Hybrid Linear Attention Done Right

Efficient Distillation and Effective Architectures for Extremely Long Contexts

Paper authors: Yingfa Chen*, Zhen Leng Thai*, Zihan Zhou, Zhu Zhang, Xingyu Shen, Shuo Wang, Chaojun Xiao, Xu Han, Zhiyuan Liu

*Equal contribution | arXiv:2601.22156v1, 29 Jan 2026

The idea

Reuse a Transformer, retain selected global attention layers, and put rotational position information in recurrent layers.

Research discussion, 15 Oct 2026. Main talk: slides 1–22 (about 25 minutes); backup: 23–27.

Paper: arxiv.org/abs/2601.22156v1

Conversion must preserve long-distance retrieval

Why convert?

Reuse pretrained Q/K/V/O weights and knowledge instead of paying for full pretraining.

- Prior conversion budgets: about 20B–400B tokens for several major methods.
- HALO: about 2.3B additional training tokens.

What can go wrong?

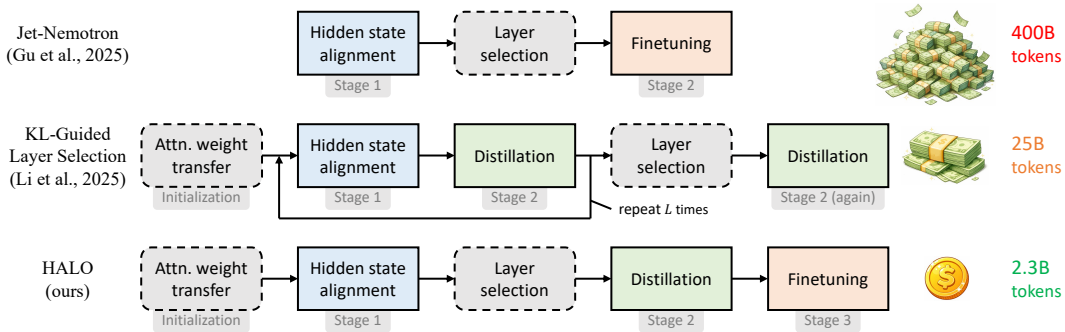
Short-task scores recover, while the model loses access to distant details.

- Fixed recurrent states compress history.
- A small number of full-attention layers preserve direct access.

Two design questions: which layers stay, and how should each type encode position?

Paper, Secs. 1–3, PDF pp. 1–4; Table 1, p. 2. SMART is a lower-budget exception, reported as >7B.

HALO separates local adaptation from global recovery



- First align candidate RNNs; then select attention layers using inference.
- Distill the assembled hybrid and finetune on longer sequences.

Paper, Fig. 2 and Sec. 4, PDF pp. 4–5. Dashed boxes: no training; solid boxes: training.

Three training stages use about 2.3B tokens

Stage	Tokens	Context	Learning rate
1. Layer alignment	320M	512	$10^{-3} \rightarrow 10^{-5}$
2. End-to-end KD	1B	512	$\eta_2 \rightarrow 10^{-5}$
3. Long-context FT	1B	16K	10^{-5}, constant

$$\mathcal{L}_1^{(l)} = \mathrm{MSE}\Big(\mathbf{Y}^{(l)}_{\mathrm{teacher}}, \mathrm{RNN}^{(l)}(\mathbf{X}^{(l-1)})\Big)$$

$$\mathcal{L}_2 = D_{\mathrm{KL}}(f_{\mathrm{orig}}(\mathbf{X}) \| f_{\mathrm{hybrid}}(\mathbf{X}))$$

- Stage 1 trains only candidate RNN layers; Stage 2 trains the hybrid.
- FineWeb-Edu, BF16, AdamW; $\eta_2 = 10^{-4}$ for the 2B student.

Paper, Eqs. (7), (10), PDF pp. 4–5; Tables 5–6, p. 14. Stage 3 loss is not explicitly specified.

Keep layers whose replacement especially harms recall

$$s_i = \frac{R_{\max} - \mathcal{R}(M^{(i)})}{C_{\max} - \mathcal{C}(M^{(i)}) + 10^{-6}}, \qquad \mathcal{I}_{\text{attn}} = \operatorname*{Top}_i\text{-}k(s_i)$$

- $M^{(i)}$: replace *only* teacher layer i with its aligned RNN.
- $R_{\max}, C_{\max}$: best recall / CSR across these replacements.
- High score: large recall drop, relatively small CSR drop.

