RACE Attention

A Strictly Linear-Time Attention Layer for Training on Outrageously Large Contexts

Paper: Sahil Joshi, Agniva Chowdhury, Amar Kanakamedala,
Ekam Singh, Evan Tu, Anshumali Shrivastava

Rice University | Independent paper discussion

arXiv:2510.04008v5 (18 April 2026)
Reading: 15 October 2026

Replace token-pair interactions with differentiable bucket summaries.

The paper asks two distinct scaling questions

Can the model preserve task quality?

Train Transformers with RACE in place of softmax attention.

- Classification up to 64K tokens.
- Causal language modeling at 128 and 1024 tokens.
- Compare accuracy and perplexity.

Can the operator process huge inputs?

Time one attention layer, one forward–backward pass.

- Up to 12M tokens on GH200.
- Up to 75M tokens on CPU.
- Batch 1, 4 heads, reported $d = 128$.

Paper, Sec. 4, Tables 1–4 and Fig. 4, PDF pp. 8–11; Table 8, p. 16.

Choose a nonnegative kernel that can be sketched

$$O_i = \frac{\sum_j S_{ij} V_j}{\sum_j S_{ij}}, \qquad S_{ij} = \left(1 - \frac{1}{\pi} \arccos \frac{Q_i^\top K_j}{\|Q_i\| \|K_j\|}\right)^\gamma.$$

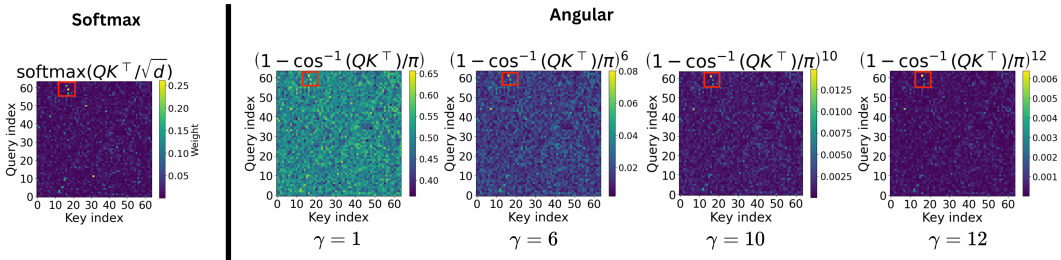
- The angular kernel is 1, $1/2$, and 0 at angles 0, $\pi/2$, and π.
- Raising it to γ increases contrast: orthogonal pairs have score $2^{-\gamma}$.
- Normalization makes each output a convex combination of values.

Three objects to keep separate

Softmax attention $\longrightarrow$ angular attention $\longrightarrow$ soft RACE approximation.

Paper, Sec. 3.1, Eqs. (3)–(4), PDF pp. 5–6.

Powers sharpen the angular similarity pattern



- Rows: queries. Columns: keys. Angular panels use $\gamma = 1, 6, 10, 12$.
- Larger powers concentrate the displayed high-similarity regions.
- Color scales differ; visual similarity does not establish kernel equivalence.

Original Fig. 3, PDF p. 6. Angular ignores vector norms; dot-product softmax does not.

A hash collision estimates a powered angular kernel

For a Gaussian hyperplane $w \sim \mathcal{N}(0, I_d)$,

$$\Pr[\operatorname{sign}(w^\top q) = \operatorname{sign}(w^\top k)] = \kappa_{\mathrm{ang}}(q, k).$$

Concatenate P independent signs:

$$h(x) = \operatorname{sign}(Wx) \in \{\pm 1\}^P, \qquad \Pr[h(q) = h(k)] = \kappa_{\mathrm{ang}}(q, k)^P.$$

- $R = 2^P$ possible buckets; the target exponent is $\gamma = P$.
- L independent tables average repeated collision estimates.
- Hard assignments obstruct ordinary backpropagation.

Paper, Sec. 2.2, PDF p. 4; Sec. 3.2, p. 7.

Soft corner assignments make hashing differentiable

$$u^{(\ell)}(x) = \tanh(W^{(\ell)}x), \qquad \phi_r^{(\ell)}(x) = \frac{e^{\beta u^{(\ell)}(x)^\top v_r}}{\sum_{r'} e^{\beta u^{(\ell)}(x)^\top v_{r'}}}.$$

- $W^{(\ell)} \in \mathbb{R}^{P \times d}$ has independent Gaussian rows.
- $v_r \in \{\pm 1\}^P$ is one of R hypercube corners.
- Each token becomes a probability vector over a small bucket bank.
- $\beta \to \infty$: hard hashing. $\beta \to 0$: uniform buckets.

