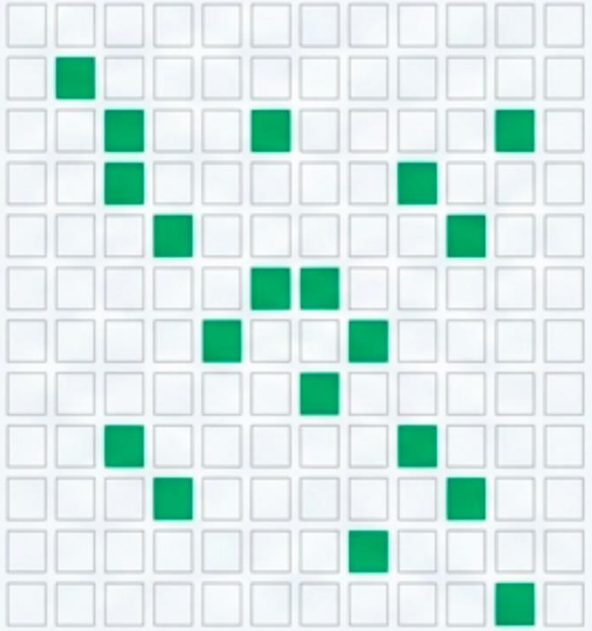
# From Compression to Selection:

## Better and Longer Video Understanding with VideoNSA

**Enxin Song**

**Zhejiang University**

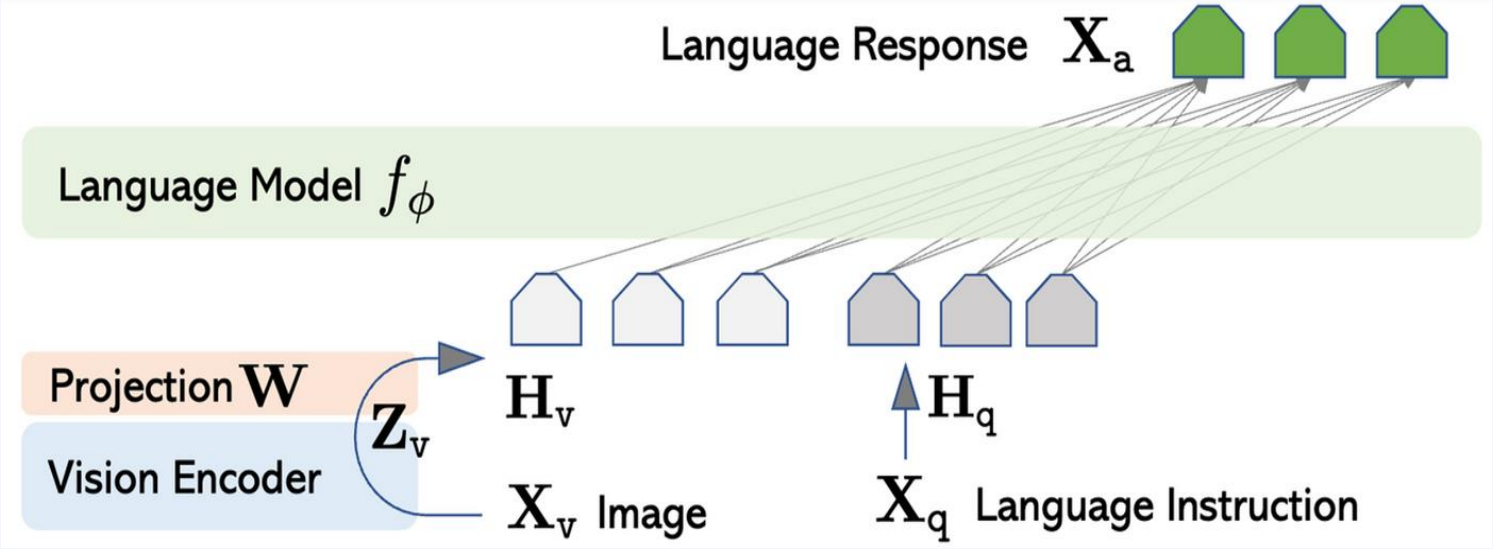
*KAUST Rising Stars in Al Symposium 2026*



# Video LLMs

## How We Connect?

- Connect ViT and LLM
- Adapt from Image LLMs
- Handle longer sequences
- May need more compute
- But less data



# Video LLMs

## Short videos, short captions — can they tell the whole story?

*Figure: Video example of MSR-VTT, which is a widely used video question answering and captioning benchmark.*
*Labeled caption: Teams are playing soccer.*

