Outlier-Aware Attention Linearization

Not All Dimensions Are Equal: Grouped Quadratic Kernels for Pretrained Attention

Paper authors: Weiqi Sun, Qianyue Wu,
Wenjia Wang, Tianyang Hu

arXiv:2610.13674v1 | Research discussion, 19 October 2026

Share coefficients according to query–key contribution scales.

A 25-minute discussion; technical details in backup slides.

Can a small kernel preserve pretrained attention?

$$y_i = \frac{\sum_{j \le i} e^{q_i^\top k_j / \sqrt{d_h}} v_j}{\sum_{j \le i} e^{q_i^\top k_j / \sqrt{d_h}}} \quad \longrightarrow \quad \widehat{y}_i = \frac{\sum_{j \le i} \kappa(q_i, k_j) v_j}{\sum_{j \le i} \kappa(q_i, k_j)}.$$

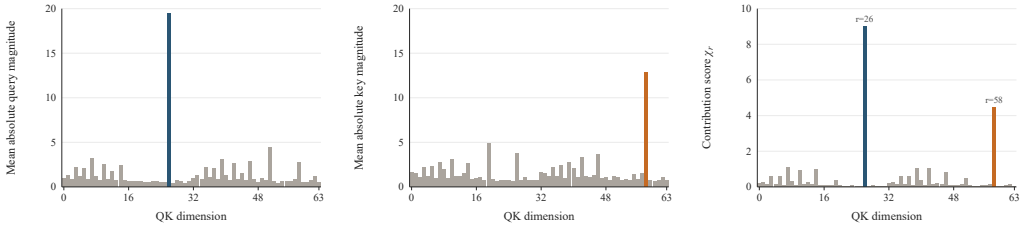
- Separable kernels permit prefix accumulation without an $N \times N$ matrix.
- Limited adaptation leaves little room to repair a poor approximation.
- A global quadratic $a + bs + cs^2$ shares coefficients across all coordinate contributions.

The paper's hypothesis

Large differences in coordinate contributions call for nonuniform coefficient sharing.

Paper, Secs. 1–3, PDF pp. 1–4. N: sequence length; d_h: Q/K head dimension.

A few coordinates dominate query–key contributions



(a) Query magnitude (b) Key magnitude (c) Dimension-wise contribution

$$\chi_{p,r} = \frac{1}{|\Omega_p|} \sum_{(i,j)\in\Omega_p} \left| \frac{q_{p,i,r} k_{p,j,r}}{\sqrt{d_h}} \right|, \qquad \chi_r = \frac{1}{P} \sum_p \chi_{p,r}.$$

Post-RoPE Q/K; causal pairs; equal weight per prompt. Coordinate products combine information from both Q and K.

Paper, Fig. 1 and Eq. (2), PDF p. 4. Highlighted coordinates: $r = 26, 58$.

Approximating only 5.5% of dimensions retains large error

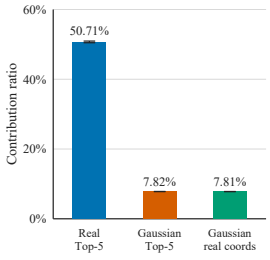
$$\widetilde{\kappa}_{ij} = \widehat{\kappa}_{ij,\mathcal{S}} \kappa_{ij,\overline{\mathcal{S}}}^{\text{exact}}, \qquad R_{\mathcal{S}} = \mathcal{E}_{\mathcal{S}} / \mathcal{E}_{\text{full}}.$$

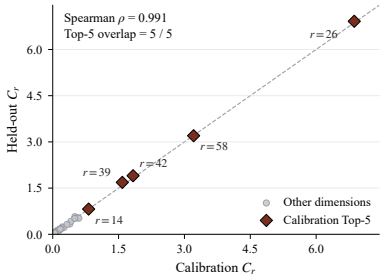
Method	High-contribution $R_{\mathcal{S}}$	Gaussian control
Performer	37.8%	4.7%
SpectraFormer	58.1%	18.1%
DiJiang	64.8%	19.4%
TaylorShift	63.9%	13.6%

- 147 selected dimensions across 42 heads, each of dimension 64.
- Separate matched-coordinate controls: $p = 1/1001$ for every method.

Paper, Table 1, PDF p. 5; App. A.2, pp. 15–16. Output MSE after normalization; not an additive error decomposition.

Representative-head diagnostics support fixed groups





Top-5 concentration is much stronger than in isotropic Gaussian controls.

Input stability: $\rho = 0.991$, Top-5 overlap 5/5. Dominant dimensions persist over 128–4096-token prefixes.

Paper, Figs. 5–6 and Table 4, PDF pp. 15, 17. Single-head diagnostics; GPT-2 evidence in Fig. 4, p. 14.

