SpargeAttention

Accurate and Training-free Sparse Attention Accelerating Any Model Inference

Research paper reading

Zhang, Xiang, Huang, Wei, Xi, Zhu and Chen

ICML 2025 | arXiv:2502.18137v8

Reading: 18 September 2026

Selective compression makes sparse prediction safer

- Predict block importance from representative tokens only when within-block similarity is high.
- Protect dissimilar blocks; then skip negligible PV updates inside online softmax.
- Combine sparsity with SageAttention quantization. Mochi on L40: 1897 s to 1037 s ($1.83\times$).
- Evidence spans language, image and video models; it is not a universal accuracy guarantee.

Paper, Secs. 3–4; Table 2, PDF p.8.

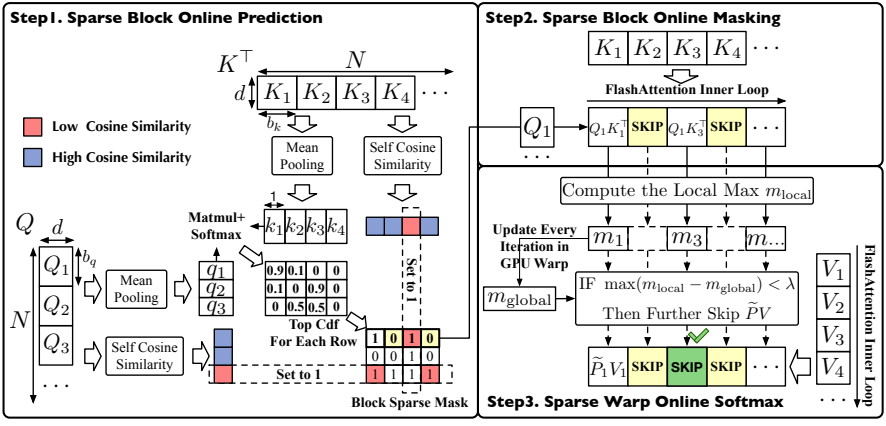
The mask must cost less than the work it saves

$$O = \mathrm{softmax}(QK^\top/\sqrt{d})V, \qquad Q, K, V \in \mathbb{R}^{N\times d}.$$

- Dense attention performs two quadratic matrix products.
- Fixed windows and attention sinks need not transfer to visual models.
- Pooling every block can hide a rare but important token.
- SpargeAttn asks whether a block is safe to summarize before trusting its mean.

Paper, Secs. 1–3.2, PDF pp.1–5.

Two filters skip different pieces of work



Paper, Fig.3, PDF p.4. Original author figure; red blocks are protected.

Online softmax maintains a running normalization

$$
\begin{aligned}
S_{ij} &= Q_i K_j^\top / \sqrt{d}, \\
m_{ij} &= \max(m_{i,j-1}, \operatorname{rowmax} S_{ij}), \\
\widetilde{P}_{ij} &= \exp(S_{ij} - m_{ij}), \\
l_{ij} &= e^{m_{i,j-1} - m_{ij}} l_{i,j-1} + \operatorname{rowsum} \widetilde{P}_{ij}, \\
O_{ij} &= \operatorname{diag}(e^{m_{i,j-1} - m_{ij}}) O_{i,j-1} + \widetilde{P}_{ij} V_j.
\end{aligned}
$$

Q_i: query block; K_j, V_j: key/value blocks. m, l: row-wise states.
Final output: $O_i = \operatorname{diag}(l_{i,T_n})^{-1} O_{i,T_n}$.

Paper, Sec.3.1, Eq.(1), PDF p.3.

