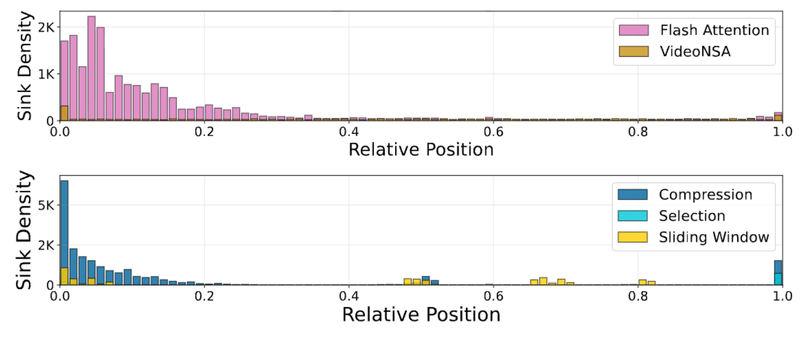


Figure 9: Relative positions of attention sinks in different branches and Flash Attention.