Computer Science Ph.D. Applicant

# Better and Longer Video Understanding

ENXIN SONG

# Research Overview

## Efficient Long-Sequence Modeling

### Long Video

- MovieChat: From Dense Token to Sparse Memory for Long Video Understanding, CVPR 2024
- MovieChat+: Question-aware Sparse Memory for Long Video Question Answering, TPAMI 2025

### Detailed Caption

- AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark, ICLR 2025

### Efficient Architecture

- AuroraLong: Bringing RNNs Back to Efficient Open-Ended Video Understanding, ICCV 2025
- VideoNSA: Native Sparse Attention Scales Video Understanding, ICLR 2026

## Benchmarking and Evaluation

### Knowledge

- Video-MMLU: A Massive Multi-Discipline Lecture Understanding Benchmark, ICCV 2025 Findings

## Applications of Generative Models

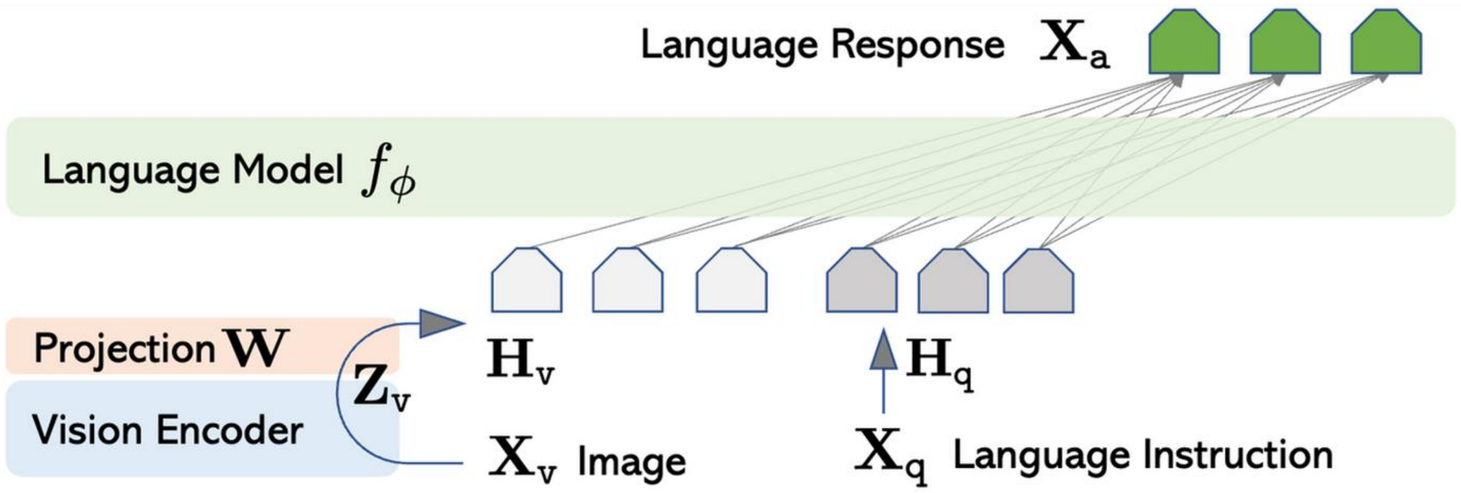
### Masked Image Modeling

- Fantasy: Transformer Meets Transformer in Text-to-Image Generation
- Meissonic: Revitalizing Masked Generative Transformers for Efficient High-Resolution Text-to-Image Synthesis, ICLR 2025

# Video LLMs

## How We Connect?

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

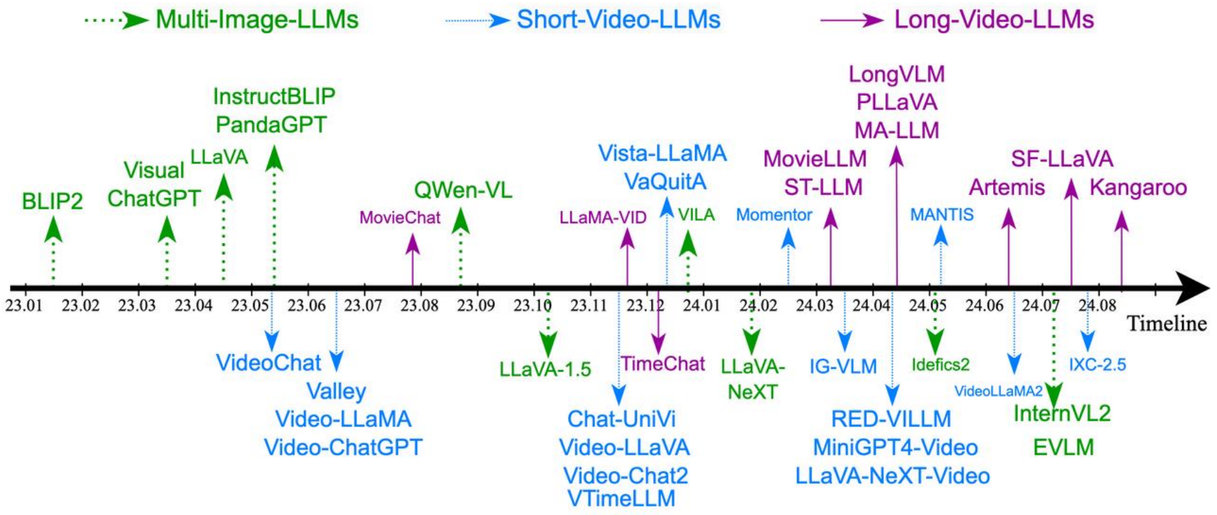


# Video LLMs

**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.*

# Long-form Video Understanding




## Why we need long-form video understand

Temporal Complexity and Granularity, Narrative Comprehension, Real-World Applications,*etc*

## What are the current challenges?

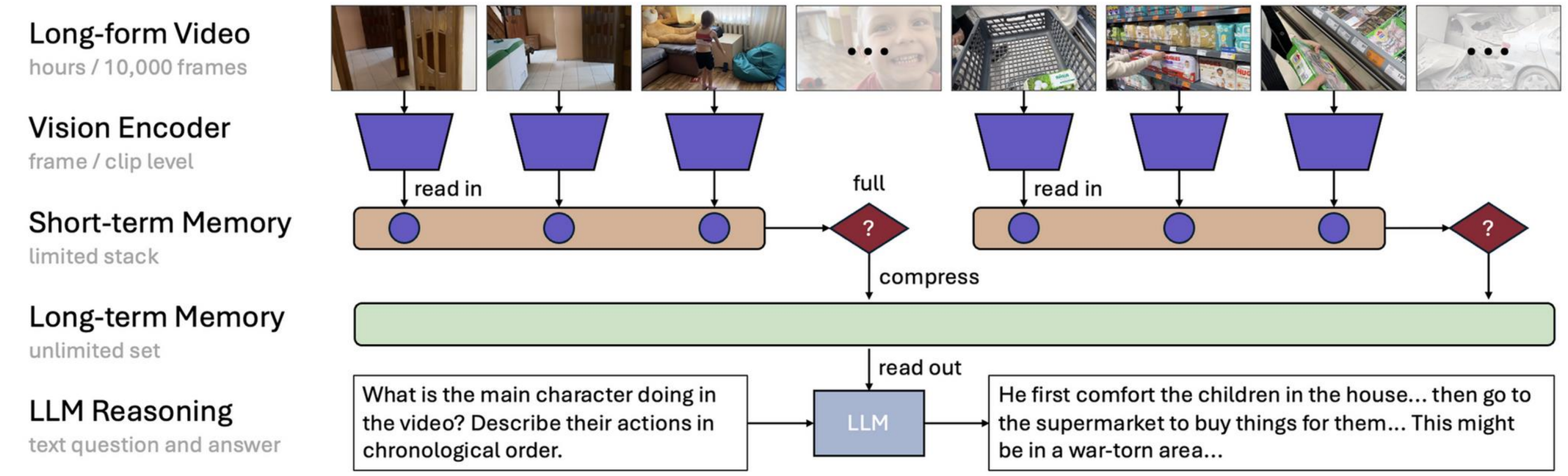
Efficiency, Training Data, *etc*

## Can we do that with current LMMs?

Yes! We found that the LMMs trained on images and short videos can be adapted to long-form video tasks even without further fine-tuning.