### Why we need long-form video understanding?

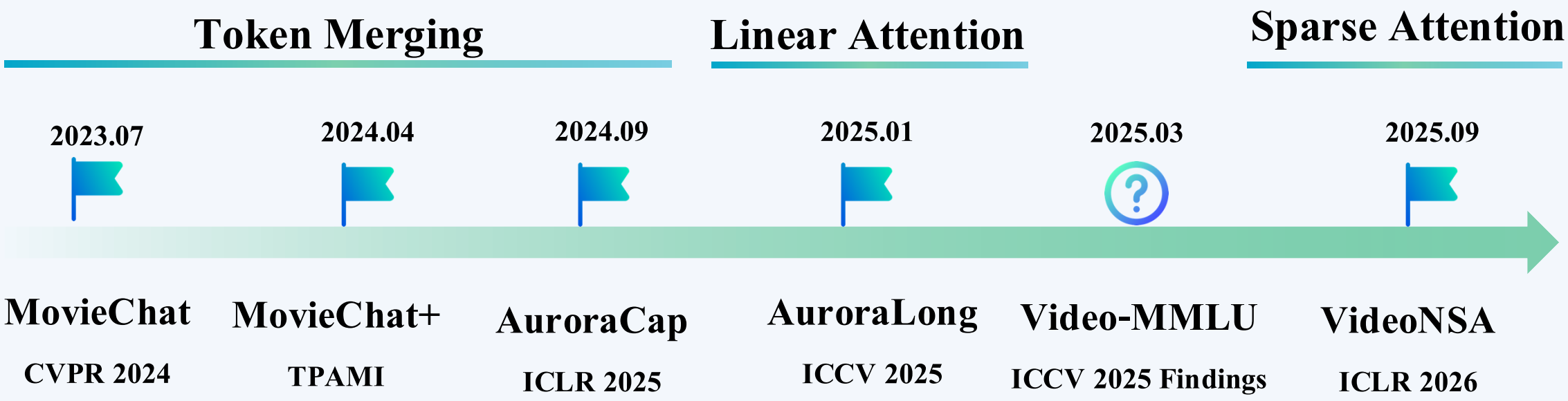
- Temporal Complexity and Granularity
- Narrative Comprehension
- Real-World Applications

### What are the current challenges?

**Efficiency**

**Training Data**

# Video LLMs


Token Merging
Linear Attention
Sparse Attention
2023.07
2024.04
2024.09
2025.01
2025.03
2025.09
MovieChat
CVPR 2024
MovieChat+
TPAMI
AuroraCap
ICLR 2025
AuroraLong
ICCV 2025
Video-MMLU
ICCV 2025 Findings
VideoNSA
ICLR 2026

# Core Paradox of Video Understanding

## Key Information vs. Compute Cost

Dense Attention

High computation cost $O(L^2)$

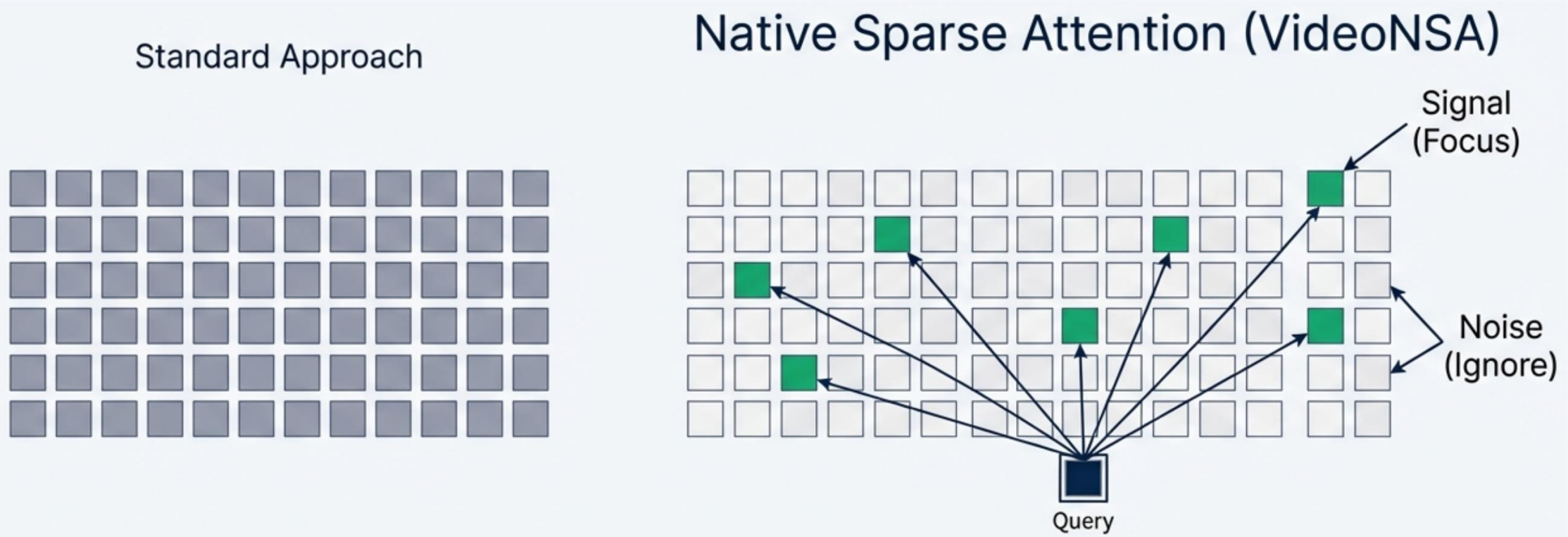
**Salient Moment** only lasts a few seconds

