Linear Attention via Orthogonal Memory

Fixed-size global statistics with local detail

Paper authors: Jun Zhang, Shuyang Jiang, Jiangtao Feng,
Lin Zheng, and Lingpeng Kong

arXiv:2312.11135v1 (18 December 2023)
Research reading presentation | 13 October 2026

arxiv.org/abs/2312.11135v1

Can each new token read a fixed-size history?

The design goal

Keep the cost per new token bounded as the context grows, while retaining useful global information and local detail.

- Compress history into r memory vectors.
- Preserve a local window of width w.
- Use local relative positions to support length extrapolation.

The tradeoff: stable long-context behavior with a restricted global memory.

Paper, Secs. 1–3, PDF pp. 1–5.

Causality exposes two different efficiency problems

Growing collections of summaries

Fixed-size chunks of length l leave more summaries to read:

$$\sum_{t=1}^{n} O(t/l) = O(n^2/l).$$

The paper uses causal LongShort as an example.

Large recurrent states

Kernel attention can retain linear recurrence:

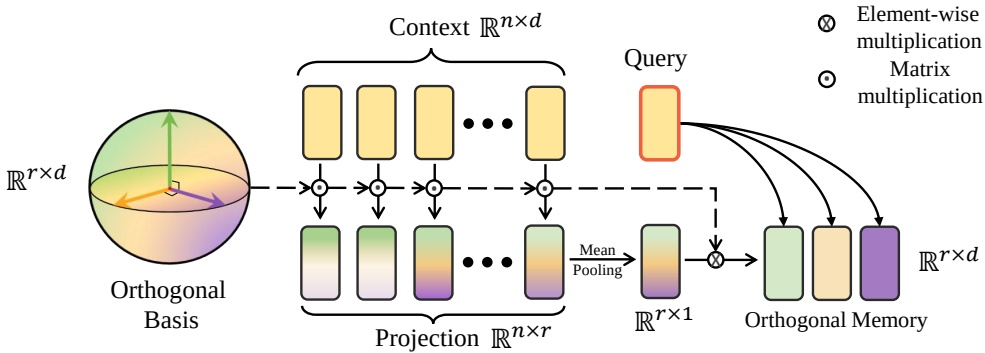
$$S_t = S_{t-1} + \phi(k_t) \otimes v_t.$$

State dimensions and saved training states increase practical cost.

LAVO chooses a small, incrementally updated statistic for the entire history.

Paper, Sec. 2, PDF pp. 2–3; Appendix A, Eqs. (7)–(11), PDF p. 15.

CODE compresses tokens into orthogonal directions



- Project token features onto r basis directions.
- Average each coordinate, then scale its basis vector.
- A query attends to these r memory slots.

Original Fig. 1, PDF p. 2; CODE definition in Sec. 3.1, PDF pp. 3–4.

Each memory slot is a scaled basis vector

Context $X \in \mathbb{R}^{n \times d}$; basis $B \in \mathbb{R}^{r \times d}$; coefficients $H \in \mathbb{R}^{r}$.

$$H = \frac{1}{n} \sum_{t=1}^{n} B x_t, \qquad \widetilde{X} = B \odot H = \mathrm{diag}(H) B.$$

$$\mathrm{LAVO}(q_t, X) = \mathrm{softmax}(q_t^\top \widetilde{X}^\top) \widetilde{X}.$$

- Row-wise broadcasting gives memory slot i: $H_i b_i$.
- Softmax is retained; it operates over r slots.
- The compressed memory supplies both keys and values.

Paper, Eq. (3) and Sec. 3.1, PDF pp. 3–4. The diagonal form is an equivalent notation.

The global state stores projected means

Direct consequence of the CODE equation

$$H = B\bar{x}, \qquad \bar{x} = \frac{1}{n}\sum_{t} x_t.$$

For a given basis, all context dependence enters through r scalars.

- Different histories with the same feature mean share this memory.
- Orthogonality separates directions; it does not prevent information loss.
- Local preprocessing and earlier layers can encode structure in x_t.

Reading analysis derived from Eq. (3), PDF p. 3. This statement concerns CODE on fixed input features.