# MovieChat

**First ever video understanding system that can take over 10,000 frames as input.**

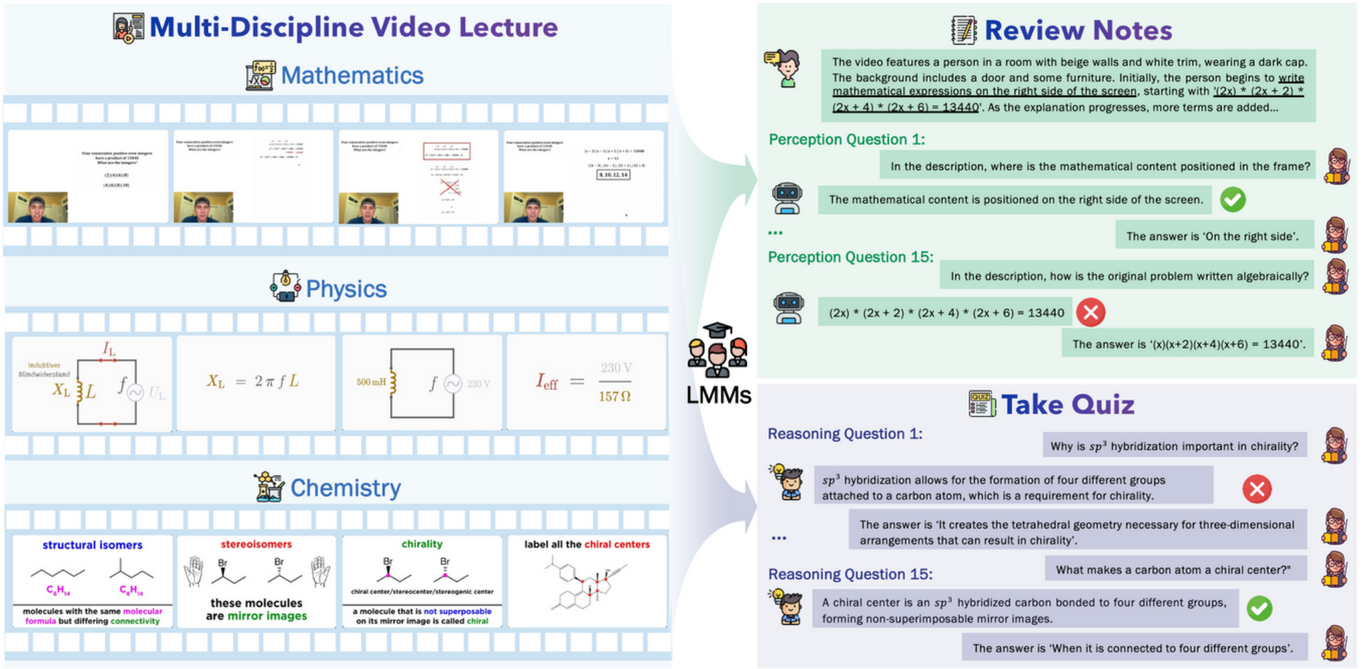


# Lecture Video Understanding

**Can video LLMs really understand real-world lectures?**

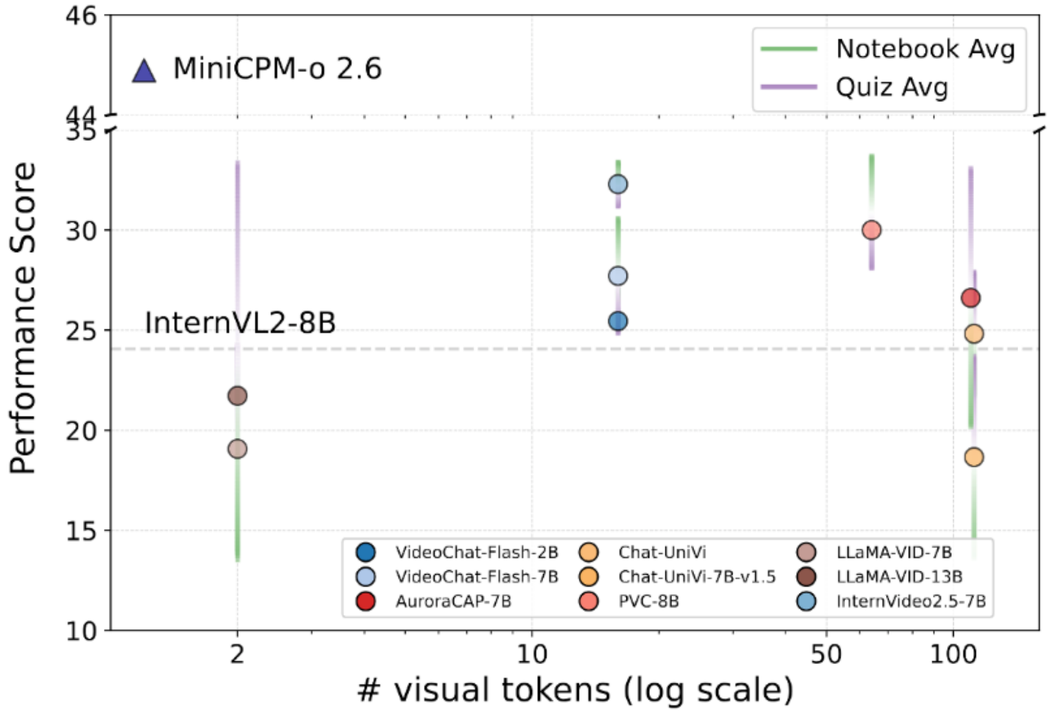
**NOT YET.**

ICCV Findings



# Video-MMLU

**Can LMMs with visual token compression sustain strong performance in complex, context-rich lecture understanding tasks like Video-MMLU?**



# 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

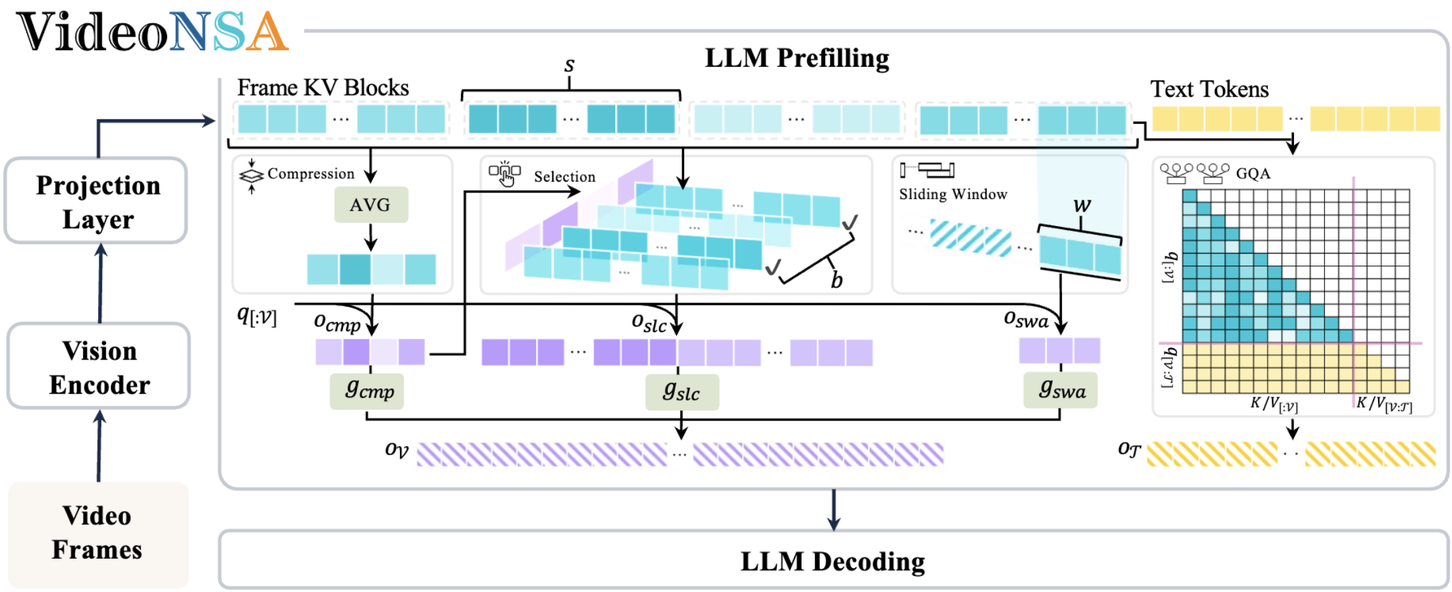
Token Compression
Key fine-grained details are lost

Humans perceive video at 60 Hz and effortlessly capture critical moments, yet current models are often limited by compute and forced to sample as low as 1 FPS.
We need a mechanism that *covers the whole scene* while *precisely focusing on key moments*.

