Scaling Linear Attention with Sparse State Expansion

Paper by Yuqi Pan, Yongqi An, Zheng Li, Yuhong Chou, Ruijie Zhu,
Xiaohui Wang, Mingxuan Wang, Jinqiao Wang, and Guoqi Li

The idea

Increase recurrent memory capacity with shared projections,
and update only the partitions selected by each token.

arXiv:2507.16577v2 (1 Oct 2025) · Reading: 2 Nov 2026
Published at ICLR 2026 with "Capacity" added to the title.
Official code: conference supplement ZIP (model and kernels; external dependencies).

A 20–30 minute research discussion. Equations, figures, and results use arXiv v2.

The bottleneck is compressed memory capacity

	Softmax attention	Linear attention
Context state	$\{K_t, V_t\}$	$S_t \in \mathbb{R}^{c \times d}$
State elements / head	$2td$	cd
Per-token state work	grows with t	fixed in t
Main compromise	cache and access cost	lossy history compression

- $c = 128$: the same element count as a 64-token KV cache.
- More model parameters and more context state are distinct resources.
- Goal: enlarge state capacity with little projection-parameter growth.

Secs. 2–4, PDF pp. 3–6. The 64-token comparison is a storage count, not equal recall capacity.

A key vector decides where information is written

$$S_t = \Lambda_t S_{t-1} + \phi(k_t)^\top v_t$$

Manage history

Λ_t: decay, retention, or erasure.
GLA uses diagonal gating;
GDN uses a delta-rule transition.

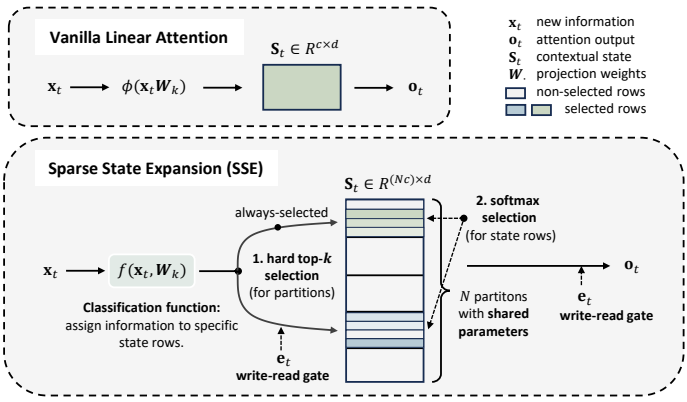
Allocate new information

$\phi(k_t)$: weights over state rows.
Interpret the key projection as a learned classifier of inputs.

A row is a learned feature bucket. Dense writes mix unrelated inputs; dense decay advances every bucket even when it is irrelevant.

Sec. 3, Eq. (6), PDF pp. 4–6. SSE experiments use GLA-style transitions.

SSE routes tokens into shared-parameter state partitions



Sec. 4.1, Fig. 4, PDF p. 7. Original author figure.

1. Hard top-k selection among N partitions.
2. Softmax allocation among rows inside each selected partition.
3. One always-active shared partition.

Inactive partitions keep their state unchanged.

One gate controls both writes and reads

$$e_t = \mathrm{softmax}(x_t W_e), \quad \mathcal{T}_t = \mathrm{TopKIndices}(e_t, k)$$

$$q_t = x_t W_q, \qquad v_t = x_t W_v, \qquad k_t = \mathrm{softmax}(x_t W_k)$$

$$S_t^i = \begin{cases} \Lambda_t S_{t-1}^i + e_t^i k_t^\top v_t, & i \in \mathcal{T}_t, \\ S_{t-1}^i, & i \notin \mathcal{T}_t, \end{cases} \qquad o_t = \sum_{i \in \mathcal{T}_t} e_t^i q_t S_t^i$$

- $S_t^i \in \mathbb{R}^{c \times d}$; $q_t, k_t \in \mathbb{R}^{1 \times c}$; $e_t \in \mathbb{R}^N$.
- Gate values carry gradients through writing and reading.
- Within-partition softmax is soft allocation, not another hard top-k.

Sec. 4.1, Eqs. (9)–(14), PDF pp. 7–8. Shared-partition contribution omitted here.

A memory decays only when its partition is visited

A local update clock

Dense recurrence: one decay per token.
Sparse recurrence: one decay per selected-partition update.