An online mean makes the causal state easy to update

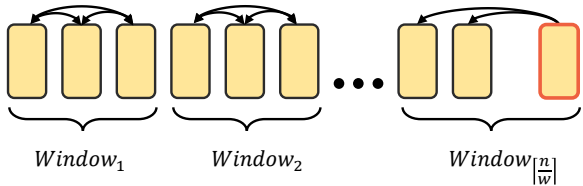
$$H_t = \frac{(t-1)H_{t-1} + Bx_t}{t} = H_{t-1} + \frac{Bx_t - H_{t-1}}{t}.$$

$$\widetilde{X}_{\le t} = \mathrm{diag}(H_t)B, \qquad F_{\mathrm{global},t} = \mathrm{softmax}(q_t^\top \widetilde{X}_{\le t}^\top)\widetilde{X}_{\le t}.$$

- Each step updates r coefficients and reads r slots.
- New information has weight $1/t$ in the mean.
- Basic CODE has no content-dependent write or forgetting gate.

Paper, Eq. (4), PDF p. 4. The update interpretation is a direct algebraic consequence.

Local attention supports both detail and compression

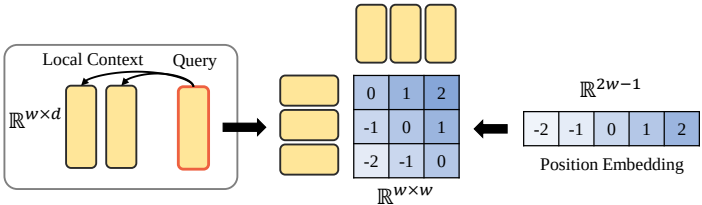


- **Local branch:** extend windows into previous context and mask to a bounded, causal neighborhood.
- **Global branch:** aggregate within historical windows using local attention before applying CODE.

$$F_t = \frac{1}{2}\left(F_{\text{local},t} + F_{\text{global},t}\right).$$

Paper, Sec. 3.2 and original Fig. 2, PDF p. 4. The authors motivate preprocessing by reduced feature variance.

Relative positions are embedded inside local attention



$$A_{ij} \leftarrow A_{ij} + P_{j-i+w-1}, \qquad P \in \mathbb{R}^{2w-1}.$$

- Local offsets remain bounded when total context length grows.
- Position parameters use $O(w)$ space; applying biases costs $O(nw)$.
- $O(w)$ describes the bias table, not all training activations.

Paper, Sec. 3.3 and original Fig. 3, PDF pp. 4–5.

Cross attention reads a compressed source sequence

Source $X \in \mathbb{R}^{n \times d}$ and target $Y \in \mathbb{R}^{m \times d}$:

$$Q = YW_q, \qquad \widetilde{X} = \mathrm{CODE}(X),$$

$$\mathrm{LAVO}(Q, X) = \mathrm{softmax}(Q\widetilde{X}^\top)\widetilde{X}.$$

- Compress the source once; each target query reads r slots.
- Remove local features: source and target lack a natural local alignment.
- Total compression and retrieval cost: $O(rd(n+m))$.

Paper, Sec. 3.4, Eqs. (5)–(6), PDF p. 5. Query multiplication is written consistently with row-vector shapes.

Linear cost requires fixed r and w

Computation	Cost including feature dimension d
Basis projection and pooling	$O(nrd)$
Reading global memory	$O(nrd)$
Local attention	$O(nwd)$
Total attention computation	$O(nd(r+w))$

- Streaming state and local cache stay bounded with history length.
- Training activations still grow with sequence length.
- The paper suppresses d; projection layers have additional cost.

Paper, Secs. 3.1–3.3, PDF pp. 3–5; explicit dimension and state accounting is reading analysis.

The central test is PG-19 length extrapolation

Language-model setup

- 12-layer GPT-2 backbone
- Hidden size 768; 12 heads
- GPT-2 BPE; 50,257 tokens
- LAVO: 122.91M parameters
- $r = 64$ memory directions

Protocol and comparisons

- Train: 8,192 context length
- Test: 8,192 / 12,288 / 16,384
- PPL: lower is better
- 125K steps; batch size 16
- Speed/memory: 1×80GB A100

Paper, Tables 1–2 and Sec. 4.1, PDF pp. 5–6. A separate experiment trains at 2,048 tokens.

LAVO trades in-range quality for stable extrapolation

Model	PPL at 8,192 ↓	12,288 ↓	16,384 ↓
FlashAttention	16.02	40.33	94.34
FA + RoPE	**15.09**	20.11	34.84
LongShort	15.52	**19.17**	26.06
ABC	31.53	86.78	147.43
Local	19.73	21.24	22.16
LAVO	19.43	19.41	**19.40**

- At training length, full attention has substantially better PPL.
- LAVO changes by only -0.03 from 8K to 16K: essentially stable.