# Efficient Video Understanding

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

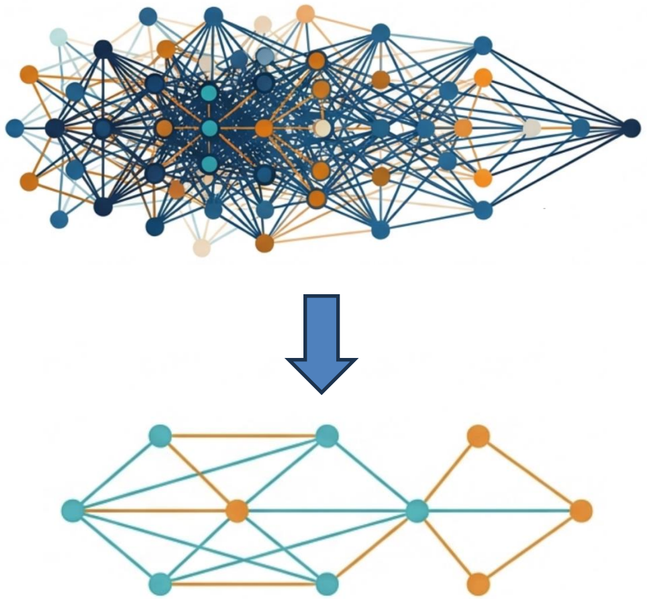
ICLR 26 submission with score 8,6,6,4



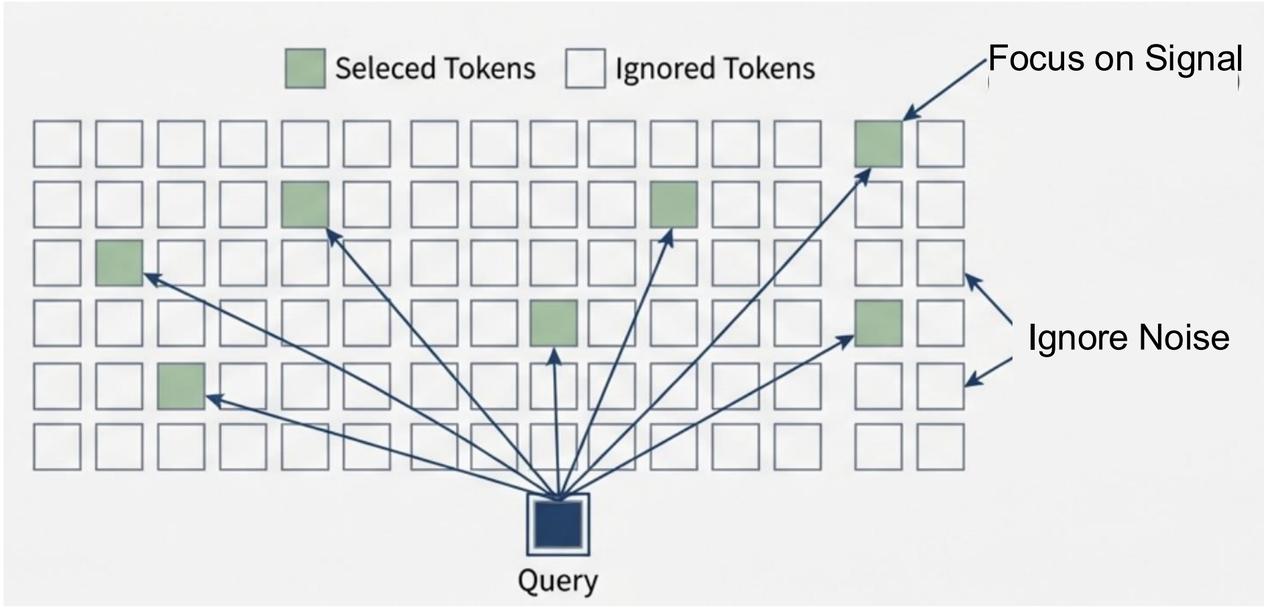

*Finished during the internship in UCSD, advised by Prof. Zhuowen Tu*

# Why use Sparse Attn?

**Token Compression**

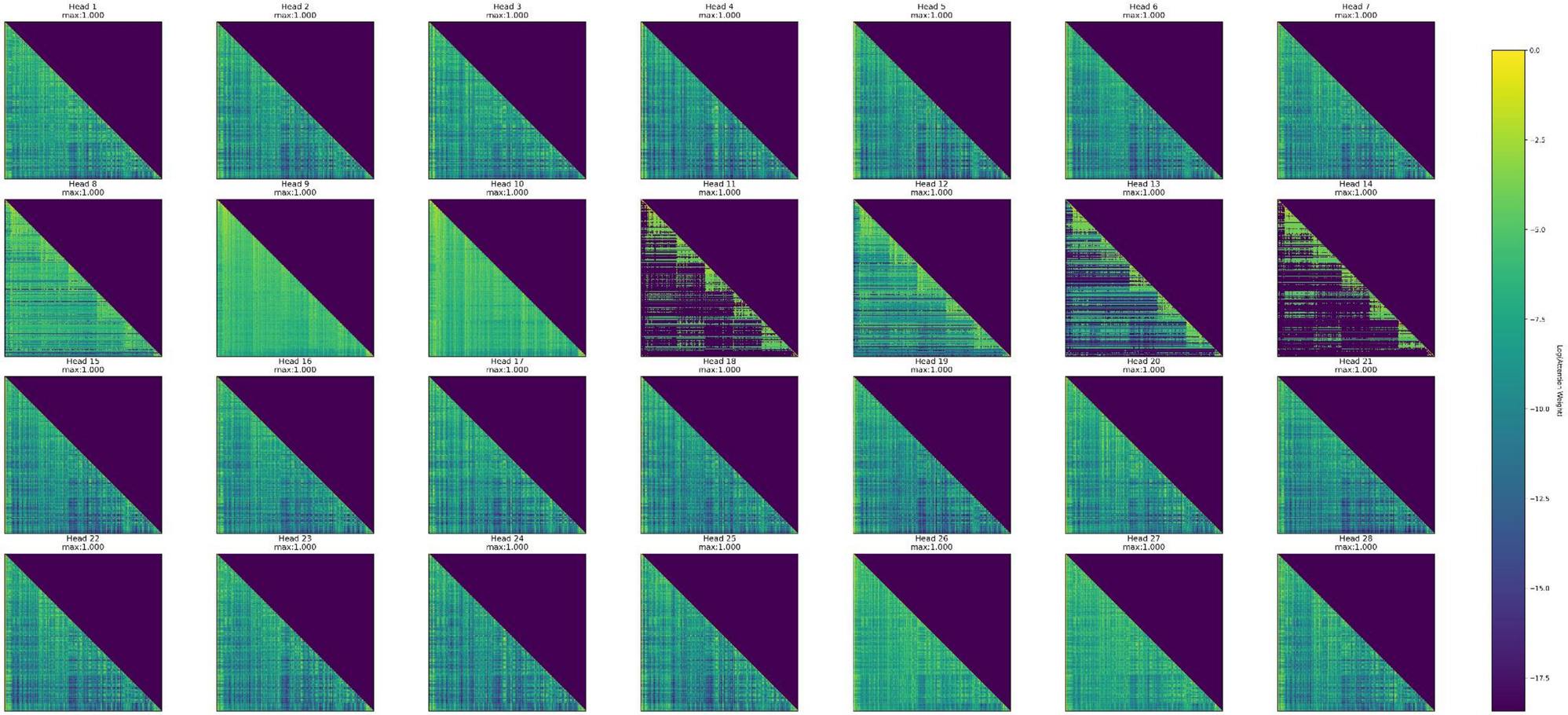


**Sparse Attention**



# Observation

Layer 28, 300 tokens

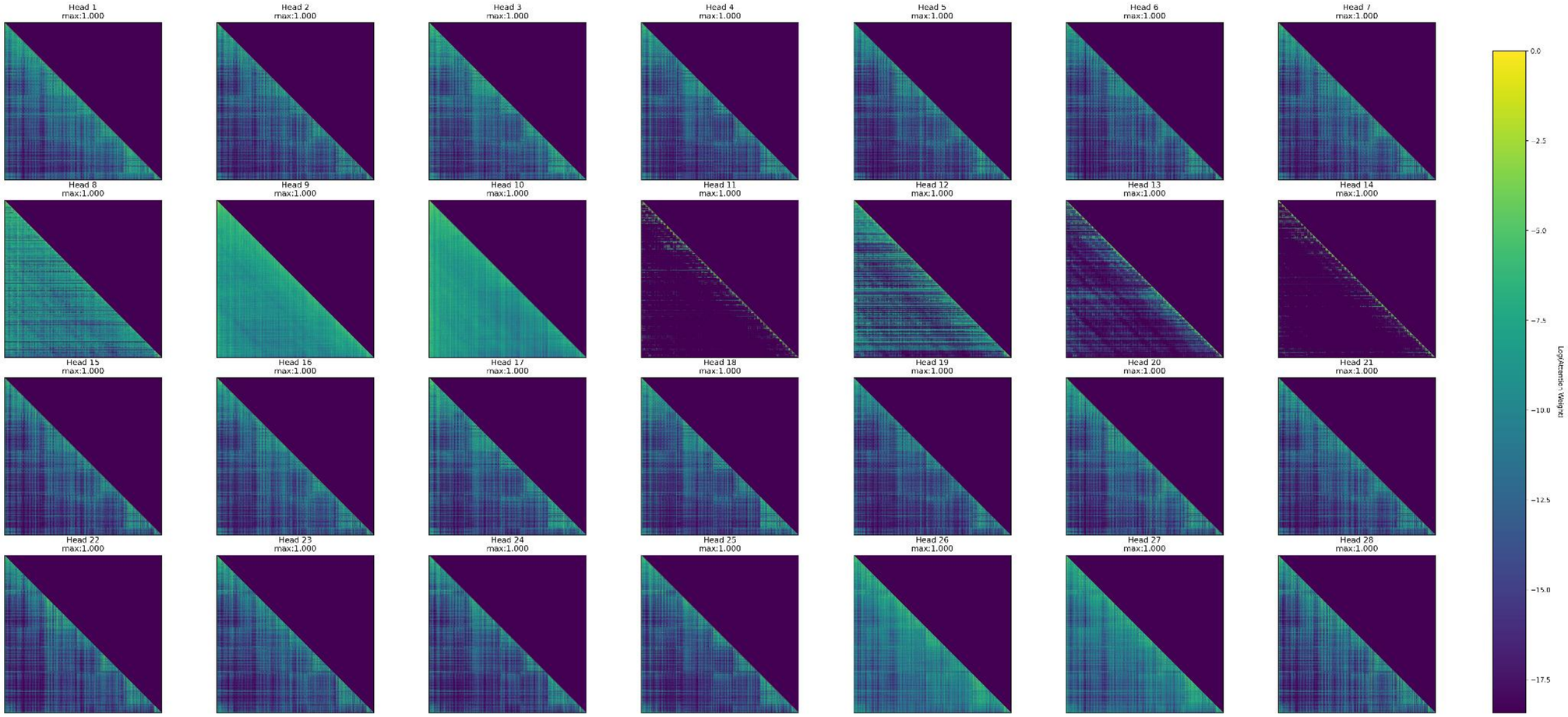



Sparse Attention Map

Unique Attention Pattern

# Observation

Layer 28, 3000 tokens

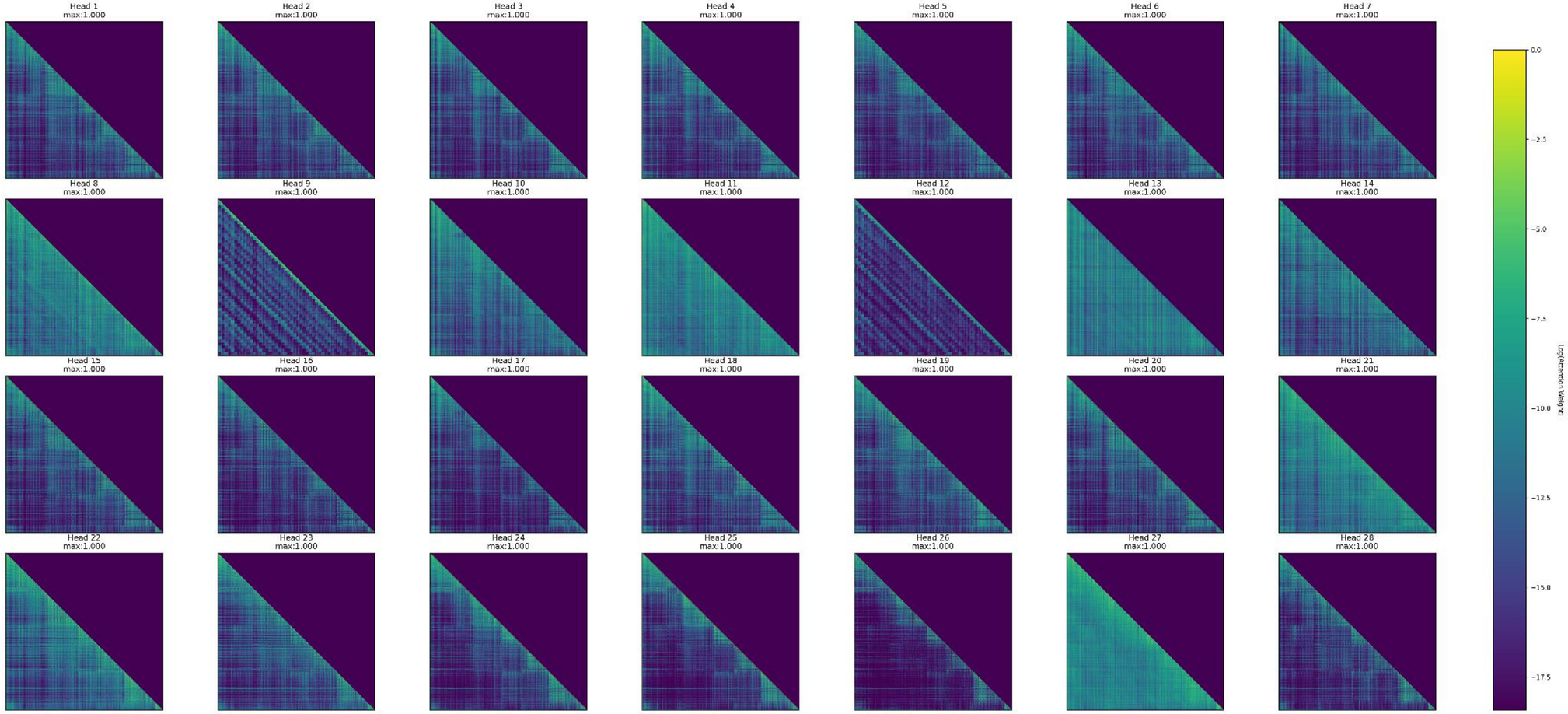



Sparse Attention Map