Teaching example	Decay events	Retained multiplier
Dense, 100 intervening tokens	100	$0.99^{100} \approx 0.366$
Partition selected 10 times	10	$0.99^{10} \approx 0.904$

Reading still requires routing back to the relevant partition.
Long idle retention does not protect against later interfering writes.

Derived illustration of Eq. (13); Appendix A, Proposition 4, PDF pp. 24–25. Not experimental data.

State storage and active computation scale differently

Store more

N routed states + 1 shared state
$\sim (N+1)cd$ elements / head

Touch fewer

k selected states + shared state
$\sim (k+1)cd$ state work / step

- `n4k1`: four routed partitions, select one; shared partition stays active.
- Shared QKV; rank-64 Q/K adaptation for the shared partition.
- Partition balance loss, coefficient 0.01.
- Router parameters and routing work still grow with N.

Secs. 4.1, 5.3; Appendix C, PDF pp. 8, 14, 27. Cost expressions are structural counts.

Varlen groups tokens while preserving causal order

Short contexts: masking

Replicate activations across partitions.
Mask inactive writes and reads.
Parallelize over batch/head axes.

Avoids many tiny subsequences that underfill chunk kernels.

Long contexts: varlen

Group by sample and partition.
Preserve time order within each group.
Build new sequence offsets.

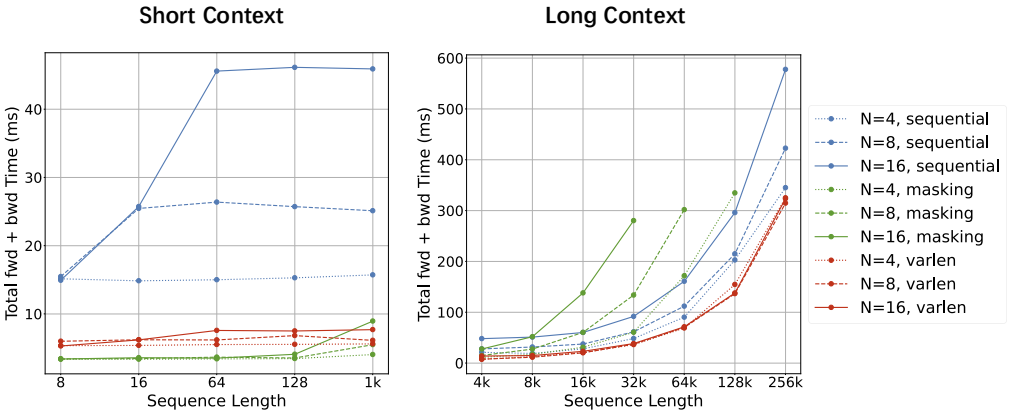
About kL routed token records; one batched operator call, then scatter and sum.

Masking must also freeze inactive decay.

Prefill uses chunks; decoding updates only selected recurrent states.

Sec. 4.2; Appendix B, Algorithms 1–2, PDF pp. 8–9, 26–27.

At fixed k, varlen scales favorably with partition count



Short sequences favor masking; long sequences favor varlen.

A100, 8 heads, dimension 128, $K = 1$, $N \in \{4, 8, 16\}$.

Fig. 6, PDF p. 9. Forward + backward operator time; not end-to-end Transformer speedup.

Each training budget answers a different question

Stage	Budget	Question
Basic pretraining	600M / 15B tokens; 2B / 100B tokens, 8K	Recall and language quality
Scale + extend	2B: 2T + 250B tokens, extend to 32K	Sufficient-data behavior
Reasoning training	80K distilled examples; 5 epochs; GRPO 230 steps	Reasoning capability
Transformer conversion	2B, about 100B tokens during 128K training	Preserve pretrained capability

MHA–SwiGLU, no convolution. SSE-H: 15 SSE + 3 softmax layers at 2B.

Sec. 5, PDF pp. 10–12. State-scaling ablations separately use 2B / 20B tokens.

Pure SSE improves recall; hybrid SSE closes more of the gap

2B / 100B tokens	Wiki.↓	Commonsense↑	Recall↑
Transformer	16.46	53.55	73.00
GLA	21.30	49.13	49.29
GDN	17.08	54.05	53.79
SSE-n4k1	**16.92**	**54.57**	**61.46**
GLA-H	17.27	53.15	67.52
SSE-H-n4k1	**16.47**	**54.48**	**70.87**

SSE gains 7.67 percentage points in recall over GDN.