Stage 1: protect blocks that are hard to summarize

$$q_i = \text{mean}(Q_i), \quad k_j = \text{mean}(K_j), \quad \hat{S} = qk^\top.$$

- Compute within-block scores s_{qi}, s_{kj}. Exclude low-similarity key columns from compressed softmax.
- Apply TopCdf to each compressed probability row, controlled by τ.
- Force $M_g[i,:] = 1$ if $s_{qi} < \theta$; force $M_g[:,j] = 1$ if $s_{kj} < \theta$.
- If $M_g[i,j] = 0$, skip both $Q_i K_j^\top$ and $\widetilde{P}_{ij} V_j$.

Paper, Secs.3.2–3.3, Eqs.(4)–(6), PDF pp.3–5.

Two definitions require care when reproducing the paper

TopCdf boundary

Text says cumulative mass "reaches" τ; pseudocode uses $\mathrm{cumsum} \leq \tau\,\mathrm{sum}$. For $(0.6, 0.3, 0.1)$ and $\tau = 0.8$, the code keeps only 0.6.

Similarity definition

The paper calls this cosine similarity but prints $\mathrm{CosSim}(X) = \mathrm{mean}(XX^\top / |\max(XX^\top)|)$. This differs from standard pairwise cosine normalization.

Check the implementation before resolving these ambiguities.

Paper, Sec.3.2, PDF pp.4–5. Example and caution: reading analysis.

Stage 2: skip PV when the local maximum is far below

For query rows I_w handled by one GPU warp:

$$\max\left(m_{\text{local}}[I_w] - m_{ij}[I_w]\right) < \lambda, \qquad \lambda < 0.$$

- Then every unnormalized weight in this warp is below e^{λ}.
- The running maximum stays unchanged; the previous output needs no rescaling.
- Skip $\widetilde{P}_{ij}[I_w]V_j$. QK and denominator updates have already happened.
- Small weights alone do not bound total output error for arbitrary V.

Paper, Sec.3.4 and Algorithm 1, PDF pp.4–5; last point: reading analysis.

Training-free still involves calibration

$$L1 = \frac{\sum |O - O'|}{\sum |O|}.$$

- Use five model inputs; search (τ, θ) under $L1 < l_1$, then λ under $L1 < l_2$.
- (l_1, l_2): Llama $(.08, .09)$; CogVideoX/Mochi $(.05, .06)$.
- Flux/SD3.5 $(.07, .08)$; Open-Sora-Plan $(.03, .035)$.
- These constrain attention output error on calibration inputs, not downstream quality on all inputs.

Paper, Secs.3.6 and 4.1, PDF pp.5–6.

Quantization contributes to the reported speedup

- The main experiments build on SageAttention with block quantization in the QK path.
- CUDA implementation and kernel fusion reduce prediction overhead.
- A speedup over dense attention combines sparsity, quantization and implementation effects.
- The p.1 footnote describes a newer SageAttention2 implementation; it does not replace every table with new measurements.

Paper, Sec.3.5, Algorithm 1 and p.1 footnote, PDF pp.1,4–5.

Hilbert ordering exposes visual locality

- Flatten visual tokens in $T \times H \times W$ along a Hilbert curve to improve within-block similarity.
- Permute Q/K/V after positional embedding; apply the inverse permutation to the output.
- For joint visual-language attention, permute visual tokens only.
- Appendix ablation: query blocks 128, key blocks 64; FlashAttention2 is the accuracy reference.
- Masks and position-dependent biases must remain consistent under permutation (reading analysis).

Paper, Sec.3.7 and Appendix A.1, PDF pp.6,14.

Evaluation tests transfer across three model families

Task	Models	Evaluation
Language	Llama3.1-8B	WikiText, LongBench, InfiniteBench En.MC, NIAH
Image	Flux.1-dev, SD3.5 large	COCO annotations
Video	CogVideoX-2B, Mochi, Open-Sora-Plan	open-sora prompts

- Quality reference: full attention. Sparse baselines: block-sparse MInference and FlexPrefill.
- Table 1 uses different achieved sparsities: compare quality and speed together.

Paper, Sec.4.1 and Table 1, PDF pp.6–7.