Training choice

The experiments learn β; it multiplies logits and acts as an inverse temperature.

Paper, Algorithm 1, PDF p. 5; trainable β, Sec. 4, p. 8.

Each bucket stores mass and a weighted value sum

For table ℓ, build $\Phi_K^{(\ell)}, \Phi_Q^{(\ell)} \in \mathbb{R}^{N \times R}$.

Mass

$$A^{(\ell)} = (\Phi_K^{(\ell)})^\top \mathbf{1}_N \in \mathbb{R}^R$$

$$A_r^{(\ell)} = \sum_j \phi_r^{(\ell)}(K_j)$$

Needed for the final denominator.

Content

$$B^{(\ell)} = (\Phi_K^{(\ell)})^\top V \in \mathbb{R}^{R \times d}$$

$$B_r^{(\ell)} = \sum_j \phi_r^{(\ell)}(K_j) V_j$$

Stores the weighted values.

Every key writes to the same small bank; queries read this bank.

Paper, Algorithm 1, PDF p. 5; Sec. 3.2, p. 7.

Normalize after combining the tables

$$\mathrm{Num} = \frac{1}{L}\sum_{\ell} \Phi_Q^{(\ell)} B^{(\ell)}, \qquad \mathrm{Den} = \frac{1}{L}\sum_{\ell} \Phi_Q^{(\ell)} A^{(\ell)}$$

$$\widehat{O} = \mathsf{diag}(\mathrm{Den})^{-1}\mathrm{Num}.$$

The implicit kernel

$$\widehat{S}_{ij} = \frac{1}{L}\sum_{\ell} \phi^{(\ell)}(Q_i)^\top \phi^{(\ell)}(K_j).$$

Associate $\Phi_Q(\Phi_K^\top V)$ to avoid forming $\widehat{S} \in \mathbb{R}^{N\times N}$.

- Ratio of summed numerators and denominators, rather than an average of ratios.

Paper, Algorithm 1, PDF p. 5. See backup for a numerical example and Fig. 7 caveat.

Causality is a prefix update of the same statistics

Treat ϕ and V_t as column vectors:

$$A_t^{(\ell)} = A_{t-1}^{(\ell)} + \phi^{(\ell)}(K_t), \qquad B_t^{(\ell)} = B_{t-1}^{(\ell)} + \phi^{(\ell)}(K_t)V_t^\top.$$

$$\widehat{O}_t = \frac{\sum_\ell \phi^{(\ell)}(Q_t)^\top B_t^{(\ell)}}{\sum_\ell \phi^{(\ell)}(Q_t)^\top A_t^{(\ell)}}.$$

- Update before reading: the current key/value is included.
- Recurrent state: $LR(d+1)$ scalars for fixed feature maps.
- Training activations and backward memory are additional costs.
- The paper implements prefix scans using OpenMP/CUDA.

Paper, Algorithm 2, PDF p. 27; implementation discussion, p. 28.

Linear cost assumes a fixed bucket budget

$$T = O(LN(dP + PR + Rd)) \simeq O(LNRd), \qquad R = 2^P.$$

(P, L)	R	LR
$(2, 2)$	4	8
$(3, 3)$	8	24
$(4, 4)$	16	64

- Core sketch workspace: $O(L(NR + Rd))$.
- Token inputs/outputs still cost $O(Nd)$.
- QKV projections and FFNs retain their own costs.

Reader's deduction

The non-causal kernel has rank at most LR. Fixed state creates a capacity trade-off.

Paper, Sec. 3.2, PDF p. 7. Rank bound follows from the concatenated feature factorization.

The claimed bound separates bias and sampling error

$$\|\widehat{O} - O\|_{\mathrm{rms}} = O\left(\frac{P}{\beta} + \sqrt{\frac{\log(N/\delta)}{L}}\right) \|V\|_F.$$

- Target: P-powered angular attention in the non-causal setting.
- Appendix assumes unit Q/K and row mass $s_{\min} \geq C_1 N$.
- Bias falls with β; the simplified variance term uses $L \geq \log(2N/\delta)$.
- Taking $L \propto \log N$ changes fixed-budget linear time to $O(N \log N\, Rd)$.

Proof caveat from this reading

Lemma 10 centers a biased soft estimator at S. Normalization control needs a bias condition as well as a table-count condition.

Theorem 2, p. 8; Sec. 6.1, p. 17; Lemma 10, p. 24; simplification, p. 27. Backups give details.