For Qwen3-1.7B

Keep 7 of 28 full-attention layers: $\{3, 21, 2, 9, 25, 6, 8\}$ in importance order, using zero-based indices.

This is a single-layer heuristic; a near-zero denominator can amplify score noise.

Paper, Sec. 4.3, Eqs. (8)–(9), PDF p. 4; Table 7, p. 15. General rule stated as $k = \lfloor L/4 \rfloor$.

HyPE places RoPE in the recurrent layers

Recurrent layers	Full-attention layers
RoPE on Q and K	NoPE
Fixed-size state	Explicit KV history
Decaying local context	Position-dependent logit scaling

- Local order is represented through recurrent dynamics and RoPE.
- Global attention avoids RoPE extrapolation at long distances.
- Its inputs still contain order information from lower layers.

Paper, Sec. 5.1 and Fig. 3, PDF p. 5. The local/global division is the authors' design explanation.

The default RNN is fixed-decay Lightning Attention

$$\mathbf{q}_t = \mathrm{Norm}(\mathbf{x}_t \mathbf{W}_q), \qquad \mathbf{k}_t = \mathrm{Norm}(\mathbf{x}_t \mathbf{W}_k), \qquad \mathbf{v}_t = \mathbf{x}_t \mathbf{W}_v,$$
$$\widetilde{\mathbf{q}}_t = \mathrm{RoPE}_t(\mathbf{q}_t), \qquad \widetilde{\mathbf{k}}_t = \mathrm{RoPE}_t(\mathbf{k}_t)/\sqrt{d_h},$$
$$\mathbf{S}_t = \gamma_h \mathbf{S}_{t-1} + \widetilde{\mathbf{k}}_t^\top \mathbf{v}_t, \quad \mathbf{o}_t = \widetilde{\mathbf{q}}_t \mathbf{S}_t.$$
$$\gamma_h = \exp(-2^{-8h/H}), \qquad h = 1, \ldots, H.$$

- Row vectors $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^{1 \times d_h}$; state $\mathbf{S}_t \in \mathbb{R}^{d_h \times d_h}$.
- The forget rate depends on the head, not the input token.
- Output normalization and a sigmoid gate precede the output projection.

Paper, Appendix A, Eqs. (18)–(19), PDF pp. 13–14. Head index is omitted on vectors and states.

Fixed decay gives a concrete local memory scale

$$\mathbf{o}_t = \sum_{i=1}^{t} \gamma_h^{t-i} (\widetilde{\mathbf{q}}_t \widetilde{\mathbf{k}}_i^\top) \mathbf{v}_i \qquad (\mathbf{S}_0 = 0).$$

A useful derivation

The distance for a $1/e$ decay is

$$\tau_h = \frac{1}{-\ln \gamma_h} = 2^{8h/H} \quad \Longrightarrow \quad \tau_{\max} = 256 \text{ tokens}.$$

- At a 4,096-token lag, even the slowest head has factor e^{-16}.
- RoPE provides relative position within this decaying direct path.
- Global attention and depth can still carry distant information.

Teaching derivation from paper, Eqs. (18)–(19), PDF pp. 13–14. This is not a hard network-wide window.

NoPE attention needs sharper logits as context grows

$$\alpha_{t,i} = \frac{\exp(s_t \mathbf{q}_t \mathbf{k}_i^\top / \sqrt{d_h})}{\sum_{j=1}^{t} \exp(s_t \mathbf{q}_t \mathbf{k}_j^\top / \sqrt{d_h})}, \qquad s_t = \frac{\ln(t+a)}{\ln a}.$$

- Apply at inference; choose a after training using document loss.
- For HypeNet-2B, $a = 500$: $s_{512} \approx 1.11$, $s_{256K} \approx 2.01$.
- Scale Q before the attention operator; no new attention kernel is needed.

Why logarithmic growth is plausible

With one relevant logit gap Δ and n equal distractors: $p_{\text{relevant}} = 1/(1 + ne^{-s\Delta})$.
Keeping p stable requires $s\Delta$ to grow with $\ln n$.