Speed, sparsity and quality have different denominators

- Speed is defined as dense operation count divided by attention latency, including mask prediction. It is not tokens/s.
- Sparsity counts skipped QK and PV block products together; it is not the fraction of zero entries in P.
- PPL and FID: lower is better. Task scores, CLIP/IR and reported video scores: higher is better.
- Table 1 speed uses the real task tensors; Llama speed/sparsity specifically use 128K NIAH.

Paper, Sec.4.1 and Table 1 caption, PDF pp.6–7.

Language quality is close while attention gets faster

Llama3.1	Full	SpargeAttn
Attention speed (paper units)	156.9	708.1
WikiText PPL ↓	6.013	6.020
LongBench ↑	38.682	39.058
InfiniteBench ↑	0.6594	0.6638
NIAH ↑	0.907	0.909

Attention speed: 4.51× at 128K; sparsity: 0.54.
Small score improvements have no reported confidence intervals.

Paper, Table 1, PDF p.7. Speed ratio calculated from table values.

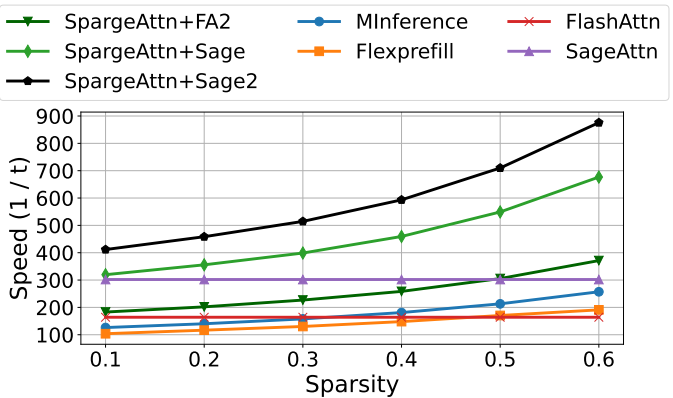
Visual metrics are largely retained, but not identical

Model	Metric	Full	SpargeAttn
CogVideoX	VQA-a ↑	80.384	78.276
Mochi	VQA-a ↑	56.472	54.179
Mochi	FScore ↑	1.681	1.807
Open-Sora-Plan	VQA-t ↑	80.60	76.91
Flux	FID ↓	166.103	163.982
SD3.5	FID ↓	166.101	166.193

"No quality loss" is the authors' summary, not literal equality.
Reported sparse baselines often degrade much more.

Paper, Table 1, PDF p.7. Selected metrics, original scales.

Kernel curves separate sparse and quantized variants



Paper, Fig.10, PDF p.9. Sequence 22K; head dimension 128. Not end-to-end speed.

End-to-end gains are smaller than attention gains

Model	GPU	Original	Sage	Sparge
CogVideoX	RTX4090	87 s	68 s	53 s
Mochi	L40	1897 s	1544 s	1037 s
Llama 24K	RTX4090	4.01 s	3.53 s	2.60 s
Llama 128K	L40	52 s	42 s	29.98 s

Mochi: 1.83× vs. original, 1.49× vs. SageAttention.
Non-attention work limits whole-model acceleration.

Paper, Table 2, PDF p.8. Ratios calculated from reported latency.

Prediction overhead shrinks relative to full attention

Length	Prediction (ms)	Full attention (ms)	Ratio
8K	0.251	6.649	3.78%
32K	0.972	106.68	0.911%
128K	8.764	1696.2	0.516%

- The absolute prediction time still grows.
- These ratios use full-attention time, not accelerated total time.
- Compressed attention remains quadratic in the number of blocks.

Paper, Table 3, PDF p.8; complexity interpretation: reading analysis.

The similarity guard prevents damaging outliers

Mochi	Without guard	With guard
VQA-a ↑	34.664	54.179
VQA-t ↑	44.722	67.219
FScore ↑	1.138	1.807

Appendix: on a selected ∼2% high-difference subset, CogVideoX L1 is 0.214 without vs. 0.0843 with the guard.

