WavePrune: One period is often enough for RoPE

Channel-wise truncation, length extrapolation, and GPU execution

Paper: Guancheng Du, Luotian Huang, Shaowen Wang,
Si Li, Kaifeng Lyu — Tsinghua University

Reading seminar 11 November 2026
arXiv:2610.06963v1 (3 October 2026)

The key claim: repeated RoPE rotations can distract attention

Idea

Keep each 2D RoPE channel's QK contribution only within its first rotation period.

- Training-free: HELMET improves on 4 of 5 models; Qwen3-8B: 35.7 $\rightarrow$ 40.0.
- Pretraining: better extrapolation from 2048 to 4096 tokens.
- A100 attention kernel, 32K, $d_h = 128$: 1.15$\times$ prefill, 1.24$\times$ decode.

Publication not found as of 2026-11-11. Preprint record.
Official code available; full README forthcoming.

Paper, Abstract; Table 1, PDF p. 8; Table 11, p. 32. This deck is a reading presentation.

Translation invariance does not require periodic repetition

$$f_c = \theta^{-2c/d_h}, \qquad w_c = \frac{2\pi}{f_c}$$

For fixed query/key content, channel c contributes

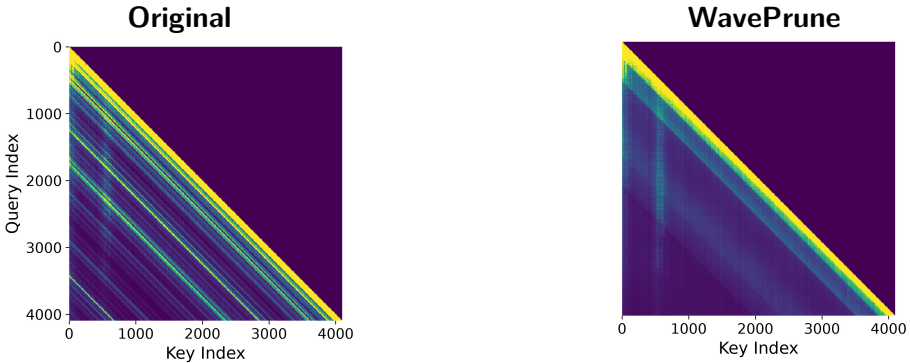
$$\ell_c(\delta) = m_c^Q m_c^K \cos(\delta f_c + \omega_c), \qquad \delta = i - j.$$

m_c^Q, m_c^K: channel magnitudes; ω_c: initial relative angle.

- A common position shift preserves δ: the useful invariance.
- Advancing δ by roughly w_c repeats the phase: aliasing.
- Many frequencies create approximate recurrences across channels.

Paper, Sec. 1–2, PDF pp. 1–3. f_c: radians/token; w_c: token distance.

Repeated phases create distant “ghost” sub-diagonals



Keys on the horizontal axis, queries on the vertical axis.

Parallel bright lines at distant lags lose strength after truncation.

Paper, Fig. 1, PDF p. 2; Qwen3-8B layer 0 head 20; mean over 100 PG19 sequences, length 4096.

WavePrune masks channel contributions before softmax

For causal positions $j \le i$:

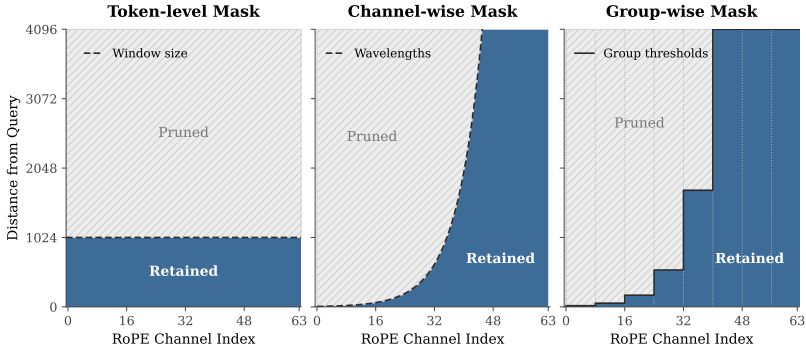
$$L^{\mathrm{WP}}_{i,j} = \frac{1}{\sqrt{d_h}} \sum_{c=0}^{d_h/2-1} \varphi_i^{(c)}(q_i)^\top \varphi_j^{(c)}(k_j)\, \mathbf{1}_{|i-j|\le w_c}.$$

$$A = \mathrm{softmax}(L^{\mathrm{WP}}), \qquad Y = AV.$$

- $\varphi_i^{(c)}$: 2D channel c rotated at position i.
- High frequencies stop contributing early; low frequencies reach farther.
- A pruned term becomes zero; causal future logits remain $-\infty$.
- Scaling stays $1/\sqrt{d_h}$; no learned pruning parameters.