Unique Attention Pattern

# Observation

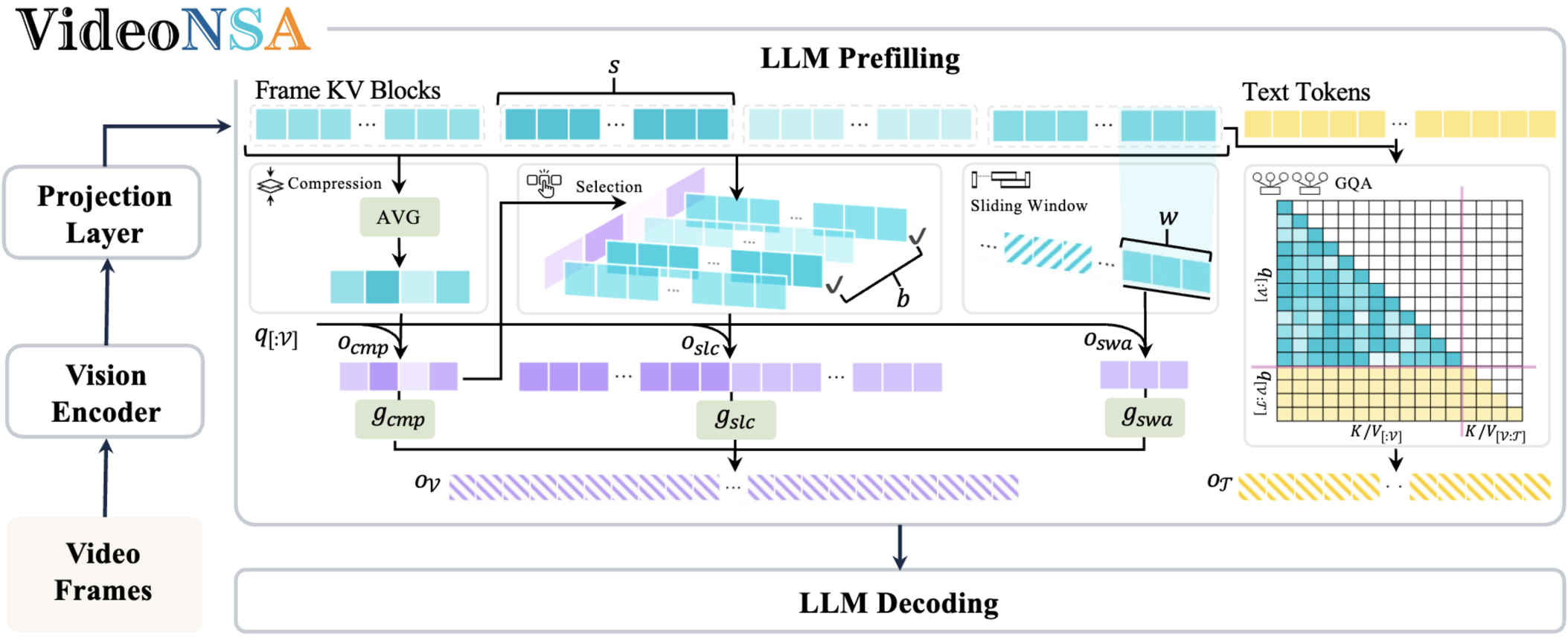
Layer 27, 3000 tokens




Sparse Attention Map

Unique Attention Pattern

# Efficient Video Understanding



# Training Recipe

| | |
|---|---|
| Base Model | Qwen2.5-VL-7B |
| Training Recipe | End-to-end |
| Dataset | 216K QA pairs of LLaVA-Video-178K |
| | Sampling at 4fps, remain 350-550 frames, max pixel = 50,176 |
| Model Setting | Block Size = 64, Block Count = 32, Sliding Window Size = 256 |
| Training Time | 4600 H100 GPU hours |

# 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

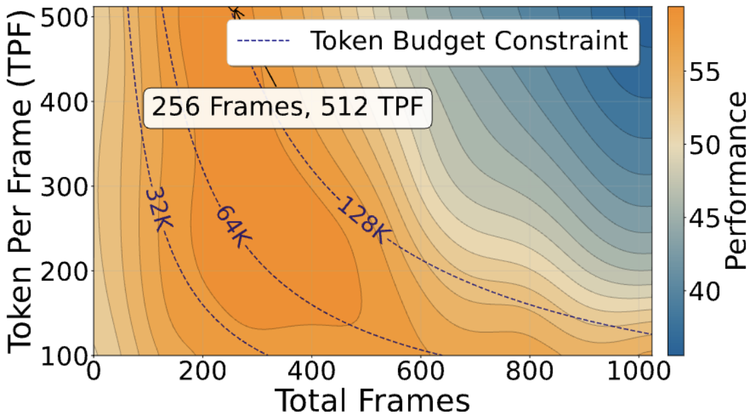