Token Compression

Key fine-grained details are lost

# The Paradigm Shift:

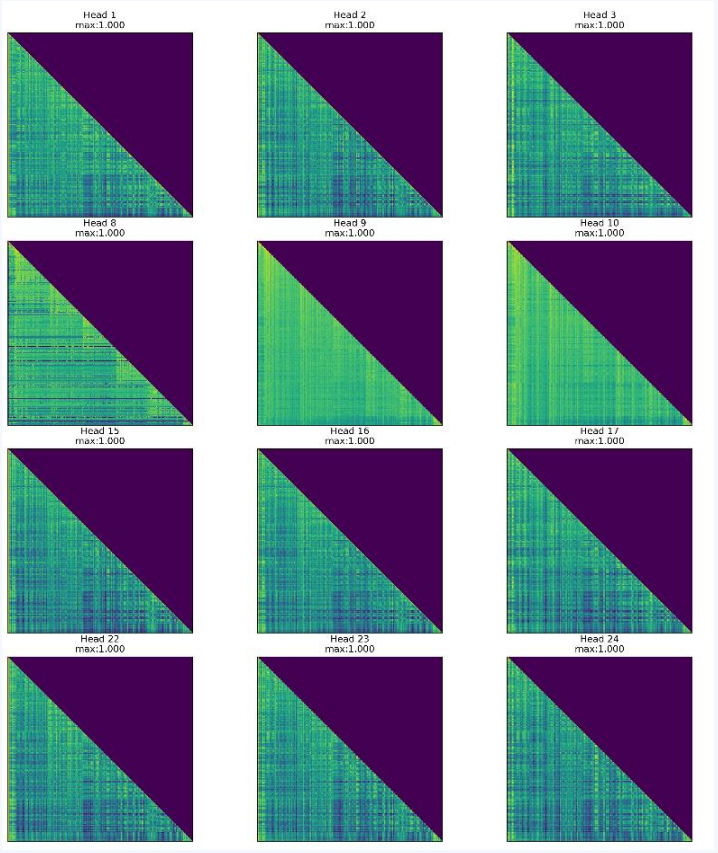
# From Passive Compression to Active Selection