Paper, Sec. 3.1, Eq. (6), PDF p. 4. Use the model's actual, possibly rescaled frequencies.

The mask varies across distance and channel index



Blue: retained contributions. Hatched: pruned contributions.

Long-distance token pairs can still interact through low-frequency channels.

Paper, Fig. 2, PDF p. 4. A zero logit still participates in softmax; this differs from a token window.

Theorem 1: even constant content has infinitely many near-maxima

Conditions and conclusion

If $q_i = q$ and $k_i = k$ at every position, define $L^\star = \sup_{\Delta \geq 0} L_\Delta$. For any $\tau > 0$, infinitely many lags satisfy

$$L_{i,i-\Delta} \geq L^\star - \tau, \qquad i > \Delta.$$

1. Logits are a finite weighted sum of cosines.
2. Wrapped phases live in a bounded multidimensional box.
3. Pigeonhole gives arbitrarily close phase returns and logit recurrences.

Paper, Theorem 1, Eq. (7), PDF p. 5; direct proof, Appendix A, p. 18.

The random extension relates recurrence to channel sensitivity

Under symmetric i.i.d. Q/K distributions (the proof uses independence):

$$\Pr(L_{i,i-\Delta} - L_{i,i-\Delta-P} \leq 2\tau) \geq \tfrac{1}{4}$$

for infinitely many P, all $\Delta \geq 0$ and $i > \Delta + P$.

- The bound is on a *one-sided* logit gap.
- Fewer sensitive mean-vector channels give a smaller recurrence bound.
- A competing logit enlarges the softmax denominator and dilutes the target.

Real hidden states depend on context; empirical attention maps establish that recurrences occur at practical lengths.

Paper, Appendix B.2–B.4, PDF pp. 19–24. Sensitivity definitions and proof details in backup.

FlashWavePrune groups 16 scalar dimensions per MMA unit

- Exact channel predicates disrupt a standard QK GEMM.
- Group 16 scalar dimensions, corresponding to 8 two-dimensional RoPE channels.
- Use the largest threshold in each group:

$$d'_g = \max_{k \in \text{group } g} d_k.$$

Why this works on Tensor Cores

Compute a standard partial GEMM per group, then apply one shared distance mask. Most groups are entirely skipped or retained.

The maximum threshold relaxes the exact mask. Most experiments use this approximation.

Paper, Sec. 3.3, Eq. (8), PDF pp. 6–7. d_k: wavelength threshold of dimension k's channel.

Two split points isolate the expensive boundary work

For a query/key block pair, distances span $[l_D, r_D]$.

Group	Condition	Execution
Fully masked	$d'_g < l_D$	Skip QK; skip K load in decode
Critical	$l_D \leq d'_g < r_D$	Group GEMM + mask
Fully retained	$d'_g \geq r_D$	Standard suffix GEMM

- Monotone thresholds give `splitL` and `splitR`.
- Specialize reachable split pairs to control branching and code size.
- Softmax and AV retain the standard FlashAttention path.

Paper, Sec. 3.3, Fig. 5, PDF p. 7; Appendix C.1, p. 25.

Evaluation separates quality, learning, and kernel efficiency

Question	Setting	Evidence
Drop-in quality	Five LLMs; 64K/128K HELMET	Task scores ↑
Learning benefit	GPT-2, 117M–774M; train 2K, test 4K	Validation loss ↓
Mechanism	Tetranacci and NIAH	Loss + attention
GPU efficiency	A100; bf16; GQA 32/8; batch 32	FA2/WP latency ↑

Baseline: the same model or architecture with original RoPE attention.

WavePrune uses group approximation unless otherwise specified.

Paper, Sec. 4, PDF pp. 7–10; Appendices D, F, G. 2K/4K denote 2048/4096 tokens here.

HELMET improves on four models, with a clear Llama exception

Model	Baseline	WavePrune	Change
Ministral3-3B	39.1	39.7	+0.6
Qwen3-8B	35.7	40.0	+4.3
Gemma3-12B	38.2	38.7	+0.5
Qwen3-32B	55.1	56.3	+1.2
Llama3.1-8B	38.2	35.2	−3.0

Six task groups, averaged over 64K and 128K; higher is better.