Relax sharing before deciding which parameters to tie

Define the coordinate contribution $x_r = q_r k_r / \sqrt{d_h}$, so $s = \sum_r x_r$.

Global quadratic

$$a + bs + cs^2 = a + \sum_r b x_r + \sum_{r,t} c x_r x_t.$$

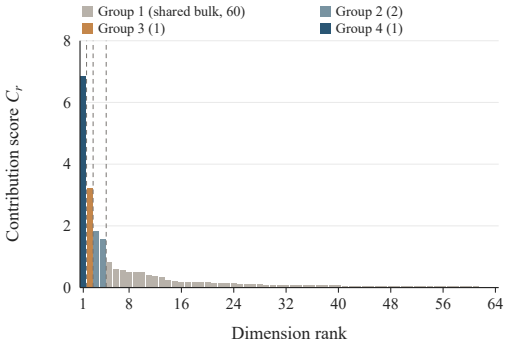
Dimension-aware quadratic

$$\kappa_{\mathrm{dim}}(q, k) = a + \sum_r b_r x_r + \sum_{r,t} C_{rt} x_r x_t, \qquad C = C^\top.$$

Linear weights b_r; squares C_{rr}; cross-coordinate interactions C_{rt}.

Paper, Sec. 4.1, Eq. (3), PDF p. 5. Setting every $b_r = b$ and $C_{rt} = c$ recovers the global kernel.

Contribution groups concentrate flexibility in the tail



$$s_g = \sum_{r \in \mathcal{G}_g} x_r$$

- Sort by contribution; inspect gaps in the high-contribution tail.
- Keep gaps stable under prompt-level bootstrap.
- Share linear coefficients within groups and quadratic coefficients across group pairs.

Qwen: 2–5 groups/head. Llama: 2–4. Groups stay fixed during calibration and model adaptation.

Paper, Fig. 2(a), PDF p. 7; App. B, pp. 17–19. Example: Qwen Layer 4, Head 0; group sizes 60, 2, 1, 1.

A squared quadratic form guarantees positive weights

Let $\bar{\mathbf{s}} = [1, s_1, \ldots, s_G]^\top$ and $\epsilon > 0$; learn a lower-triangular L.

OAL kernel

$$\kappa_{\mathrm{OAL}}(q, k) = \epsilon + \bar{\mathbf{s}}^\top L L^\top \bar{\mathbf{s}} = \epsilon + \|L^\top \bar{\mathbf{s}}\|_2^2 > 0.$$

With $M = LL^\top$, map back to coordinate coefficients:

$$a = \epsilon + M_{00}, \qquad b_r = 2M_{0,g(r)}, \qquad C_{rt} = M_{g(r),g(t)}.$$

- Positive denominator for every nonempty causal prefix.
- $(G+1)(G+2)/2$ parameters per head: 3 at $G = 1$, 21 at $G = 5$.
- $G = 1$ gives the learned global reference.

Paper, Eq. (7), PDF p. 7; App. D, pp. 21–22. Pointwise positivity does not imply Mercer positive semidefiniteness.

HD evaluation moves token sums inside a key moment

For value channel ℓ:

$$\sum_j (q_i^\top k_j)^2 V_{j,\ell} = q_i^\top \underbrace{\left(\sum_j V_{j,\ell} k_j k_j^\top \right)}_{d_h \times d_h \text{ key moment}} q_i.$$

Equivalently, with $(D_\ell)_{jj} = V_{j,\ell}$:

$$[(QK^\top) \odot (QK^\top)] V_{:,\ell} = \operatorname{diag}\Big(Q[K^\top D_\ell K] Q^\top \Big).$$

- Accumulate key moments once; read them with each query.
- Evaluate only diagonal entries; never build the outer $N \times N$ matrix.
- Restrict the inner sum to $j \leq i$ for causal attention.

Paper, Eq. (4), PDF p. 5; derivation in App. D, p. 21. HD = Hadamard–diagonal.

Three prefix statistics evaluate the chosen kernel exactly

Augment values: $u_j = [v_j; 1] \in \mathbb{R}^{d_v+1}$.

$$H_{0,i} = \sum_{j \le i} u_j, \quad H_{1,i,r} = \sum_{j \le i} k_{j,r} u_j, \quad H_{2,i,rt} = \sum_{j \le i} k_{j,r} k_{j,t} u_j.$$

$$\widetilde{u}_i = a H_{0,i} + \frac{1}{\sqrt{d_h}} \sum_r b_r q_{i,r} H_{1,i,r} + \frac{1}{d_h} \sum_{r,t} C_{rt} q_{i,r} q_{i,t} H_{2,i,rt},$$

$$\widehat{y}_i = \widetilde{u}_{i,1:d_v} / \widetilde{u}_{i,d_v+1}.$$

The last channel accumulates the normalization denominator.

Approximation comes from replacing softmax; the recurrence adds none.