The Transformer recall gap is 11.54 points for SSE, 2.13 for SSE-H.

Table 1, PDF p. 10. Recall: mean of FDA, SWDE, SQuAD. Commonsense: six-task mean.

Harder 8K retrieval tasks reveal the memory advantage

2B / 100B tokens	S-NIAH-1	S-NIAH-2	S-NIAH-3
Transformer	100.0	100.0	100.0
GLA	87.6	23.2	16.2
GDN	100.0	8.2	9.0
SSE-n4k1	**100.0**	**85.2**	**62.2**
GLA-H	100.0	100.0	70.8
SSE-H-n4k1	**100.0**	**100.0**	**97.4**

Pure SSE still loses accuracy as context grows.

Sparse state expansion and full-attention layers provide complementary gains.

Table 2, PDF p. 11. Zero-shot RULER, 8K columns; all scores in percent.

After 2T-token training, task-level gains remain mixed

Broad evaluation after 32K extension

Ten-task average: 45.6 vs. 45.1
SSE-H vs. Transformer

- MMLU: 54.5 vs. 52.6
- C-Eval: 59.7 vs. 55.9
- GSM8K: 49.2 vs. 50.6

Selected RULER tasks at 32K

S-NIAH-3: 90.2 vs. 67.4
MQ-NIAH: 44.6 vs. 28.6
MV-NIAH: 40.6 vs. 18.1

S-NIAH-1 is lower: 93.6 vs. 99.4.
Multi-query/value recall remains difficult in absolute terms.

Table 3, PDF p. 11; Table 11, PDF p. 28. 2T pretraining + 250B extension tokens.

Conversion preserves more retrieval than GLA-H

2B model	Quest	Qampari	Table Query	Average
Transformer	49.1	71.5	60.6	60.4
Converted GLA-H	25.6	56.6	48.2	43.5
Converted SSE-H	35.6	70.6	50.0	52.1

- Conversion during 128K training, using about 100B tokens.
- Evaluation at 32K: +8.6 points over GLA-H, −8.3 vs. Transformer.
- Qampari is largely preserved; Quest still loses 13.5 points.

Sec. 5.1, Table 4, PDF pp. 11–12. SSE-H uses n4k1.

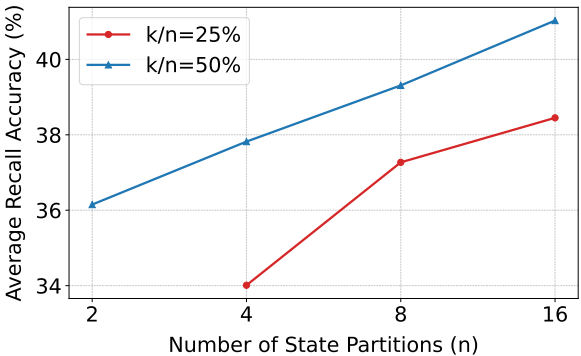
Matched training supports comparable mathematical reasoning

Model	AIME24	AIME25	MATH500	Olympiad	AMC23
Qwen3-1.7B Thinking	48.3	36.8	93.4	–	–
R1-Distill-Qwen-7B	55.5	39.2	92.8	–	–
Authors' Transformer-2B	64.1	52.8	93.0	83.3	92.0
Authors' SSE-H-2B	**64.5**	**50.2**	**92.1**	**85.7**	**91.4**

- AIME gains over Qwen3: +16.2 / +13.4 points; training differs.
- Matched Transformer difference: +0.4 / −2.6 points.
- AIME uses avg@32, not pass@32; temperature 1.0, top-p 0.7.

Sec. 5.2, Table 5, PDF p. 12. "Olympiad" abbreviates OlympiadBench.

State scaling improves recall while active work also grows



- 2B models, 20B tokens.
- More partitions improve recall at fixed k/N.
- Larger k/N also improves recall in the plotted range.
- Fixed ratio means k grows with N. This is not the fixed-K runtime comparison.

Sec. 5.3, Fig. 7, PDF pp. 12–13. Original figure; limited configurations, not a fitted scaling law.

Softmax keys and shared parameters both improve recall

600M / 15B tokens unless noted	Recall mean	Change
SSE with SiLU keys	24.23	reference
SSE with softmax keys	31.16	+6.93
SSE with shared parameters	31.16	reference
Independent parameters, 890M total	25.95	−5.21

- More independent parameters do not replace useful memory organization.
- Unshared non-embedding parameters rise from 300M to 580M.
- Commonsense changes only from 42.91 to 42.99 in that comparison.