Qwen3-8B: Recall +6.0, ICL +10.0, Cite −0.5 points.

An average gain can include individual task regressions.

Paper, Table 1, PDF p. 8. Adapted HELMET protocol; Qwen thinking disabled, YaRN above 32K.

The evaluation protocol matters for interpreting the scores

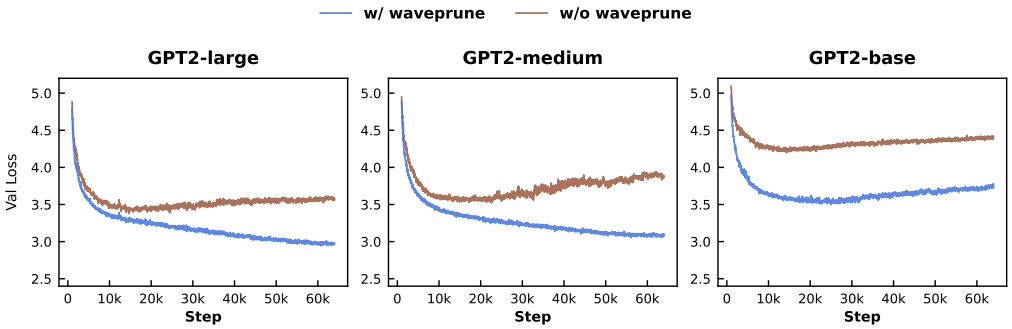
- Six groups: Recall, RAG, Re-rank, Cite, LongQA, ICL.
- NarrativeQA uses ROUGE-1 F1; summarization is omitted.
- Output caps increase to 200 for RULER MV and 300 for JSON KV.
- Table 1 reports cross-task standard deviations.

Complementary general-task results

Eight-benchmark mean scores rise modestly on the four reported models:
Ministral 47.47 $\rightarrow$ 48.10; Qwen3-8B 63.77 $\rightarrow$ 64.09;
Gemma 55.77 $\rightarrow$ 56.15; Qwen3-32B 52.05 $\rightarrow$ 52.91.

Paper, Appendix D, PDF pp. 27–28; Table 7, p. 31. Table 7 does not include Llama3.1-8B.

Pretraining with WavePrune improves 2K-to-4K extrapolation



OpenWebText; GPT-2 with RoPE base 10^4; train length 2048.

Horizontal: training step. Vertical: validation loss at length 4096.

Blue improves extrapolation at all three model sizes.

Paper, Fig. 6, PDF p. 9; Appendix G, p. 29. 64K training steps; about 15.7B training tokens.

The best result uses WavePrune in both training and evaluation

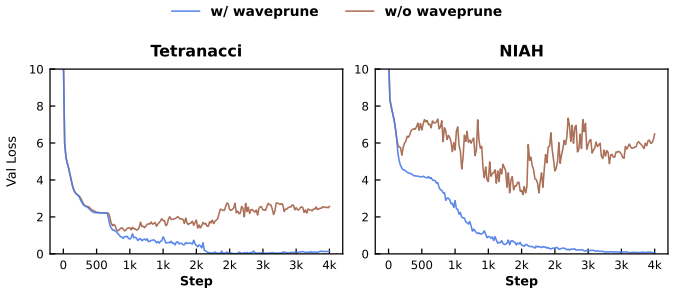
Pretraining	Evaluation	Medium loss	Large loss
RoPE	RoPE	3.8564	3.5618
RoPE	WavePrune	3.3314	3.1797
WavePrune	RoPE	4.3849	4.5162
WavePrune	WavePrune	**3.0953**	**2.9536**

Validation length 4096; lower is better.

- Inference-only integration already helps.
- Learning with the rule gives an additional benefit.
- Removing the rule after learning hurts: the model adapts to it.
- At length 2048, matched train/eval losses remain comparable.

Paper, Table 10, PDF p. 32. GPT2-medium 345M; GPT2-large 774M.

Controlled tasks test local dilution and distant interference



- Tetranacci: predict a sum of the previous four terms, modulo 1000.
- NIAH: recall a seven-digit number hidden in irrelevant text.
- GPT2-base; train 2048, validate 4096; both losses improve.

Paper, Sec. 4.3; Fig. 7, PDF p. 10; Appendix F, pp. 28–29. Each dataset: 500K samples, 99:1 split.