The evaluation spans kernels, tasks, and hardware

Comparison	Role
FlashAttention2/3	Exact softmax; optimized GPU execution
Linear / Performer-256	Simple positive / random feature maps
Linformer-128	Sequence projection approximation
Angular ($\gamma = 8$)	Alternative kernel, still quadratic pairwise evaluation
YOSO-16	Related hard-hash estimator

- Accuracy experiments: single A100, best-validation checkpoint.
- Metrics: classification accuracy ↑; language-model perplexity ↓.
- Tiny Stories uses masked LM; WikiText-103 and PTB use causal LM.

Paper, Sec. 4, PDF pp. 8–10; Table 5, p. 15.

A moderate sketch can preserve task performance

Method	CIFAR-10 Acc.% ↑	QNLI Acc.% ↑	Tiny Stories PPL ↓
RACE (2,2)	63.7	60.7	4.2
RACE (3,3)	62.5	60.7	3.2
RACE (4,4)	**65.9**	**61.1**	**2.6**
FlashAttention2	61.44	**61.1**	2.7
Linear	60.0	60.7	7.0
Performer-256	64.9	61.0	10.0

- Contexts: CIFAR-10 1024, QNLI 2048, Tiny Stories 1024.
- The smallest RACE budget loses quality on Tiny Stories.

Selected rows from Table 2, PDF p. 9. All values are reported results.

Causal RACE matches the reported WikiText baseline

WikiText-103 @1024	PPL ↓
RACE (2,2)	23.9
RACE (2,3)	23.4
RACE (3,3)	21.9
RACE (3,4)	21.5
RACE (4,4)	**20.9**
FlashAttention2	**20.9**
Angular ($\gamma = 8$)	19.0

Paper, Table 4, PDF p. 9; Table 8, p. 16.

PTB @128

RACE (4,4): **53.4** PPL
FlashAttention2: 55.4 PPL

- More tables help in the shown fixed-P WikiText comparisons.
- Angular has a different exponent and quadratic computation.

At 64K, RACE trains in 584 s versus 1645 s

Arxiv	RACE Acc.%	FA2 Acc.%	RACE train	FA2 train
16K	71.3	69.8	82.4 s	95.7 s
32K	90.6	89.7	289 s	471 s
64K	97.92	97.00	584 s	1645 s

Fixed RACE $(P, L) = (3, 3)$; times are per epoch on A100 40 GB.

64K comparison

+0.92 percentage points; FA2/RACE training-time ratio **2.82**×.

- Same-class documents are concatenated to build each example.
- Dataset sizes differ by length; the rising accuracy is not a controlled context-length ablation.

Table 1, PDF p. 9; preprocessing, pp. 15–16. Ratio calculated from reported epoch times.

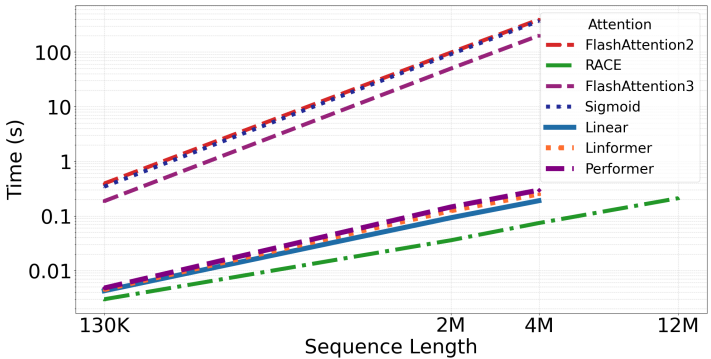
Food-101 supports a same-batch comparison with FA2

Method	Batch	Train/epoch	Test/epoch	Acc.%
RACE (2,2)	8	891 s	37 s	42.4
RACE (3,3)	8	950 s	40 s	**43.5**
RACE (4,4)	8	1042 s	42 s	40.3
FlashAttention2	8	2600 s	95 s	42.1
Performer-256	1	2546 s	105 s	42.4

- 51-class subset: 20,000 training / 2500 test images.
- 512×512 images, patch/stride 4: 16,384 tokens.
- RACE (3,3) vs. FA2: $+1.4$ points and $2.74\times$ training-time ratio.
- Linear, Linformer, and Performer used batch 1 after OOM at 8.

Table 3 and batch-size exception, PDF p. 9; dataset construction, p. 16.

The GPU operator reaches 12M tokens at a small budget



GH200, 96 GB

One layer; forward + backward.