Paper, Eqs. (5)–(6), PDF p. 6; Algorithm 1, p. 8; equivalence in Eq. (25), p. 21.

Grouping, output calibration, and adaptation are separate

1. Group dimensions from pretrained statistics

40 short prompts; per-layer, per-query-head assignments.

2. Fit the normalized attention output

$$L^\star = \arg\min_L \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \sum_i \|\widehat{y}_i(L; x) - y_i^{\mathrm{softmax}}(x)\|_2^2.$$

Frozen Q/K/V and groups; 8 training + 2 validation WikiText-2 blocks.

3. Adapt the converted model

Jointly optimize calibrated L and LoRA on WikiText-103 next-token prediction; keep groups fixed.

Paper, App. B, PDF pp. 19–20, Eq. (20); App. F.2, p. 25. Blocks contain 4096 tokens.

A compact kernel can require a large recurrent state

Quantity	HD implementation
Time	$O(Nd_h^2(d_v+1))$
Logical state	$O(d_h^2(d_v+1))$

Symmetric packing reduces the state to

$$\left(1 + d_h + \frac{d_h(d_h+1)}{2}\right)(d_v+1) \text{ entries.}$$

At $d_h = d_v = 64$ and $G = 5$

139,425 state entries; 21 learned coefficients.
Grouping leaves the coordinate-pair states intact.

Paper, Sec. 4.4, PDF pp. 7–8; App. D, pp. 21–22. State entries calculated from the paper's formula; not peak GPU memory.

Frozen Q/K/V isolate approximation fidelity

- Qwen2.5-0.5B Base; all query heads in zero-based Layers 4, 12, 20.
- 16 held-out WikiText-2 sequences, each 4096 tokens.
- Evaluate before and after the frozen output projection W_o.

$$\mathrm{MSE} = \frac{1}{B}\sum_b \frac{\|\widehat{Y}_b - Y_b\|_F^2}{n_b}, \quad \mathrm{RelMSE} = \frac{1}{B}\sum_b \frac{\|\widehat{Y}_b - Y_b\|_F^2}{\|Y_b\|_F^2}.$$

Here n_b counts scalar output elements; $B = 16$.

Aggregation matters

RelMSE normalizes by each sequence's exact output energy, then averages. After W_o, the denominator also uses the projected output.

Paper, Sec. 5.1, PDF pp. 8–9; App. B–C, pp. 20–21. Lower is better for both metrics.

OAL improves all four frozen-output error metrics

	Before W_o		After W_o	
Method	MSE	RelMSE	MSE	RelMSE
Performer	0.103	0.569	0.038	0.710
SpectraFormer	0.106	0.585	0.041	0.772
DiJiang	0.086	0.474	0.033	0.631
TaylorShift	0.088	0.490	0.034	0.649
OAL	**0.078**	**0.432**	**0.031**	**0.595**

Relative to DiJiang

Pre-projection RelMSE: -8.86%. Post-projection: -5.71%.
The abstract's 8.9% refers to the pre-projection metric.

Paper, Table 2, PDF p. 8. Percent reductions calculated from reported values. Learned Global is not included in this table.

Model conversion uses a matched, limited LoRA budget

Backbone	Replaced layers	Retained softmax
Qwen2.5-0.5B	3–20	First/last 3 layers
Llama-3.2-1B	1–14	First/last layer

- Same ranges for all converted methods; layer indices are zero-based.
- LoRA on Q/K/V/O in every layer, including retained softmax layers.
- Rank 8, $\alpha = 16$, dropout 0.05; 512 updates, batch 1, length 4096.
- WikiText-103 next-token training; 2,097,152 input tokens; seed 42.

These are hybrid models: retained softmax layers still have quadratic prefill computation.

Paper, App. F, PDF pp. 22–25. Softmax reference also receives LoRA. Best validation NLL selects the checkpoint.

Qwen: grouping improves every reported quality metric

Method	PPL ↓	PIQA ↑	HellaSwag ↑	ARC-E ↑
Softmax (LoRA)	10.89	70.18	51.38	62.63
Performer	37.03	57.78	30.37	38.93
SpectraFormer	34.78	59.19	31.77	41.84
DiJiang	36.05	59.90	31.47	41.41
TaylorShift	39.50	59.90	31.18	40.78
Learned Global	31.71	61.15	32.43	41.41
OAL	**27.50**	**64.69**	**37.20**	**50.38**

Against Learned Global: PPL -13.28%; task gains $+3.54$, $+4.77$, $+8.97$ percentage points.

Paper, Table 3, PDF p. 10, Qwen2.5-0.5B. WikiText-103 test PPL; zero-shot normalized accuracy (%). Bold: best converted model.