**Do learned sparse attention weights remain beneficial in dense attention settings?**

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

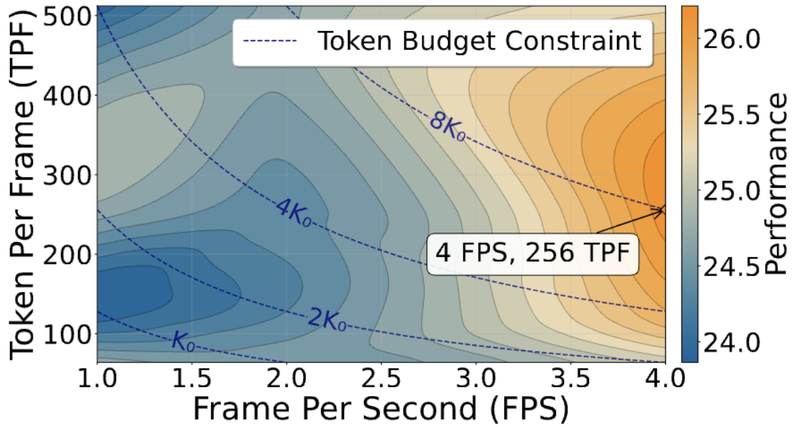
| Model | Long Video Understanding | | | | Temporal Reasoning | Spatial Understanding |
|---|---|---|---|---|---|---|
| | LongVideoBench | $\text{MLVU}_{Test}$ | TimeScope | LongTimeScope | Tomato | VSIBench |
| Qwen2.5-VL-7B | 58.7 | 51.2 | 81.0 | 40.7 | 22.6 | 29.7 |
| Dense-SFT | 57.8 (-1.5%) | 51.2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

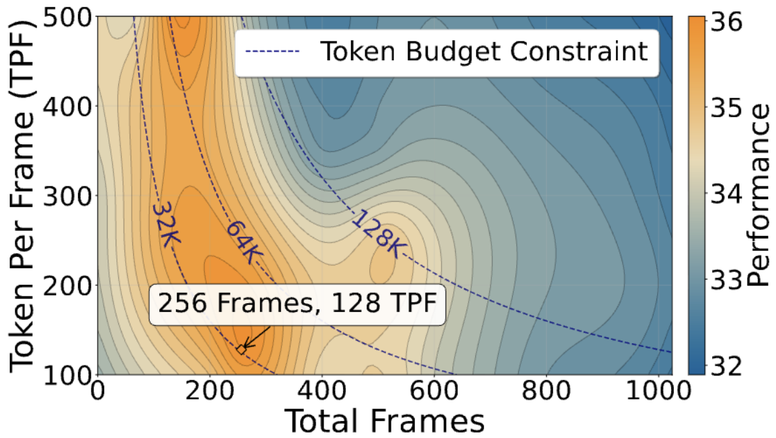
**How far can VideoNSA scale in context length?**