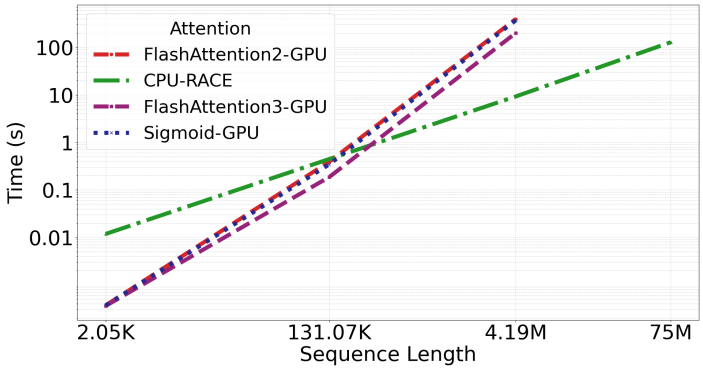
Maximum shown:

- (2,2): 12M
- (3,3): 10M
- (4,4): 7M

At ∼4M: authors report ∼0.1 s for RACE vs. ∼550 s for FA2.

Original Fig. 4d and endpoints in Fig. 4a–c, PDF p. 10; reported timing, p. 11. Operator timings.

Algorithmic savings can outweigh GPU acceleration



- RACE: Xeon Gold 5220R CPU.
- Baselines: GH200 GPU.
- Crossover near 131K.
- CPU RACE reaches 75M, roughly 100 s.

At ∼4M, reported CPU-RACE speedup: ∼40× vs. GPU-FA2, ∼20× vs. GPU-FA3.

Original Fig. 4e, PDF p. 10; Sec. 4, p. 11. The separate CPU 33M "under 10 s" claim conflicts with plotted CPU curves and is not used here.

More buckets do not guarantee better trained models

Task	Metric	Configuration 1	Configuration 2
Tiny Stories @512	PPL ↓	(5,4): **2.7**	(5,5): 5.1
FashionMNIST @784	Acc.% ↑	(2,5): **87.7**	(4,5): 85.7
Food-101 @16K	Acc.% ↑	(3,3): **43.5**	(4,4): 40.3

- P changes both target-kernel sharpness and the exponential bucket count.
- L changes sampling budget and the model's training computation.
- A fixed-input approximation bound does not imply monotonic trained-task quality.

Paper, Table 3, PDF p. 9; Tables 6–7, p. 16.

Additional tasks show mostly modest accuracy gains

Task	Best reported RACE	FA2	Difference (points)
ListOps @2000	41.9	41.4	+0.5
Text Retrieval @8000	80.9	80.5	+0.4
Yahoo @256	67.2	67.2	0.0
IMDB @512	81.3	80.0	+1.3
SST-2 @1024	79.4	78.5	+0.9

- Accuracy in percent; best RACE configuration selected separately per task.
- ListOps uses (2,2); Retrieval/Yahoo/SST-2 use (4,4).
- YOSO-16: 41.0 on ListOps and 80.2 on Retrieval.

Paper, Tables 9–10, PDF p. 16.

RACE offers a concrete design for compact attention state

1. **Mechanism:** differentiable angular hashing, bucket mass/value sums, and query-dependent normalization.
2. **Evidence:** competitive task scores in the tested settings, with strong single-layer scaling at long sequences.
3. **Trade-off:** finite bucket capacity, exponentially growing 2^P, and non-monotonic training outcomes.
4. **Theory:** useful bias–variance decomposition; normalization proof needs correction, and causal guarantees remain open.

Paper, Secs. 3–5 and Appendix 6. Proof caveat and rank interpretation are from this reading.

Discussion: connect the operator to a full long-context model

- Can the denominator analysis be repaired with explicit soft-bias conditions?
- What happens at matched state capacity and matched target exponent?
- How much operator speedup survives full causal training and decoding?

Primary source

S. Joshi, A. Chowdhury, A. Kanakamedala, E. Singh, E. Tu, A. Shrivastava. *RACE Attention: A Strictly Linear-Time Attention Layer for Training on Outrageously Large Contexts*. arXiv:2510.04008v5.

Versioned paper | Author repository

Main presentation ends here. The following slides provide derivations and protocol details.