Paper, Eq. (11), p. 5; Table 6, p. 14. Bottom block: teaching model. Current-query index corrects Eq. (17).

The architecture also changes normalization and heads

Modification	Purpose / implementation
RNN QK normalization	Control Q/K scale before rotation and state updates.
RNN GQA $\rightarrow$ MHA	Copy KV projections to independently trained heads.
Output gates	Normalize mixer output; multiply by sigmoid gate before O projection.
No short convolution	Added convolution failed for 1.7B conversion and did not help 8B.

More parameters, less KV traffic: the authors report roughly 10% parameter growth.

Paper, Sec. 5.2, PDF p. 6; Appendix A, p. 13; Appendix D, p. 14. Attention output gates are added at Stage 2.

Separate conversion quality from architecture potential

	Qwen3 conversion	Training from scratch
Model scale	1.7/4/8B teachers; 2/5/9B students	About 500M
Training tokens	About 2.3B added	20B
Training context	512, then 16K	4K
Attention placement	HALO selection	1 attention + 3 RNNs
Main question	Affordable transfer	Length generalization

- CSR: seven zero-shot downstream tasks; higher is better.
- NIAH: RULER Single-1/2/3; unspecified NIAH denotes their mean.
- Long-context Qwen3 evaluation uses YaRN beyond 40,960 tokens.

Paper, Secs. 6.1–6.3, PDF pp. 6–7; Appendix E.2, p. 16; Appendix H, pp. 18–20.

At 256K, retrieval gains vary across the three tasks

Model	Train tokens	Single-1	Single-2	Single-3
Qwen3-1.7B + YaRN	—	17.0	19.2	19.0
Jet-Nemotron-2B	400B	0.0	0.0	0.0
KL-LS-GDN-3B	25B	14.8	10.4	11.0
HypeNet-2B	**2.3B**	**99.8**	**86.2**	**48.8**

- Very large gains at extreme lengths with a smaller conversion budget.
- Single-3 remains difficult: 48.8% is far from perfect recall.
- At 64K Single-3, the teacher still wins: 98.4 vs. 72.6.

Paper, Table 2, PDF p. 7. Accuracy (%), higher is better. These are published-model comparisons.

Table 2 single-task means do not exactly match Table 4 aggregates; keep the two tables separate (backup slide 27).

Layer selection changes recall much more than CSR

Selection inside HALO	CSR	NIAH 128K	NIAH 256K
HALO score	**55.9**	**79.9**	**74.3**
Jet-Nemotron rule	55.0	63.7	56.2
KL-LS rule	55.3	58.3	44.3
Uniform across depth	54.0	61.9	50.9
Uniform in latter half	55.8	39.2	40.4

- Same conversion pipeline; swap the selection rule.
- HALO vs. latter-half uniform: only 0.1 CSR points apart, but 33.9 points apart on 256K retrieval.
- Teacher CSR is 58.5: the student retains a 2.6-point gap.

Paper, Table 4 and Sec. 6.5, PDF p. 8. Full Jet-Nemotron / KL-LS training procedures are not rerun here.

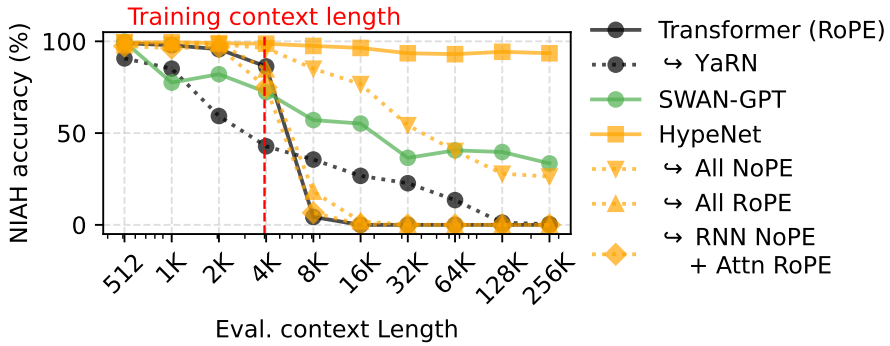
QK normalization and global NoPE matter most at 128K

Converted 2B configuration	CSR	NIAH 128K
Full HypeNet	**55.9**	**79.9**
Remove RNN RoPE	53.8	47.9
Add attention RoPE	55.8	19.7
Remove RNN QK norm	55.3	17.3
Keep RNN shared KV	55.8	83.5
Remove RNN output gate	55.6	74.5
Remove attention output gate	55.4	80.9

- Tiny CSR changes can hide 60+ point retrieval losses.
- Independent KV heads and attention gating do not win at every length.