Paper, Table 2, PDF p. 6. Selected rows; all models here train at length 8,192.

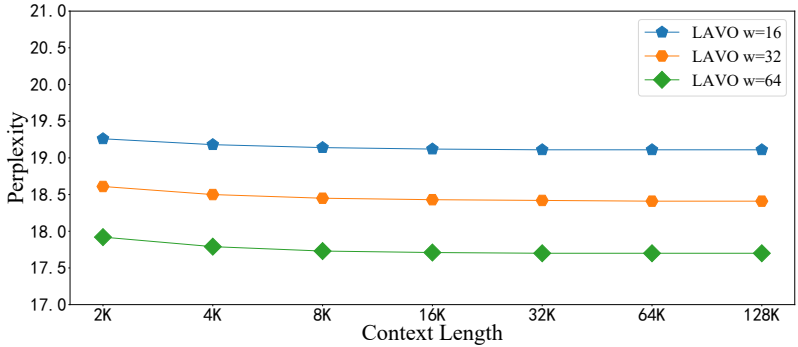
The speed advantage emerges at longer sequences

Model	8K speedup ↑	16K speedup ↑	16K memory (MB) ↓
Vanilla	1.00×	1.00×	25,747
FlashAttention	**14.00**×	18.70×	**9,133**
FA + RoPE	12.40×	16.96×	9,193
Local	5.00×	9.10×	11,909
LAVO	11.49×	**22.00**×	9,337

At 16K, LAVO is about 1.18× faster than FlashAttention in this table, with slightly higher memory use. FlashAttention is faster at 8K.

Paper, Table 2, PDF p. 6; 1×80GB A100, relative to vanilla at each length. The prose says 20.3×; this slide follows the table's 22.0×. Timing scope is not fully specified.

Training at 2K supports stable evaluation up to 128K



- Separate run: train at 2,048; batch size 64; test on one 80GB A100.
- Larger local windows help; gains from length largely plateau after 16K.
- This measures language-model PPL, not distant fact retrieval.

Paper, Sec. 4.3 and original Fig. 4, PDF p. 8. Authors report reaching 128K without IO optimization.

Speech results depend on the attention pattern

LJSpeech with Transformer-TTS; MCD and MSD measure acoustic distortion.

Pattern	Model	MCD ↓	MSD ↓
Causal	Vanilla	4.095	2.199
	S4D	4.030	2.189
	LAVO	**4.010**	**2.179**
Noncausal	LongShort	**3.913**	**2.136**
	LAVO	3.964	2.155

LAVO leads the causal comparison and ranks second in the noncausal table.

Paper, Tables 3 and 5, PDF p. 7; task description in Sec. 4.1, PDF p. 6. Selected rows.

Local structure explains much of the summary gain

Multi-News; maximum input / summary lengths: 4,096 / 400.

Pattern	Model	R-1 ↑	R-2 ↑	R-L ↑
Causal	S4D	**34.90**	**6.65**	**31.98**
	Local	33.50	6.27	30.74
	LAVO	34.16	6.16	31.16
Noncausal	Vanilla	34.61	6.35	31.66
	Local	38.50	10.54	35.39
	LAVO	**39.01**	**10.77**	**35.70**

Noncausal R-1: +4.40 over vanilla, but +0.51 over local. Causal R-2: local exceeds LAVO.

Paper, Tables 4 and 6, PDF p. 7. Selected rows; numeric values take precedence over the table's R-2 underlining.

Compressed source memory also helps cross attention

Point cloud completion

PCN dataset; PoinTr backbone.

Metric	Vanilla	LAVO
CDL1 ↓	7.425	**7.200**
CDL2 ↓	0.231	**0.216**
F-score ↑	0.779	**0.794**

Time-series forecasting

Informer; averaged evaluation settings.

Metric	Vanilla	LAVO
ETT MSE ↓	1.138	**1.123**
ETT MAE ↓	0.775	**0.765**
Weather MSE ↓	0.478	**0.444**
Weather MAE ↓	0.508	**0.492**

LAVO leads all reported metrics against vanilla, ABC, Performer, and cosFormer.

Paper, Tables 7–8, PDF p. 7; Sec. 4.2, PDF p. 8. ETT averages three datasets; both datasets average univariate/multivariate settings.