Backup: the proof propagates kernel error through a ratio

$$\Delta = \widehat{S} - S = \frac{1}{L}\sum_{\ell}(\widehat{S}^{(\ell)} - \mathbb{E}\widehat{S}^{(\ell)}) + \widetilde{B}.$$

$$\|\Delta\|_2 \le \|\widetilde{B}\|_2 + \frac{4N}{\sqrt{L}}\sqrt{\log\frac{2N}{\delta}} + \frac{4N}{3L}\log\frac{2N}{\delta}.$$

Let $D = \mathrm{diag}(S\mathbf{1})$, $\widehat{D} = D + E$, $A = D^{-1}S$.

$$\widehat{A} - A = \widehat{D}^{-1}\Delta + (\widehat{D}^{-1} - D^{-1})S.$$

- Soft-hash bias: $\|\widetilde{B}\|_2 = O(NP/\beta)$, plus an exponential tail.
- Denominator stability requires $\|E\|_2 \le s_{\min}/2$.
- Then $\|\widehat{O} - O\|_F \le \|\widehat{A} - A\|_2 \|V\|_F$.

Theorem 6, pp. 19–20; Lemma 8, pp. 21–23; Lemma 11, p. 25; Theorem 13, p. 26.

Backup: Lemma 10 omits the soft-hash bias

Define $b = \max_{ij} |\mathbb{E}\widehat{S}_{ij} - S_{ij}|$. With probability at least $1 - \delta$,

$$\frac{\|E\|_2}{N} \le b + \sqrt{\frac{\log(2N^2/\delta)}{2L}}.$$

- Lemma 10 uses $b = 0$, contradicting the preceding soft-bias analysis.
- A sufficient repair is $b \le C_1/4$ and $L \ge (8/C_1^2)\log(2N^2/\delta)$.

Counterexample to the lemma as written

All Q/K equal one unit vector; $P = 2$, $\beta = 0.1$. Then $S_{ij} = 1$, but

$$\widehat{S}_{ij}^{(\ell)} = \|\phi^{(\ell)}(x)\|_2^2 \le \max_r \phi_r^{(\ell)}(x) \le e^{0.4}/4 < 0.373.$$

After averaging, $\|E\|_2 > 0.627N > N/2$ for every L.

Reader's analysis of Lemma 10, PDF p. 24. This refutes that lemma's condition, not every output bound.

Backup: bucket masses change the answer

Teaching example: one table, two buckets, two scalar values.

$$\phi(K_1) = (0.8, 0.2), \quad \phi(K_2) = (0.1, 0.9), \quad V = (2, 10).$$

$$A = (0.9, 1.1), \qquad B = (2.6, 9.4), \qquad \phi(Q) = (0.75, 0.25).$$

$$\widehat{O} = \frac{0.75(2.6) + 0.25(9.4)}{0.75(0.9) + 0.25(1.1)} = \frac{4.3}{0.95} \approx 4.5263.$$

- Equivalent pairwise weights: $(0.65, 0.30)$ before normalization.
- Mixing per-bucket means instead gives 4.3030, a different operator.

Teaching example derived from Algorithm 1, PDF p. 5. Fig. 7, p. 28 depicts within-bucket normalization; use Algorithm 1 for the precise global-ratio definition.

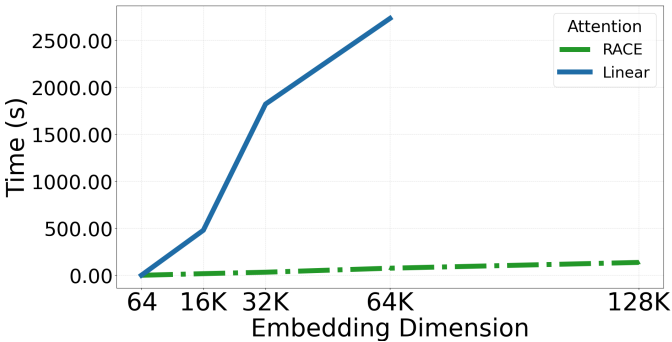
Backup: task models differ from the scaling operator

Dataset	Layers	Heads	Config. d	Batch	Epochs
WikiText-103	8	8	512	16	100
PTB	1	2	128	16	70
Tiny Stories	6	4	384	32	150
Arxiv	4	4	256	2	100
Food-101	8	8	512	8	100
Text Retrieval	4	2	384	1	20

- d here is the model configuration reported in Table 5.
- Arxiv/Food use gradient accumulation 16/4.
- AdamW; usually 1% warmup then linear decay. Constant LR below 2×10^{-4}.
- Food baseline batch exceptions are shown in the main results slide.

Paper, Table 5 and optimizer discussion, PDF p. 15.

Backup: fixed feature width explains dimension scaling



Fixed $N = 65536$.

Generic feature attention:

$$O(Nmd).$$

$m = d$ gives $O(Nd^2)$.

RACE uses $m = LR$, fixed by its sketch budget.

This comparison does not imply every random-feature method must be quadratic in d.

Original Fig. 6d, PDF p. 17. General feature-width interpretation is the reader's deduction.