Paper, Table 3, PDF p. 8. Same Qwen3-1.7B conversion setting; scores in percent.

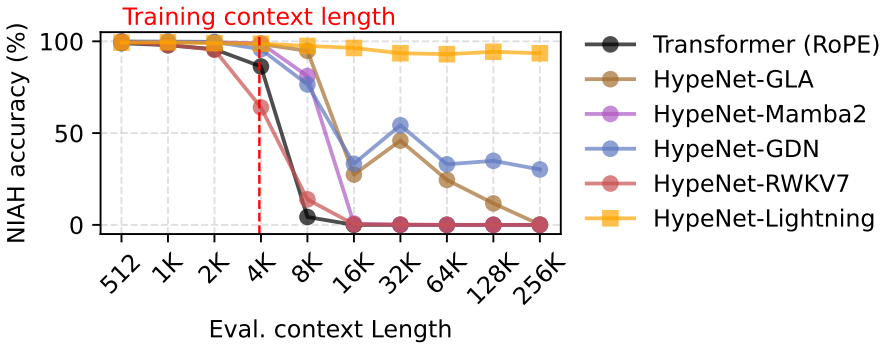
From scratch, HyPE reaches 64 times the training length



- About 500M parameters, 20B tokens, 4K training context.
- 93.5% NIAH at 256K; this is a separate from-scratch model.

Paper, Fig. 4 and Sec. 6.3, PDF p. 7. Red line: training length; orange solid line: complete HypeNet.

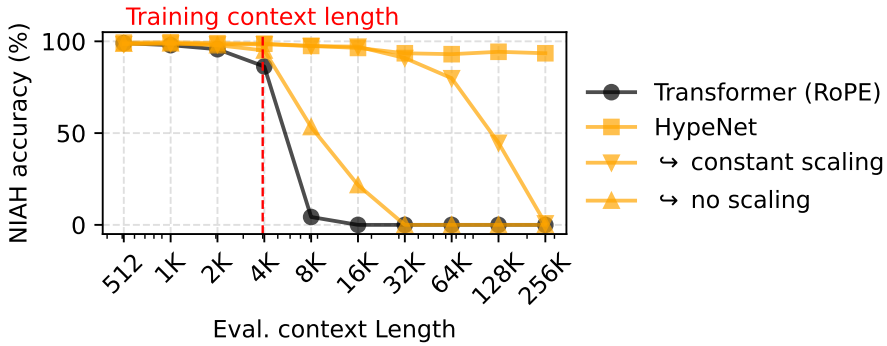
Lightning has the strongest extrapolation in this setup



- Authors' hypothesis: fixed decay gives more stable length behavior.
- QK norms and other module settings also differ across variants.

Paper, Fig. 5, PDF p. 7; discussion, p. 8; Table 15, p. 20. All variants omit short convolutions.

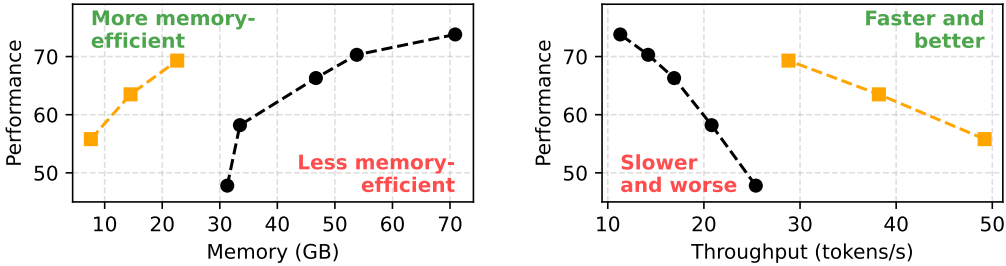
Dynamic scaling sustains the longest-range performance



- No scaling fails early; constant $s = 1.5$ delays the decline.
- Position-dependent $s_t = \log_a(t + a)$ maintains high accuracy.