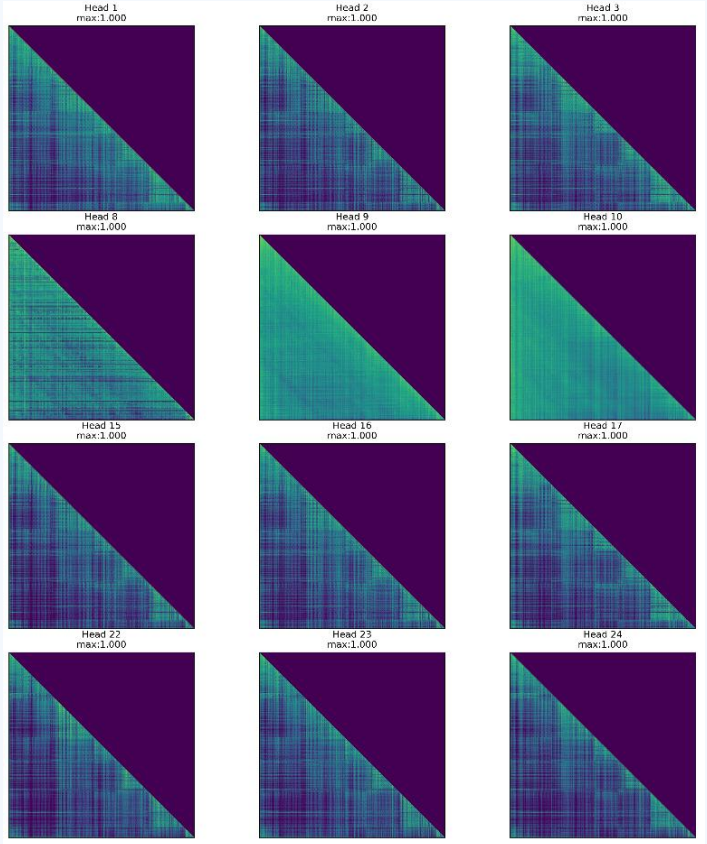
# Observation

**Sparse Attention Map**

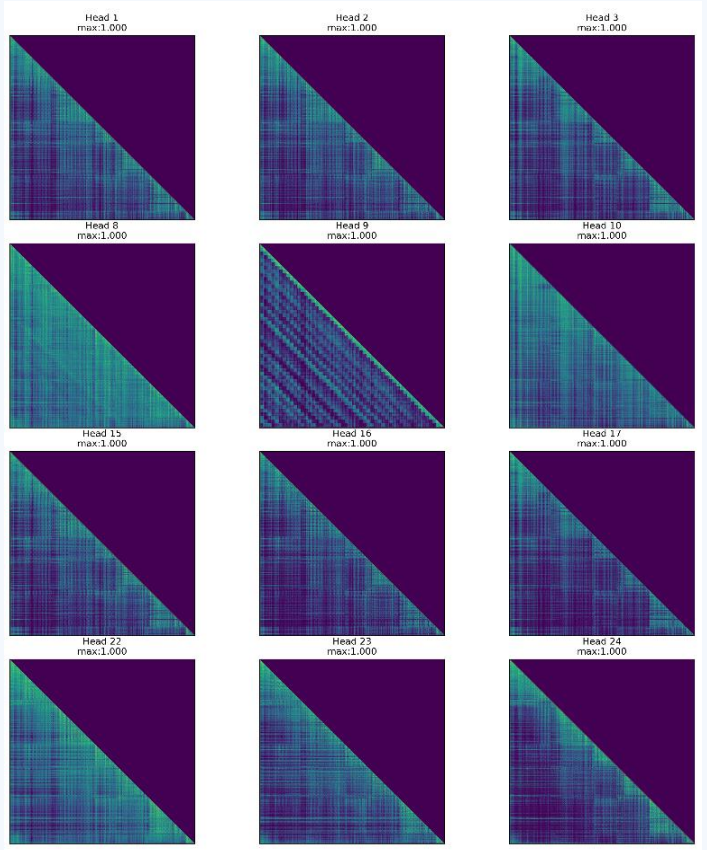
**Unique Attention Pattern**




**Layer 28, 300 tokens**




**Layer 28, 3000 tokens**

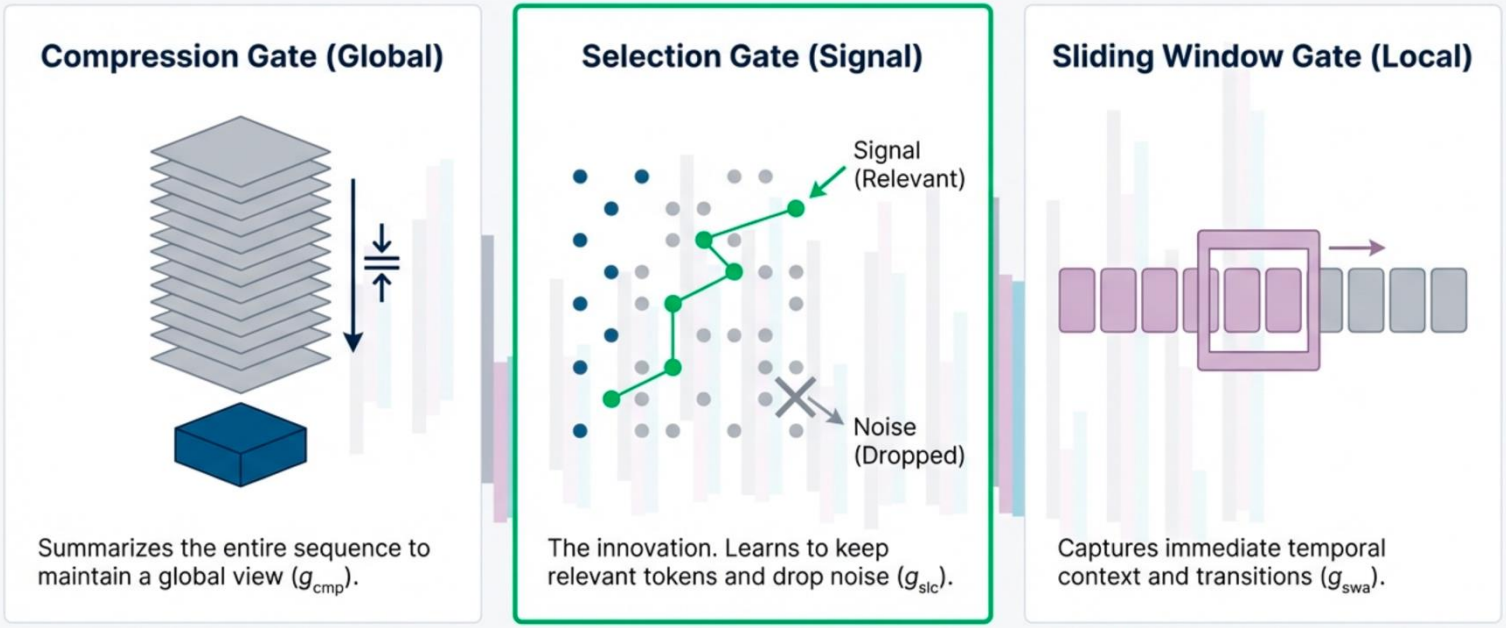



**Layer 27, 3000 tokens**

# Efficient Video Understanding

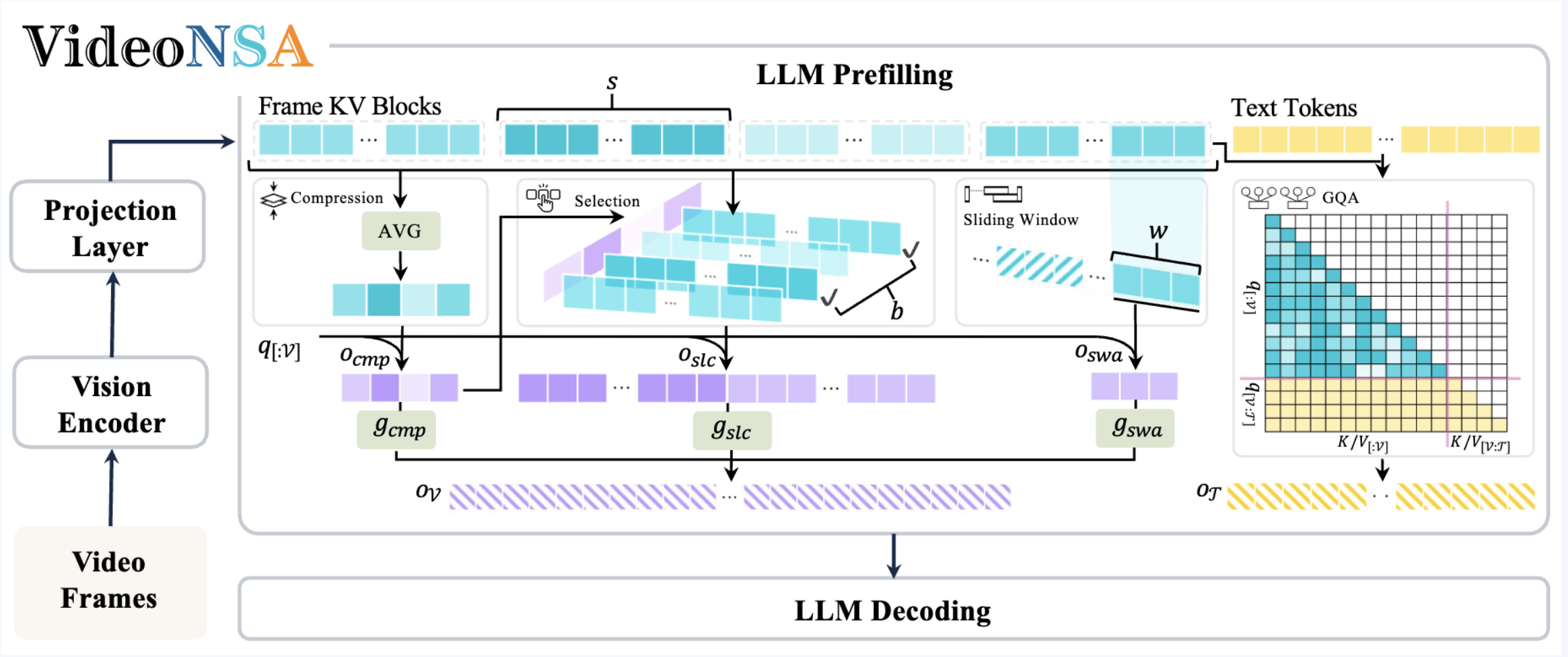
**VideoNSA: Native Sparse Attention Scales Video Understanding**

**ICLR 2026**




*Finished during the internship in UCSD, advised by Prof. Zhuowen Tu*

# Efficient Video Understanding



# VideoNSA

Table 1: Results on long video understanding, temporal reasoning and spatial understanding tasks. LVB, LTS for LongVideobench (Wu et al., 2024) and LongTimeScope (Zohar et al., 2025).

| Model | Long-form Video | | | | Temporal | Spatial |
|---|---|---|---|---|---|---|
| | LVB | $\text{MLVU}_{test}$ | TimeScope | LTS | Tomato | VSIBench |
| LLaVA-OneVision-7B (Li et al., 2024a) | 56.3 | – | – | – | 25.5 | 32.4 |