Attention probes support both proposed mechanisms

Local dependencies: Tetranacci

Attention to neighbors within 15 tokens, averaged across 12 layers:
Baseline: $0.161 \rightarrow 0.123$, retention 76.7%.
WavePrune: $0.179 \rightarrow 0.148$, retention 82.8%.

Long-range retrieval: NIAH

Across four illustrated samples, the first retrieval heads attend to the target more consistently with WavePrune.
Sample losses: WP $0.030, 0.005, 0.003, 0.001$;
baseline $6.328, 4.734, 10.840, 5.767$.

The NIAH table contains inactive WavePrune heads too; the improvement is in the overall retrieval pattern.

Paper, Tables 13–14, PDF p. 33. Retention percentages are reported by the paper from unrounded values.

Llama reveals when high-frequency truncation removes useful signals

Observed failure

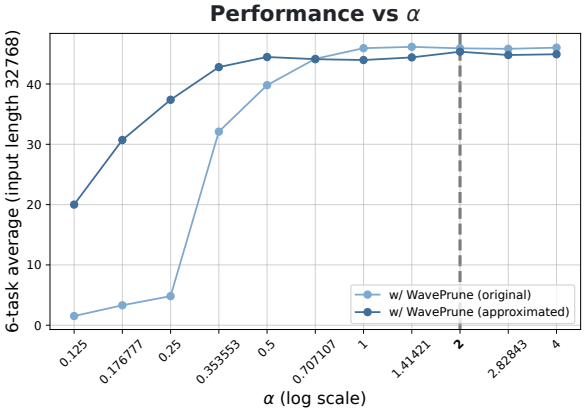
Llama3.1-8B, 32K NIAH: layer 16 head 21 retrieves a middle-context needle through a channel of wavelength about 571 tokens.

- The useful query–key distance greatly exceeds that wavelength.
- WavePrune removes exactly this long-range signal.
- Larger thresholds, fewer channels/layers, and ProLong continuation did not recover the baseline.

Frequency is a scale prior; learned computation can cross that scale.

Paper, Appendix E, PDF p. 28; Table 1, p. 8. HELMET mean: 38.2 $\rightarrow$ 35.2.

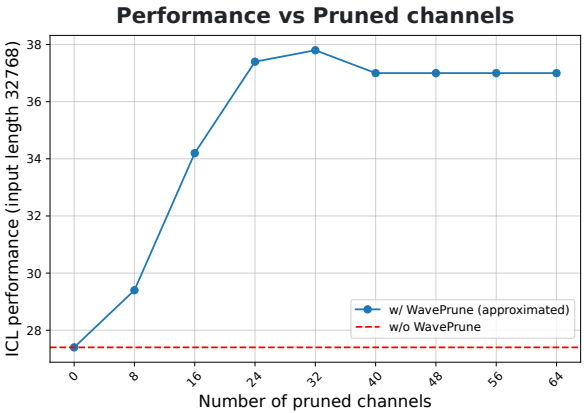
One period is a useful default; grouping has an accuracy regime



$w_c = \alpha\pi/f_c$; $\alpha = 2$ means one period. Near this default, the two implementations are close; aggressive pruning separates them.

Paper, Fig. 8a, PDF p. 11. Qwen3-8B, 32K, six-task HELMET average; exact peaks slightly below 2.

Benefits span several frequency groups, then saturate



Truncate the first n highest-frequency channels. Beyond 40, remaining group windows exceed 32K and truncation has no further effect.

Paper, Fig. 8b, PDF p. 11; Table 12, p. 32. Qwen3-8B, 32K, ICL only; grouped implementation.

Kernel speedups depend on length and head dimension

Phase	8K	16K	32K	64K
Prefill, $d_h = 128$	1.10×	1.13×	1.15×	1.19×
Decode, $d_h = 128$	1.13×	1.20×	1.24×	1.27×
Prefill, $d_h = 64$	1.02×	1.05×	1.09×	1.10×
Decode, $d_h = 64$	1.04×	1.05×	1.05×	1.05×

A100; bf16; batch 32; GQA 32 query heads / 8 KV heads.
Speedup = FA2 latency / WavePrune latency.

- QK FLOP savings favor prefill; fewer K reads also help decode.
- At 128K, H128 decode reaches 1.28×; both prefill kernels OOM.
- These are attention-kernel measurements; KV eviction is future work.

Paper, Sec. 4.4, PDF pp. 9–10; Table 11, p. 32.