Tables 6 and 8, PDF pp. 13–14. All differences are percentage points.

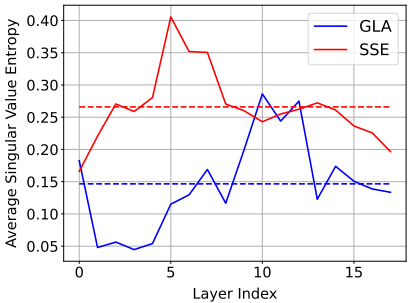
The gate works best when used on both sides

SSE gating ablation, 2B / 100B tokens	Recall mean	Commonsense
Write + read	**56.63**	**53.63**
Neither	51.87	53.01
Write only	50.92	52.82
Read only	53.96	52.91

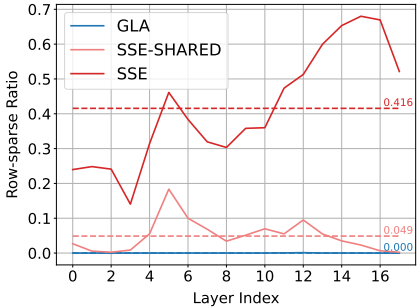
- Joint write/read weighting beats either isolated use.
- The full row here is 56.63; Table 1 reports 61.46.
- v2 does not explain that difference: compare variants within this table.

Table 7, PDF p. 14; Table 1, PDF p. 10.

State diagnostics support more selective memory use



Energy spread over singular directions is higher on average for SSE.



Key values below 10^{-5}: routed 41.6%, shared 4.9%, GLA $\sim$0.

Near-zero key entries are not exact hard zeros.

Representation diversity and retrieval accuracy are separate measurements.

Figs. 5 and 8, PDF pp. 8, 13. Original figures.

SSE makes state capacity an explicit design variable

1. **Mechanism:** route information, pause irrelevant decay, and share parameters across larger recurrent memories.
2. **Evidence:** stronger retrieval than linear baselines; hybrid models retain competitive post-trained reasoning.
3. **Tradeoff:** larger stored states and routing work; hybrid layers still require context-growing KV caches.

Questions for follow-up

Fixed-compute state scaling; state-aware routing;
SSE combined with delta-rule transitions; checkpoint-based state expansion.

Secs. 5–6, PDF pp. 10–14. SSE-on-GDN is future work in v2; GDN is already a baseline.

Paper and implementation sources

- Yuqi Pan et al.
  Scaling Linear Attention with Sparse State Expansion.
  arXiv:2507.16577v2, revised 1 October 2025.
- ICLR 2026 official proceedings:
  Scaling Linear Attention Capacity with Sparse State Expansion.
- Author implementation in the official supplement:
  SSE-H modeling/configuration and masking kernel code.
  External project dependencies remain; not a standalone training package.

Publication/code checked on 2 November 2026.
Figures are author originals; tables select and retypeset cited results.

Backup: the classification argument is geometric

$$C^i = \{x : xW_k^i > k_{th}^i\}, \qquad \|x\|^2 = d$$

$$\theta^i = \arccos \frac{k_{th}^i}{\sqrt{d}\|W_k^i\|} \quad \Longrightarrow \quad x_r x_s^\top > d\cos(2\theta^i), \quad x_r, x_s \in C^i.$$

- A high threshold restricts each class to a narrow cone.
- Between-class separation additionally needs separated classifier directions.
- Proposition 2 studies mixing under orthogonal W_v, nonnegative weights, and separated classes.

Appendix A, Propositions 1–2, PDF pp. 21–23. These assumptions restrict the theoretical claim.

Backup: two theoretical claims need their conditions

Output distinguishability: Proposition 3

The stated row-inner-product bound omits normalization present in the proof.
If every state row is $(2, 0)$, output cosine is 1 but row inner products are 4.
The proof also assumes nonnegative queries; SSE uses a linear query.

Unbounded retention: Proposition 4

With all-ones query/key, a channel whose decay stays exactly 1 can retain its contribution across an arbitrarily long idle interval. The threshold must not exceed that contribution; later writes can still interfere.

Appendix A, PDF pp. 23–25. Counterexample is reader analysis of the v2 statement.