Paper, Appendix F.1, PDF pp. 16–17; Fig. 8, p. 17. From-scratch, 4K-context setting.

Less KV memory offsets the extra parameters



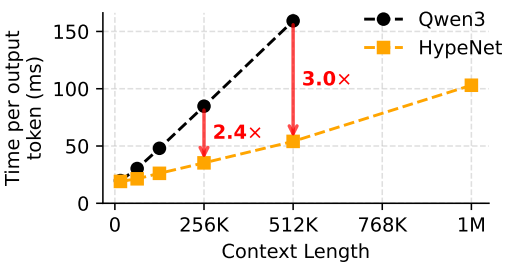
Black: Qwen3 Orange: HypeNet

- At 128K, HypeNet offers a better quality–memory–throughput tradeoff.
- For the 2B student, only 7 of 28 layers keep growing KV caches.
- Full attention remains: prefill still contains a quadratic term.

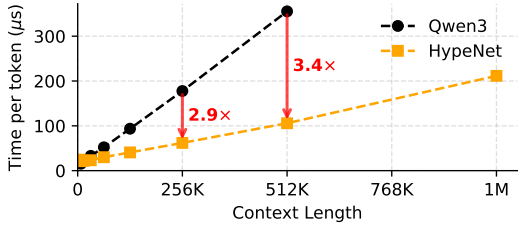
Paper, Fig. 1, left/center panels, PDF p. 2. Original vector panels cropped; legend restated above. BF16, batch 1.

At 512K: 3.0x faster decode, 3.4x faster prefill

Decode: milliseconds / output token



Prefill: microseconds / input token



- Single A800, batch 1, BF16; Qwen3-1.7B vs. HypeNet-2B.
- Qwen3 runs out of memory at 1M; HypeNet has runtime measurements.
- The reported retrieval evaluations extend to 256K.

Paper, Fig. 1 right, p. 2; Fig. 6 and Sec. 6.6, PDF pp. 8–9; efficiency setup, p. 7.

Training is affordable; selection still costs inference

Component	Token budget	A800 GPU-hours
Stage 1	320M training	10.0
Layer selection	234M inference	Not reported
Stage 2	1B training	43.4
Stage 3	1B training	37.7
Three training stages	About 2.3B	91.1

- Selection evaluates 28 variants on about 8.36M tokens each.
- The 91.1-hour sum excludes the unspecified selection runtime.
- These budgets reuse a pretrained teacher; searches add further cost.

Paper, Table 8 and Appendix E, PDF p. 15; Qwen3-1.7B conversion. Reported FLOPs have an unresolved factor mismatch.

The strongest evidence is for synthetic retrieval

- **General ability:** CSR drops from 58.5 to 55.9 for the 2B conversion.
- **Retrieval:** 256K Single-3 reaches 48.8%, with substantial room left.
- **Behavior:** FineWeb-Edu conversion may weaken instruction following and post-training alignment.
- **Efficiency:** speedups are measured at batch 1 on an A800; the hybrid still contains full attention.
- **Reproduction:** v1 has inconsistent layer counts and cross-table aggregates; avoid mixing those entries.

A useful next test

Evaluate real long-document tasks and recover instruction behavior; isolate fixed decay under matched module configurations.

Paper, Tables 2–4, pp. 7–8; Limitations, p. 9; Tables 6–7, pp. 14–15. Next test: reader's analysis.

Three lessons for designing converted hybrids

1. **Choose what to preserve.**
   Retrieval-sensitive layer selection protects useful global access.

2. **Assign position information by function.**
   Local recurrent RoPE and global NoPE work together with dynamic scaling.

3. **Evaluate the full tradeoff.**
   Low conversion cost and long-input speed come with measurable quality and behavior constraints.

Yingfa Chen et al., Hybrid Linear Attention Done Right: Efficient Distillation and Effective Architectures for Extremely Long Contexts.

arXiv:2601.22156v1 Author-linked code: THUNLP/hybrid-linear-attention.