Forward wins require careful specialization; backward is harder

- At 32K, removing split specialization makes H128 prefill 1.47× slower than the final kernel.
- Removing suffix-only QK computation makes it 1.46× slower.
- Removing suffix-only K loads makes H128 decode 1.14× slower; H64 has a 0.99× exception.

Backward	8K	16K	32K
Overhead vs. FA2	+8.58%	+3.81%	+0.18%

Backward needs masks before MMA and shares work across dQ, dK; full split-pair specialization causes code and local-memory costs.

Paper, Appendix C, Tables 3–4, PDF pp. 25–27.

The contribution connects a positional mechanism to executable sparsity

1. A per-frequency finite-support rule can preserve relative positions while reducing periodic distractions.
2. Quality improves for several LLMs and small-model extrapolation; Llama exposes a concrete failure mode.
3. A grouped approximation translates QK sparsity into practical GPU gains.

Next experiments suggested by these results

Head-specific thresholds that preserve retrieval signals; larger-scale pretraining; actual KV-capacity reclamation.

Paper, Sec. 6, PDF pp. 11–12; Appendix E. Future experiments are proposals, not reported results.

References and relation to concurrent work

Primary paper

Du, Huang, Wang, Li, and Lyu (2026).
WavePrune: One period is often enough for RoPE.
arXiv:2610.06963v1 Official implementation

- Complements frequency rescaling such as YaRN; uses updated frequencies.
- Concurrent ATFlash also uses wavelength windows and grouping.
- Authors distinguish WavePrune by a one-period default, ghost-diagonal analysis, and pretraining experiments.

Paper, Sec. 5, PDF pp. 10–11. Publication not found as of 2026-11-11; code files present, full README pending.

Backup: the recurrence proof uses phase returns

$$L_\Delta = \frac{1}{\sqrt{d_h}} \sum_c m_c^Q m_c^K \cos(\Delta f_c + \omega_c)$$

$$|L_{\Delta+P} - L_\Delta| \le \frac{1}{\sqrt{d_h}} \sum_c m_c^Q m_c^K |\operatorname{wrap}(P f_c)|.$$

- Partition the bounded phase box into finitely many small cells.
- Infinitely many positions share one cell; their differences give infinitely many almost-periods.
- Start within $\tau/2$ of the supremum and shift by a $\tau/2$ almost-period.

Paper, Appendix A, PDF p. 18. Constant Q/K; a raw full-dimensional bound can be extremely large.

Backup: sparsity refines the recurrence bound

$$t_c = \frac{2\pi m_c^Q m_c^K}{\sqrt{d_h}}, \qquad \sigma_\tau(t) = \min_{\|z-t\|_1 \le \tau} \|z\|_0$$

$$P_{\min} \le \prod_{i=1}^{\sigma_\tau(t)} \left\lceil \frac{\sigma_\tau(t) t_i^s}{\tau} \right\rceil .$$

- m_c^Q, m_c^K: channel magnitudes of mean vectors; t_i^s: sorted sensitivities.
- Absorb weak channels into the error budget; finely cover strong channels.
- Four reflected inner products average to the mean-vector inner product, giving a probability lower bound of $1/4$.

Paper, Appendix B.2–B.4, PDF pp. 19–24. Visual contribution sparsity need not equal sensitivity sparsity.

Backup: a competing logit limits softmax concentration

Let L_t be the target logit and L_g a ghost competitor. If $L_t - L_g \le 2\tau$, then

$$A_t \le \frac{e^{L_t}}{e^{L_t} + e^{L_g}} = \frac{1}{1 + e^{L_g - L_t}} \le \frac{1}{1 + e^{-2\tau}}.$$

Taking $\tau = \frac{1}{2}\log((1-\epsilon)/\epsilon)$ gives

$$\Pr(A_{i,i-\Delta} \le 1 - \epsilon) \ge \tfrac{1}{4}$$

once a suitable competing lag is available.

Paper, Corollary 1, Appendix B.2.3, PDF p. 20. $0 < \epsilon < 1/2$; same distributional conditions as Theorem 2.

Backup: native Qwen3 group thresholds explain the plateau

Group end channel	7	15	23	31
Wavelength (tokens)	28	160	900	5063
Group end channel	39	47	55	63
Wavelength (tokens)	28472	160114	900388	5063255

Values rounded down, native RoPE base 10^6.

At 32K, the groups ending at channels 47, 55, and 63 retain all causal distances.

Activating truncation there changes nothing.

Paper, Table 12, PDF p. 32. With frequency rescaling, thresholds must be recomputed.