(a) Information Scaling of LongVideoBench



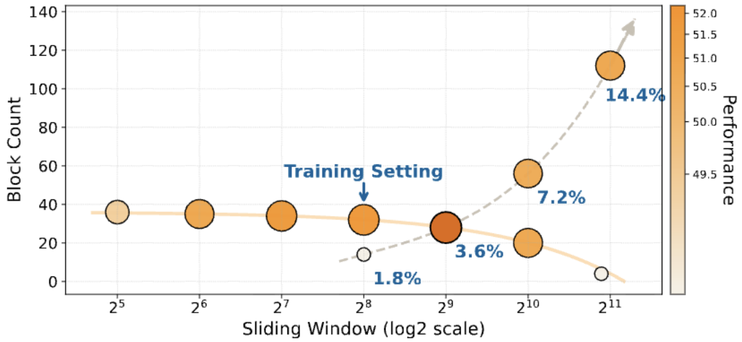

(c) Information Scaling of Tomato



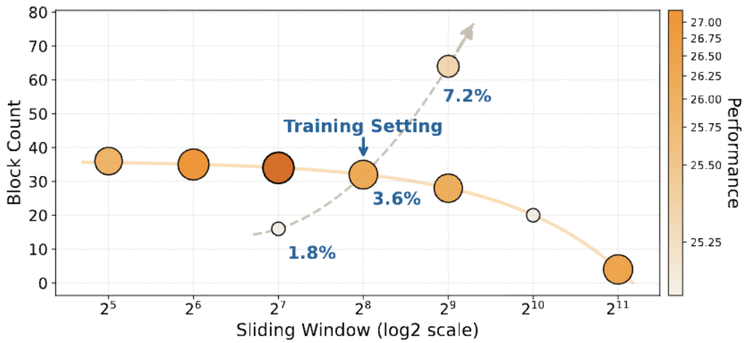

(d) Information Scaling of VSIBench

# VideoNSA

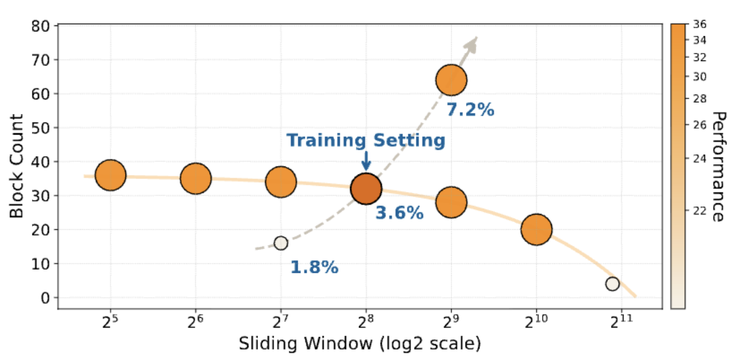
**How to allocate the attention budget?**