Context dissection has the larger ablation effect

Configuration	R-1 ↑	R-2 ↑	R-L ↑
LAVO	**39.19**	**10.82**	**35.95**
Without EPE	39.01	10.77	35.70
Without EPE and CD	35.29	6.80	32.29

- EPE: embedded position encoding; CD: context dissection.
- The final row **retains local features**; it removes dissection.
- R-1 losses: 0.18 for EPE, then 3.72 for CD conditional on no EPE.

Paper, Table 9, PDF p. 8; Sec. 4.3, PDF p. 9. Caption/metrics identify summarization, but the prose references the LM setup. The full protocol is inconsistent; these values are presented as the table reports them.

LRA gains are concentrated in three tasks

LAVO uses $r = 32$ and $w = 256$ for all five tasks.

Task / score ↑	LAVO	Best in table	Best model
Text	**65.28**	65.28	LAVO
ListOps	**40.57**	40.57	LAVO
Retrieval	80.17	82.06	Skyformer
Pathfinder	**72.14**	72.14	LAVO
Image	40.55	48.87	Reformer
Average	**59.74**	59.74	LAVO

Average margin: +2.37 over vanilla; +0.35 over Skyformer.

Paper, Appendix C, Table 11, PDF p. 16.

A useful tradeoff: bounded statistics plus local detail

1. **Mechanism.** Average projections provide a compact recurrent state; softmax reads a fixed number of memory slots.
2. **Evidence.** PG-19 PPL stays stable beyond training length; cross-attention and other sequence tasks also benefit.
3. **Tradeoff.** Mean pooling loses distinctions between histories; local structure is essential, and in-range LM quality is weaker.

Next question: can selective writes preserve rare distant information at the same state budget?

Synthesis of Secs. 3–4. Author limitation: short-sequence speed trails FlashAttention (Sec. 6, PDF p. 9). The proposed next question is reading analysis.

Paper reference and discussion guide

Jun Zhang, Shuyang Jiang, Jiangtao Feng, Lin Zheng, Lingpeng Kong.

Linear Attention via Orthogonal Memory. arXiv:2312.11135v1, 18 December 2023.

https://arxiv.org/abs/2312.11135v1

- Mechanism: Sec. 3, PDF pp. 3–5.
- Language modeling and 128K: Table 2 and Fig. 4, PDF pp. 6, 8.
- Additional evidence: Tables 3–11, PDF pp. 7–8, 15–16.

Figures reproduced from the author's v1 TeX source. Tables are selected and reformatted from the paper. Algebraic interpretations are labeled as reading analysis. Backup slides follow.

Backup: orthogonal slots can still hide a collision

Take $B = I_2$ and two contexts:

$$X^{(a)} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \qquad X^{(b)} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Both give $\bar{x} = (1,1)^\top$, $H = (1,1)^\top$, and $\widetilde{X} = I_2$.

$$\mathrm{LAVO}(q, X^{(a)}) = \mathrm{LAVO}(q, X^{(b)}) \quad \text{for every fixed } q.$$

- Memory rows are orthogonal, yet the contexts are indistinguishable.
- The full model may distinguish them through local/contextual features.

Teaching example derived from Eq. (3), PDF p. 3; not an experiment from the paper.

Backup: simulated generation has a separate speed table

Dummy inputs; recurrent evaluation from prefix length 1 to the target length.

Model	8,192 ↑	12,288 ↑	16,384 ↑
Vanilla	1.00	1.00	1.00
Local	3.35	6.86	10.34
LongShort	0.70	1.71	2.94
FlashAttention	**7.94**	**13.10**	15.01
FA + RoPE	5.73	10.28	12.66
LAVO	3.71	10.80	**20.81**

Implementation-sensitive speedups relative to vanilla; S4D is omitted. At 16K, LAVO / FlashAttention ≈ 1.39.

Paper, Appendix B, Table 10, PDF p. 15. These values are distinct from Table 2 and are not end-to-end serving throughput.

Backup: three details that affect interpretation

Maintaining the basis

Orthogonal initialization is described, but freezing B versus training it with an orthogonality constraint is not clearly specified.

Attributing the improvement

There is no controlled orthogonal-versus-nonorthogonal memory ablation. Table 9 removes CD only after EPE has already been removed.

Matching the reported configurations

Table 6's LAVO row equals Table 9's "without EPE" row. Table 9's caption and its surrounding protocol description disagree.

Paper, Sec. 3.1, PDF p. 3; Tables 6 and 9, PDF pp. 7–8; Sec. 4.3, PDF p. 9.