Backup: selected gate mass is usually below one

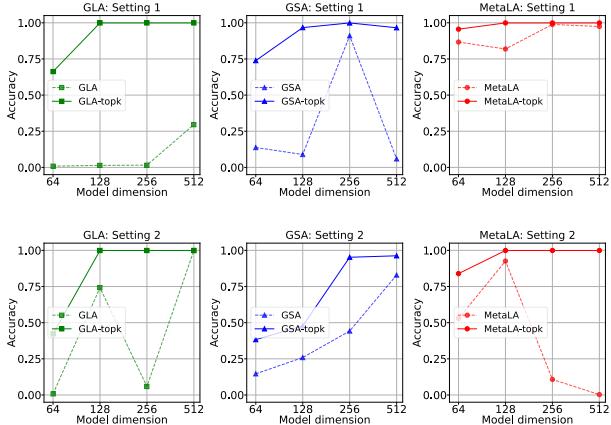
$$e_t = \mathrm{softmax}(x_t W_e), \qquad \sum_j k_t^j = 1$$

$$\boxed{\sum_{i \in \mathcal{T}_t} \sum_j e_t^i k_t^j = \sum_{i \in \mathcal{T}_t} e_t^i \leq 1}$$

- Full-partition softmax followed by top-k does not renormalize selected weights.
- v2 p. 8 describes total classification intensity as 1; Eqs. (9)–(14) generally give less than 1.
- The conference supplement retains original softmax gate values.
- Restoring a sum of 1 would change both scaling and gate gradients.

Sec. 4.1, PDF pp. 7–8; official supplementary model code. Equation-based clarification.

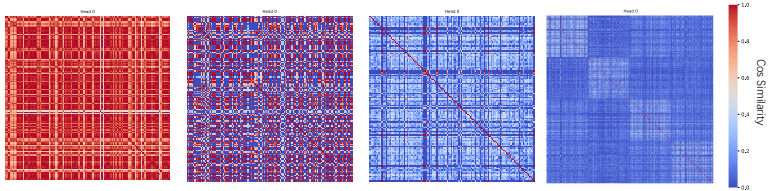
Backup: row-sparse variants improve synthetic recall



GLA, GSA, MetaLA; solid lines: top-k variants.
Setting 1 inserts zeros; setting 2 inserts random noise. Two layers, two heads, $k = d/4$.

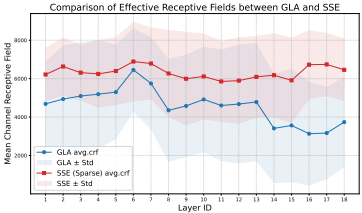
Appendix A, Fig. 9, PDF p. 25. Original figure.

Backup: row similarity and effective receptive fields



(a) Linear Classifier (b) SiLU Classifier (c) Softmax Classifier (d) SSE Classifier

Lower off-diagonal similarity supports more distinct state rows.



Receptive-field probe: pretrained 2B models, 8K training, last 128 tokens, gating threshold 0.001.
SSE has broader mean fields across layers; shaded bands are standard deviations.

Fig. 3, PDF p. 5; Appendix C, Fig. 10, PDF p. 28. Original figures.

Backup: all six RULER tasks at 32K

Task	Transformer-2B	SSE-H-n4k1-2B
S-NIAH-1	99.4	93.6
S-NIAH-2	70.8	84.4
S-NIAH-3	67.4	90.2
MK-NIAH-1	55.4	64.6
MQ-NIAH	28.6	44.6
MV-NIAH	18.1	40.6

SSE-H wins five of six tasks in this comparison.

Both models remain far from perfect multi-query/value recall.

Appendix C, Table 11, PDF p. 28. Zero-shot; after long-context extension.

Backup: 12B conversion uses a different hybrid pattern

Layer pattern

SSE $\rightarrow$ SWA $\rightarrow$ SSE $\rightarrow$ SWA $\rightarrow$ Softmax
Convert during 128K training, then supervised distillation.

12B model	AIME24	AIME25	MATH500	Olympiad	AMC23
Transformer	75.7	58.7	95.8	82.7	97.2
SSE-H-n4k1	74.3	57.3	96.0	85.3	96.3

Similar overall mathematical performance, with mixed task differences.
SWA denotes sliding-window attention; this is not the 2B five-to-one pattern.

Appendix C, Table 10, PDF p. 27.