| LLaVA-Video-7B (Zhang et al., 2024b) | 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., 2024b) | 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 |

## Baselines:

- AWQ Model
- SFT with Same Dataset
- Token Compression
- Sparse Attention

## Competitive results.

## Scalable with better data.

# VideoNSA

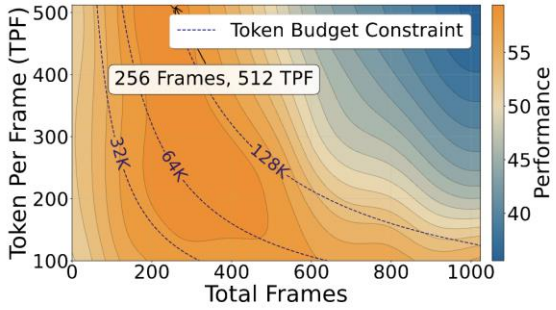
## Do learned sparse attention weights remain beneficial in dense attention settings?

Table 3: Ablation study on transferring sparse attention weights to dense attention across tasks.

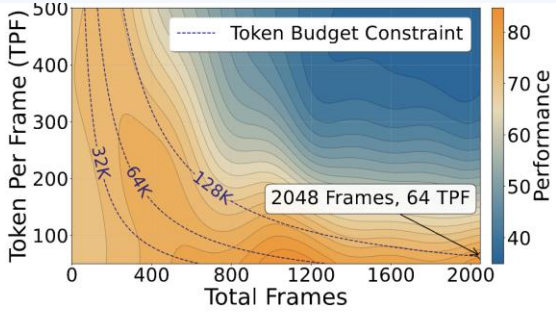
| Model | Long Video Understanding | | | | Temporal Reasoning | Spatial Understanding |
|---|---|---|---|---|---|---|
| | LongVideoBench | $\text{MLVU}_{Test}$ | TimeScope | LongTimeScope | Tomato | VSIBench |
| Qwen2.5-VL-7B | 58.7 | 51.2 | 81.0 | 40.7 | 22.6 | 29.7 |
| Dense-SFT | 57.8 (-1.5%) | 51.2 (+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%)** |

# VideoNSA

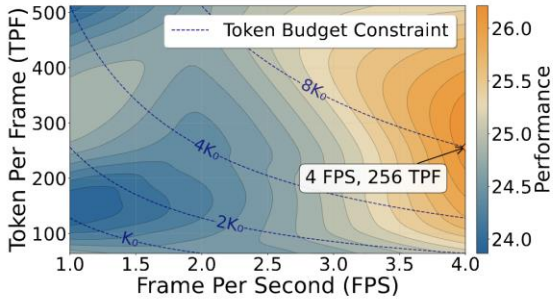
## How far can VideoNSA scale in context length?