(a) Attention Scaling of MLVU



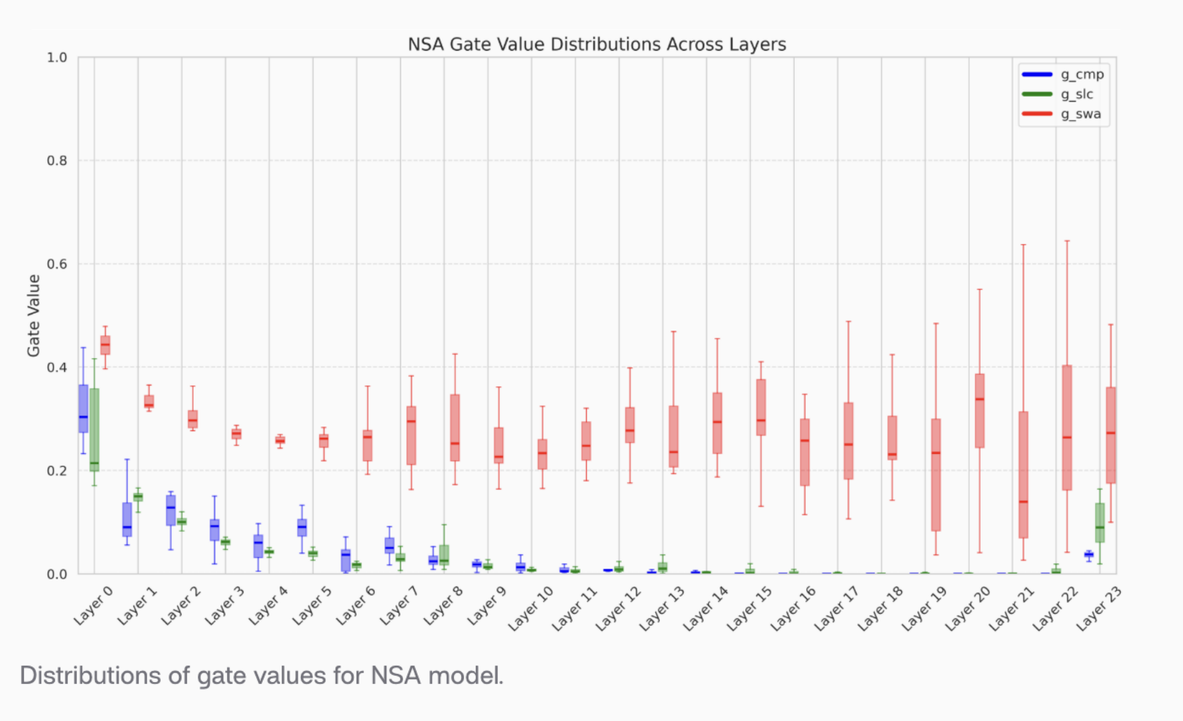

(c) Attention Scaling of Tomato



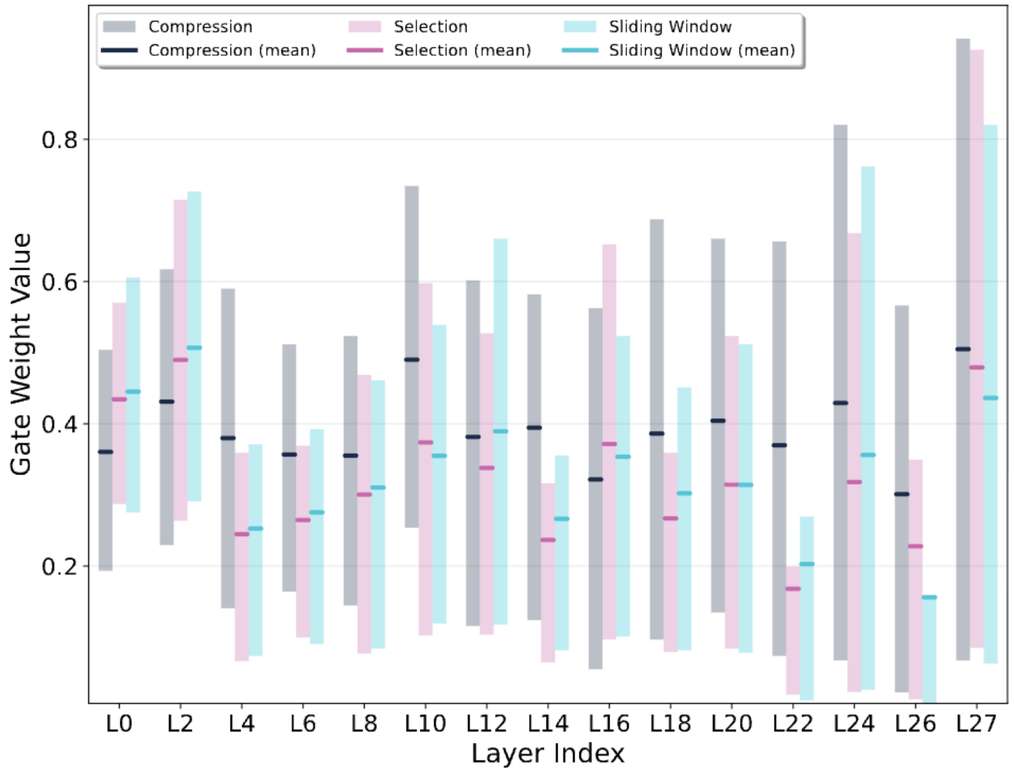

(d) Attention Scaling of VSIBench

# VideoNSA

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

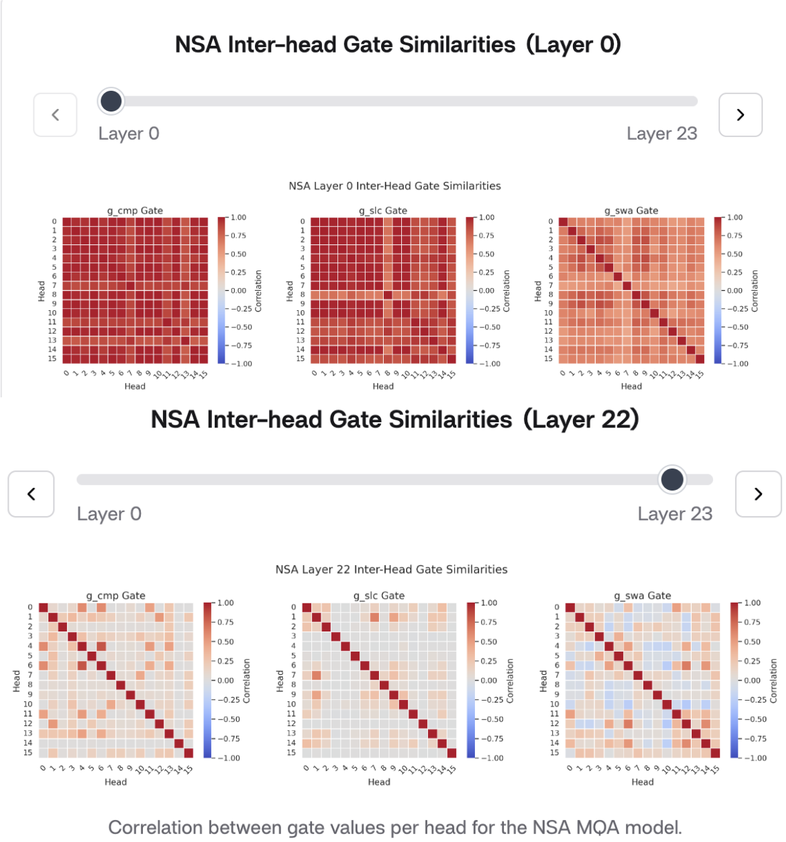



Distributions of gate values for NSA model.

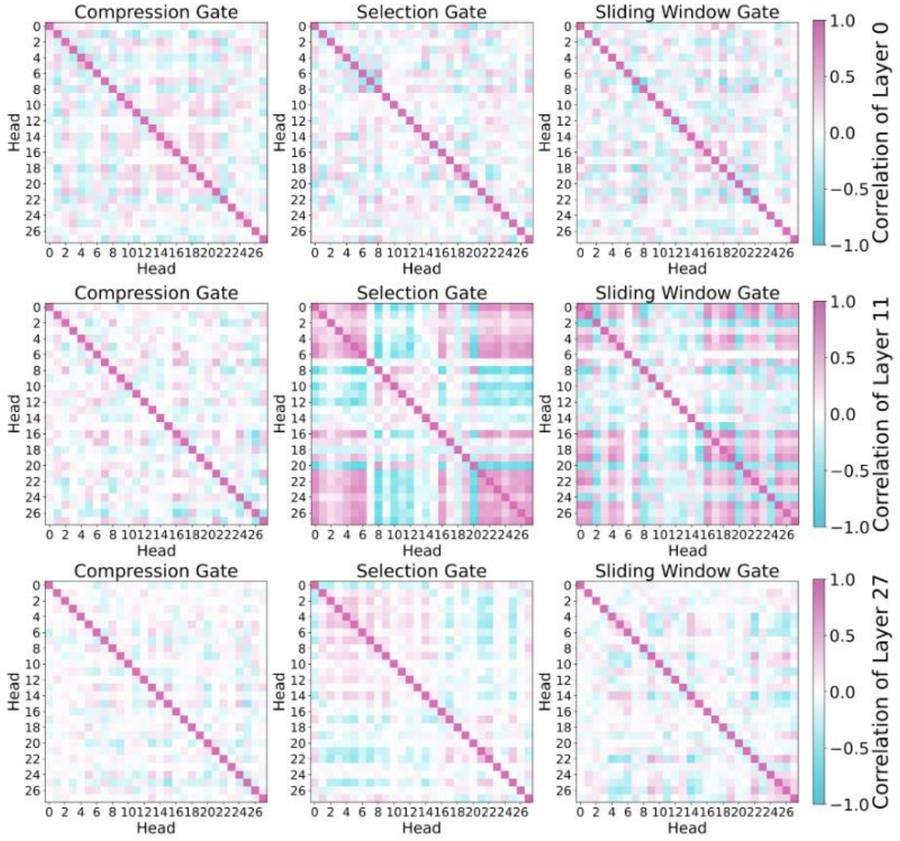


# VideoNSA

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

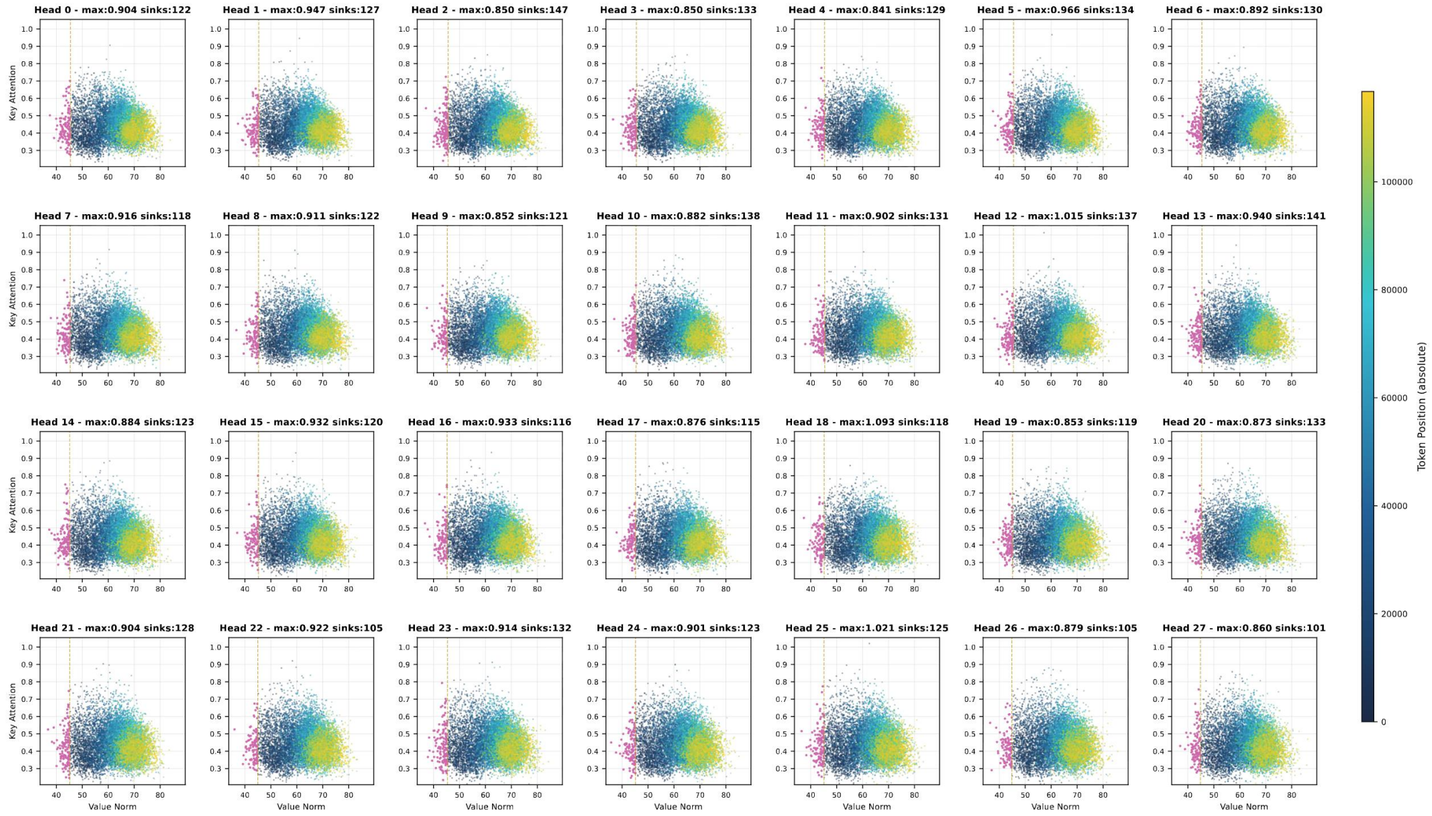



Correlation between gate values per head for the NSA MQA model.