This selected subset is not the population average.

Paper, Tables 5 and 10; Appendix A.2, PDF pp.9,14–15.

Hilbert ordering trades a little accuracy for sparsity

Mochi	Sim-q	Sim-k	L1 ↓	Sparsity
Random	.321	.019	.0414	.048
Row-major	.551	.390	.0307	.363
Hilbert	.572	.479	.0389	.392

Row-major to Hilbert: sparsity increases by 2.9 percentage points.
Random ordering retains accuracy by falling back to more computation.

Paper, Tables 4 and 9, PDF pp.9,14. Ablation settings differ from Table 1.

The two stages overlap; longer contexts expose more sparsity

Llama3.1, 128K NIAH

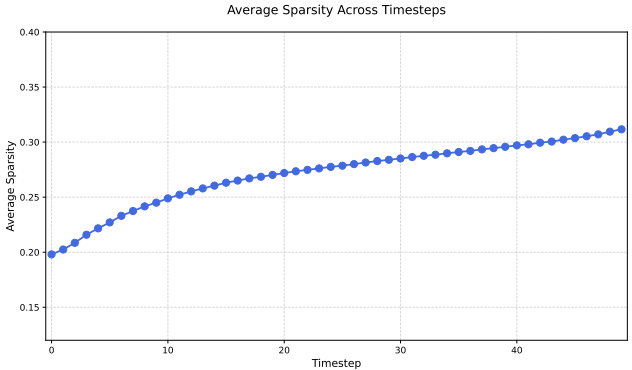
Only M_g: 51.2%; only M_{pv}: 27.7%; together: 54%. The incremental gain over stage 1 is 2.8 percentage points.

Length	8K	16K	24K	48K	128K
Sparsity	6.8%	26.4%	35.7%	49.8%	54%

These are measured trends under the paper's accuracy constraint, not a guarantee for every longer input.

Paper, Tables 6–7, PDF p.9.

Diffusion sparsity varies over sampling steps



Layer and head variation is also substantial; prompt averages vary less.

Paper, Appendix A.4, Figs.14–17, PDF pp.17–18. Shown: original Fig.15.

What the evidence supports, and what remains open

- Supported: selective protection plus two-stage skipping is practical across the tested model families.
- Calibration, quantization and CUDA implementation are part of the result.
- No arbitrary-input error theorem; some quality metrics decline; repeated-run uncertainty is not reported.
- Reproduction needs exact threshold semantics, kernel versions, baseline settings and hardware details.
- Discussion: can omitted mass and value norms provide a stronger error control?

Paper, Secs.3–4 and Appendix; limitations and question: reading analysis.

Reference and scope of this reading

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, Jianfei Chen.

SpargeAttention: Accurate and Training-free Sparse Attention Accelerating Any Model Inference.
ICML 2025; arXiv:2502.18137v8, revised 19 November 2025.

`https://arxiv.org/abs/2502.18137v8`
`https://github.com/thu-ml/SpargeAttn`

Read from the version-matched PDF, author TeX and figures.
Experiments are reported by the paper, not independently reproduced.

Reading prepared 18 September 2026. Main talk: about 25–30 minutes.

Backup: medium-length retrieval and calibration gaps

- Appendix Table 11 evaluates NIAH across 16–28K; the speed row is labeled 24K.
- Full to Sparge: 156.9 to 443.6 TOPS; NIAH 0.838 to 0.863; sparsity 0.36.
- Section 4.1 says MInference 30%/70%, while main table labels are mostly 0.3/0.5; the discrepancy is unresolved.
- The full calibration grid, cost, sample sensitivity and head-wise configuration are not specified.

Paper, Table 11, PDF p.15; Secs.3.6,4.1,4.3, PDF pp.5–9.