Backup: longer layer alignment is not always better

Stage 1 tokens	CSR	NIAH 16K	NIAH 128K
100M	55.2	80.9	79.9
320M	55.4	88.2	80.9
625M	55.4	90.0	75.6
1.3B	55.1	81.0	61.2

- More local output matching does not ensure a better final hybrid.
- At 128K, 1.3B alignment tokens score 19.7 points below 320M.
- Compare within this ablation: its reference row differs from the main model.

Paper, Table 9, Stage 1 ablations, and Appendix F.1.1, PDF p. 17. Scores in percent.

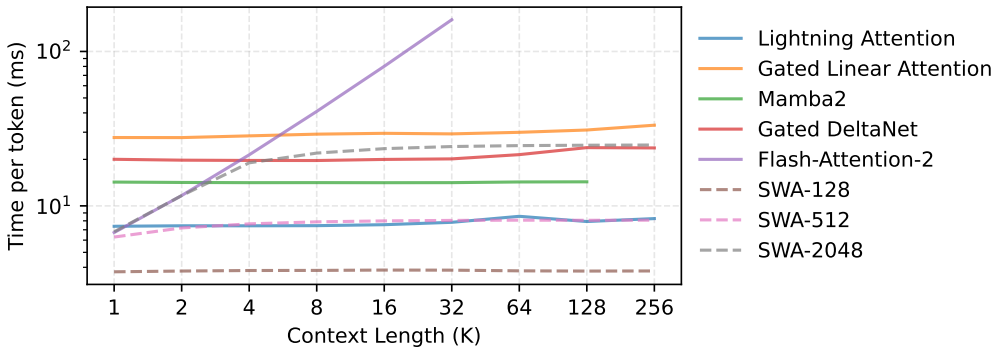
Backup: the learning rate changes the quality balance

Stage 2 maximum LR	CSR	NIAH 128K
10^{-5}	46.9	60.7
3×10^{-5}	55.5	54.9
10^{-4}	56.4	68.7
3×10^{-4}	46.0	36.1
10^{-3}	36.8	75.1

- 10^{-4} gives the best CSR in this group.
- Optimizing only long-context recall would favor a different setting.
- These are Table 9 ablations, not the final-model scores in Tables 3–4.

Paper, Table 9, Stage 2 ablations, and Appendix F.1.1, PDF p. 17.

Backup: Lightning also has an efficient kernel



- Lightning is close to sliding-window attention with window 512.
- This is a mixer benchmark, separate from whole-model speedups.

Paper, Fig. 7, PDF p. 16; Appendix E.1. Single A800-80GB, BF16; vertical axis is logarithmic (units as labeled).

Backup: interface compatibility does not ensure convergence

$$\mathbf{S}_t = \mathbf{F}_t \mathbf{S}_{t-1} + \mathbf{k}_t^\top \mathbf{v}_t, \qquad \mathbf{y}_t = \mathbf{q}_t \mathbf{S}_t \mathbf{W}_o^\top.$$

- Q/K/V/O or equivalent projections support weight transfer into several recurrent families.
- **Mamba2:** the paper uses a multi-head adaptation to inherit attention projections.
- **KDA:** Stage 2 gradients became infinite after a few steps; reducing the learning rate did not fix the reported run.
- **Module comparison:** QK normalization, activations and gate parameterizations differ; all tested from-scratch RNNs omit short convolution.

Paper, Appendix G, PDF pp. 17–18; Tables 10–11, p. 18; Table 15, p. 20.

Backup: use each v1 result under its own stated setting

- **Aggregates:** Table 2 HypeNet means are 80.8 at 128K and 78.3 at 256K; Table 4 reports 79.9 and 74.3.
- **Layer counts:** Table 6 reverses the 2B RNN/attention counts. Larger-model totals are inconsistent; the 4B-teacher row in Table 7 boxes 8 layers despite the stated $k = \lfloor 36/4 \rfloor = 9$.
- **Attention formula:** Eq. (17)'s denominator should use the current query $\widetilde{\mathbf{q}}_t$, not $\widetilde{\mathbf{q}}_j$.
- **Training details:** Stage 3's loss is unspecified; Table 8's total FLOPs do not equal tokens times its reported FLOPs/token.

The notes retain the original table values and label the softmax correction.

Paper, Tables 2 and 4, pp. 7–8; Eq. (17), p. 13; Tables 5–8, pp. 14–15. Arithmetic comparisons: reader's analysis.