# VideoNSA

## Do learnable sparse mechanisms induce dynamic attention sinks?



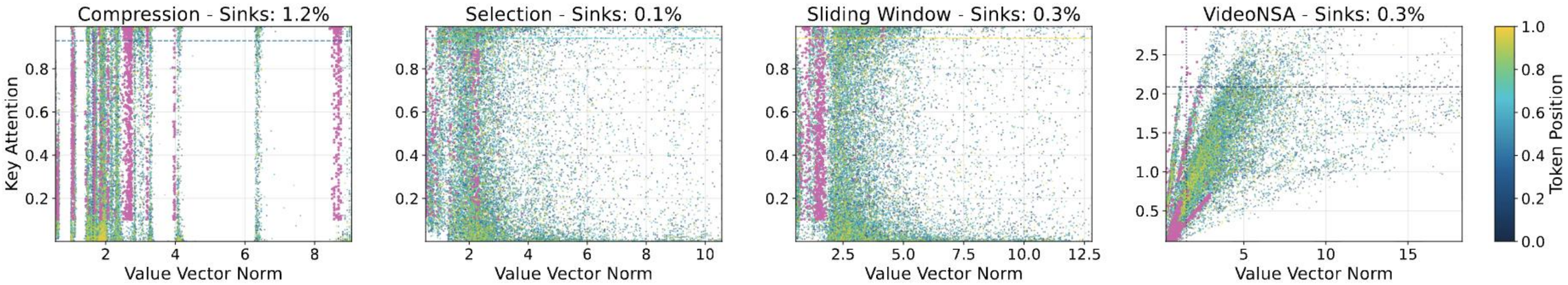

Attention Sink Distribution of Layer 14 in 128K Flash Attention

In decoder-only transformers, a disproportionate amount of attention is often allocated to the first few tokens, which act as attention sinks and absorb excessive attention mass as a byproduct of softmax normalization.

Prior studies show that attention sinks arise from massive activations and unusually small key and value norms, so attention directed to these tokens contributes little to the residual state.

# VideoNSA

**Do learnable sparse mechanisms induce dynamic attention sinks?**



# VideoNSA

**Do learnable sparse mechanisms induce dynamic attention sinks?**

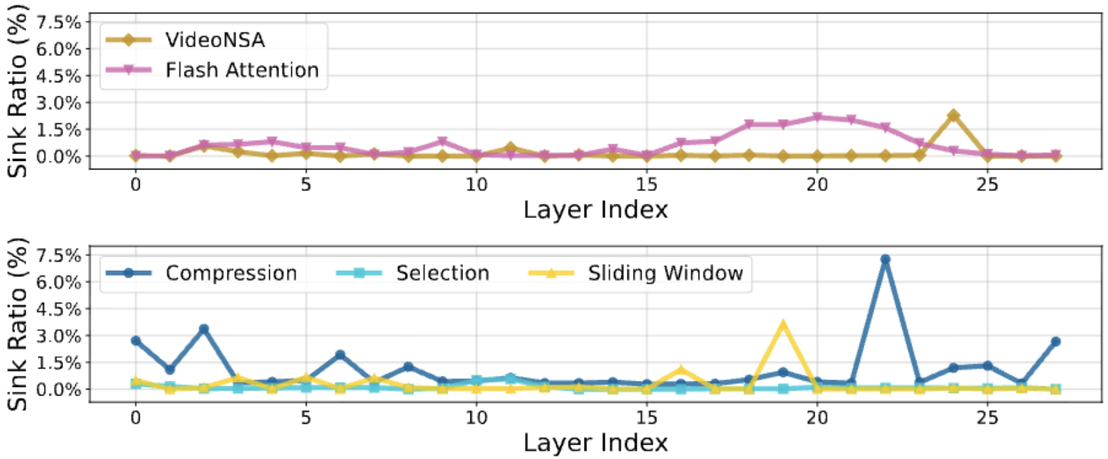



Figure 8: Layer-wise attention sink ratio distribution in different branches and Flash Attention.

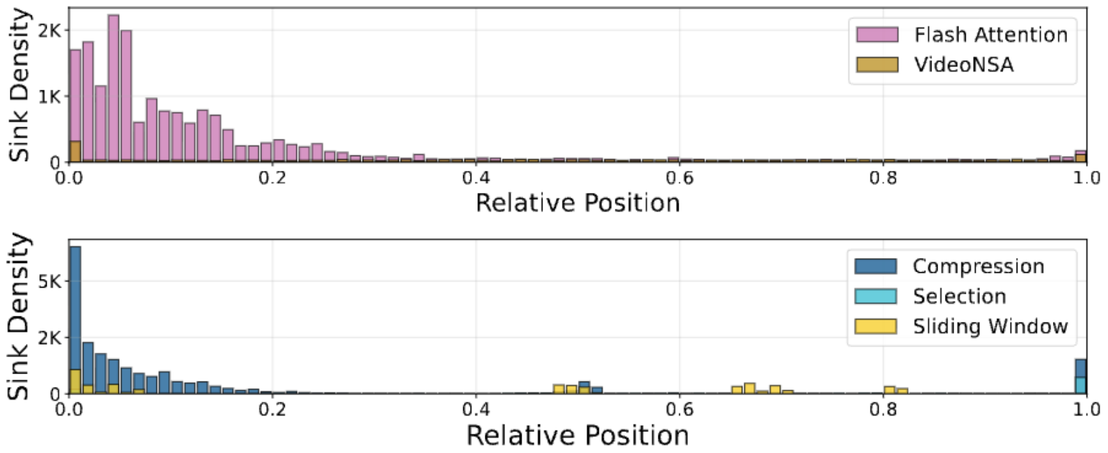



Figure 9: Relative positions of attention sinks in different branches and Flash Attention.

# 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

# Shortage & Improvement

- **Currently, we limit the frame resolution and sampling rate.**

  It would be better to train with dynamic settings based on the video content.

- **The speed and GPU cost do not have significant advantage.**

  Qwen2.5-VL 7B has 28 query head and 4 KV head, not suitable for Triton.

  Copy and padding requires additional costs.

  More efficient kernel brings better utilization **(e.g. Scalable-Flash-Native-Sparse-Attention is 6X faster than flash attn)**

# Research Interest

## Long-context Encoding

1. How an AI system can perceive and understand extremely long multimodal contexts, such as an entire day of human activity or the full history of a project.
2. Currently we do compression, selection and sliding window. What is Next?
3. How we process the conflict between training time and test time?

## Long-context Decoding

1. How an AI system should evolve its internal state over time to generate coherent long-horizon multimodal trajectories？
2. Why do models drift or contradict themselves even with sufficient context?
3. How to generate effective long-context multimodal outputs？

## Reliable Evaluation

1. How can we design reliable and persistent evaluation metrics for vision and multimodal models, analogous to perplexity (PPL) in NLP?
2. How can we build dynamic benchmarks that evolve over time to prevent memorization and overfitting to static test sets?

# Graduate Plan

**1.Research Direction**

- Try research on **long-context generation** during the PhD.

**2.Research Output**

- Aim to submit 1–2 papers per year during the PhD.
- Start writing blogs to organize research thoughts and findings.

**3.Open-source Project**

- Build a large open-source project, such as a training or evaluation framework, to support long-context research and reproducible experiments.

**4.Internships**

- leverage industry resources to work on real and meaningful problems, and use these experiences to clarify what I want to work on next

# Questions

1. How can physical video help improve LLMs ability on physics?
2. In your view, what kinds of research questions or work can only or best be done in academia? And how do you see the current trend of PhD students leaving academia for industry roles? Why you choose academia?

---

1. After my first year in the PhD program, when you look back, what concrete outcomes or progress would make you feel that admitting me was the right decision?
2. Are there kinds of work that you generally discourage students from doing? Is that more about problem choice, methodology, or how the work is framed?
3. Computing Recourse
4. Funding
5. Graduation requirements / Mentoring Style

# THANK YOU