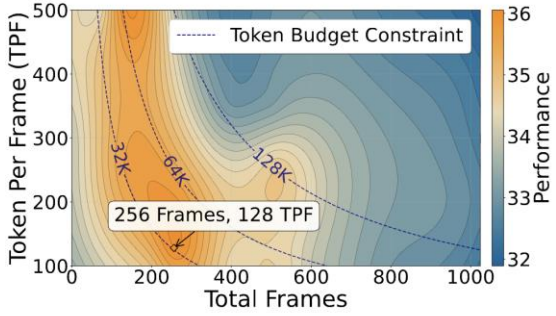

(a) Information Scaling of LongVideoBench




(b) Information Scaling of TimeScope

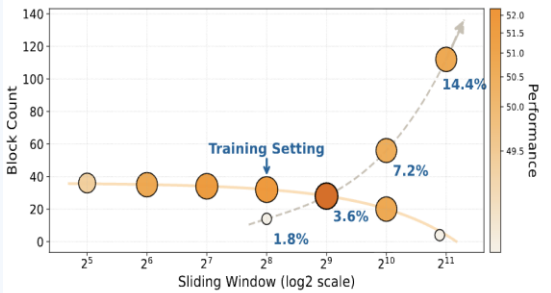



(c) Information Scaling of Tomato

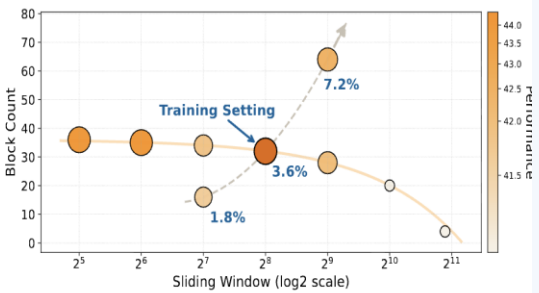



(d) Information Scaling of VSIBench

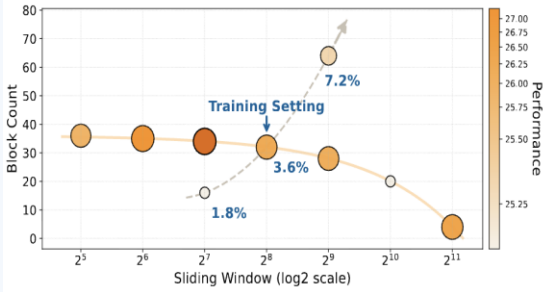
## How to allocate the attention budget?



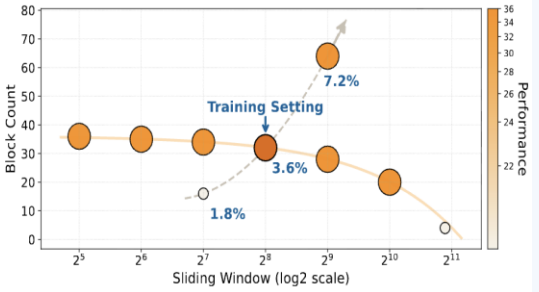

(a) Attention Scaling of MLVU




(b) Attention Scaling of LongTimeScope



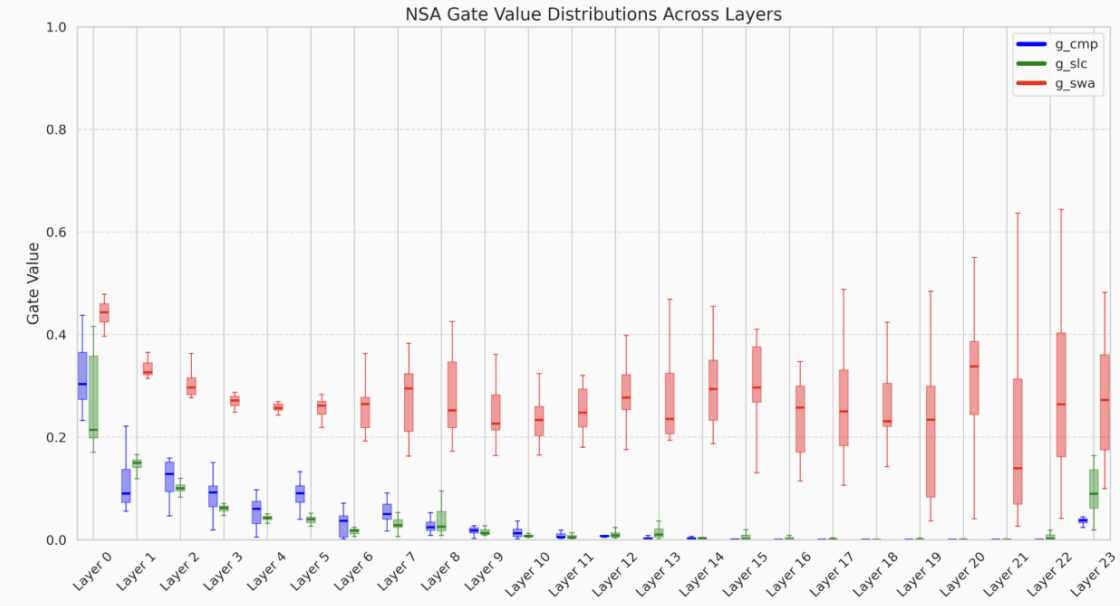

(c) Attention Scaling of Tomato



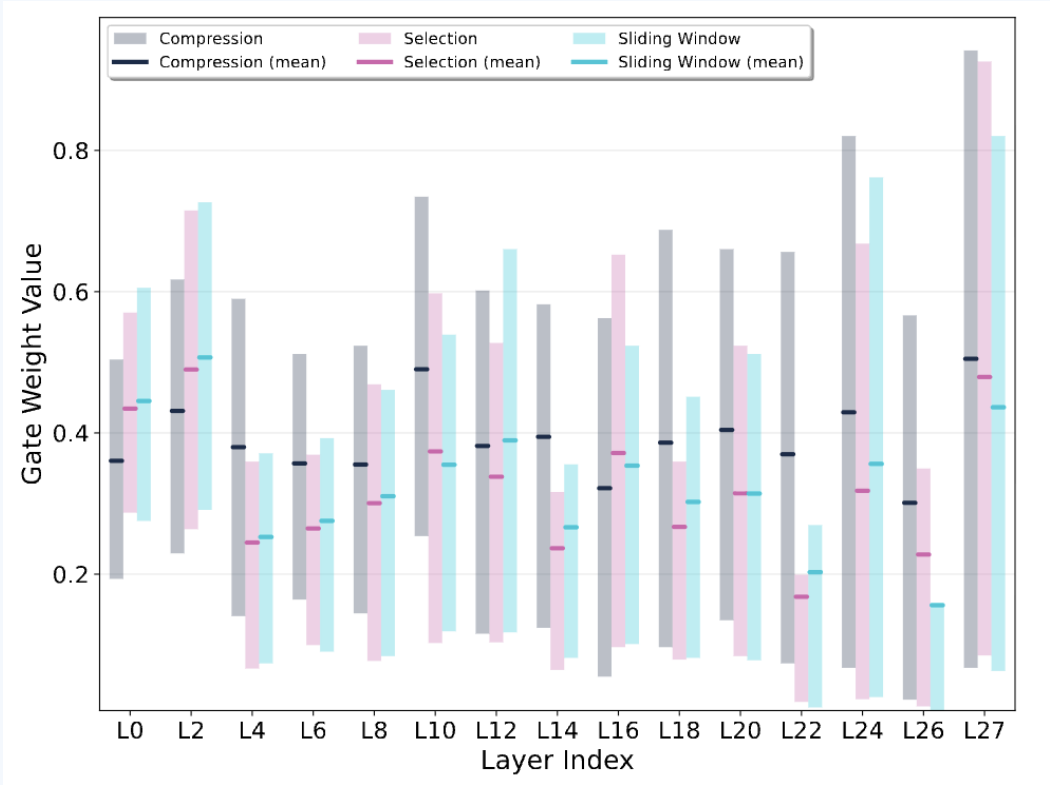

(d) Attention Scaling of VSIBench

# VideoNSA

## What roles do compression, selection, and sliding-window gates play in VideoNSA?

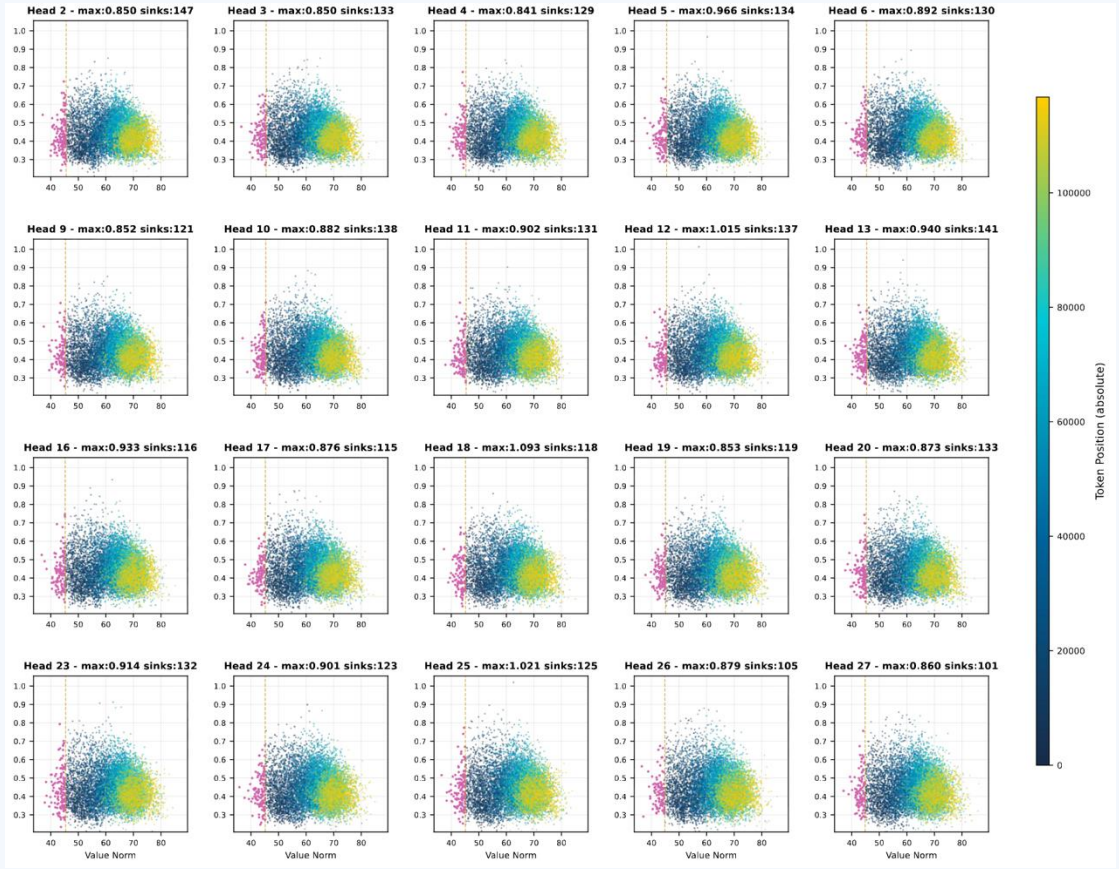



Distributions of gate values for NSA model.

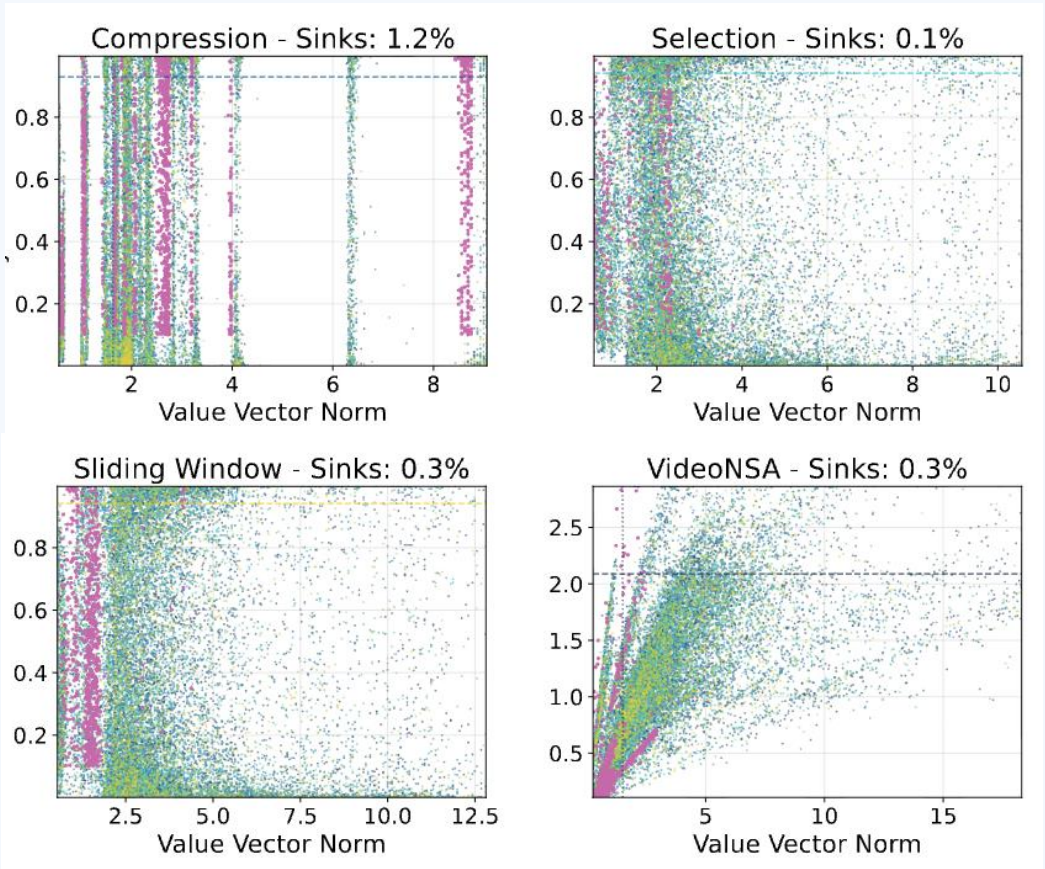


# VideoNSA

## Do learnable sparse mechanisms induce dynamic attention sinks?




*Attention Sink Distribution of Layer 14 in 128K Flash Attention*




*Attention Sink Distribution of Layer 14 in 128K VideoNSA*

# VideoNSA

Webpage: https://enxinsong.com/VideoNSA-web/

Github: https://github.com/Espere-1119-Song/VideoNSA

Paper: https://arxiv.org/abs/2510.02295

Model: https://huggingface.co/Enxin/VideoNSA

Dataset: https://huggingface.co/datasets/Enxin/VideoNSA-data

# VideoNSA


**Thank you for your attention!**

**Welcome Questions and discussions.**