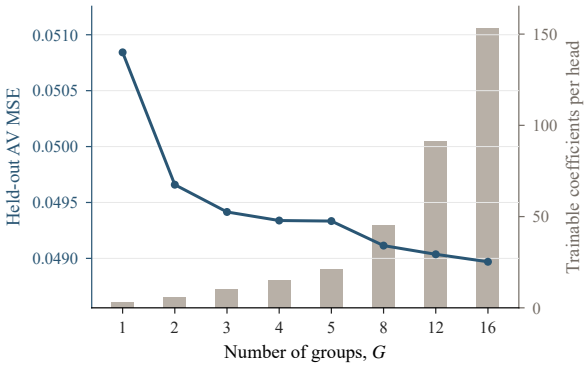
Llama: gains transfer, but their size depends on the task

Method	PPL ↓	PIQA ↑	HellaSwag ↑	ARC-E ↑
Softmax (LoRA)	8.19	74.92	62.84	59.43
Performer	50.39	57.07	27.42	32.58
SpectraFormer	45.09	59.36	30.41	36.49
DiJiang	56.27	58.92	28.75	36.32
TaylorShift	62.67	57.40	28.95	35.77
Learned Global	36.03	57.40	35.02	33.46
OAL	**32.99**	**63.66**	**35.42**	**41.67**

Against Learned Global: PPL -8.44%; task gains $+6.26$, $+0.40$, $+8.21$ percentage points. One adaptation seed.

Paper, Table 3, PDF p. 10, Llama-3.2-1B. WikiText-103 test PPL; zero-shot normalized accuracy (%). Bold: best converted model.

A small group budget captures most of the diagnostic gain

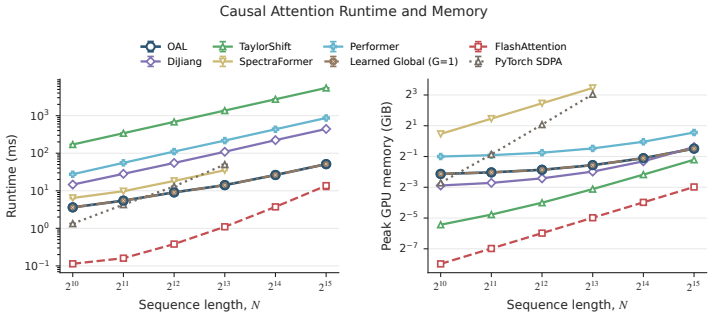


- One head: Qwen Layer 4, Head 0.
- Fixed G: split at the largest $G-1$ contribution gaps.
- Same data and fitting protocol across budgets.
- More groups lower error, with diminishing returns.

This diagnostic differs from bootstrap-based automatic group selection.

Paper, Fig. 2(b), PDF p. 7; App. B, p. 20. Left axis: held-out output MSE; right axis: learned coefficients/head.

OAL tracks Global in cost; FlashAttention remains ahead



A100 40GB; forward only; BF16 inputs; $B = 1$, $H_q = 14$, $H_{kv} = 2$, $d_h = d_v = 64$. OAL $G = 5$ and Global $G = 1$ share the HD backend.

Paper, Fig. 3, PDF p. 9; App. E, p. 22. Precision and optimization differ across other methods. Memory includes inputs and workspace.

The results establish fidelity gains under specific tradeoffs

- **Quality:** OAL leads the evaluated converted methods, while LoRA-softmax stays substantially stronger.
- **Efficiency:** sequence-linear operator scaling coexists with expensive quadratic head states and slower measured execution than FlashAttention.
- **Comparison scope:** adapted small-parameter kernels; no MLP feature-map baseline such as Hedgehog.
- **Next discriminating test:** compare contribution, random, and uniform partitions at the same group budget.

What to carry forward

Allocate approximation freedom according to pretrained representation statistics, then account separately for state and execution costs.

Paper, Sec. 7, PDF p. 11; Apps. D–F, pp. 21–25. Final experiment proposal is this reading's analysis.

Three takeaways for attention conversion

1. Contribution heterogeneity is associated with where linearization error concentrates.
2. Grouped positive quadratics provide a compact, output-calibrated extension of global quadratic attention.
3. Exact HD recurrence preserves the selected kernel; practical conversion quality and speed remain open challenges.

Weiqi Sun, Qianyue Wu, Wenjia Wang, Tianyang Hu.
Not All Dimensions Are Equal: Outlier-Aware Attention Linearization with Grouped Quadratic Kernels.
arXiv:2610.13674v1, 9 October 2026.

Author-provided code: github.com/wuqianyue09/OAL.

Original paper figures and reported experimental values are attributed at their point of use.

Backup: bootstrap support determines natural boundaries

Fix the contribution ordering π; search only the top-$\lceil \alpha d_h \rceil$ tail and its preceding boundary.

$$\rho_k = \frac{\chi_{\pi_{k+1}}}{\max(\chi_{\pi_k}, \varepsilon)}, \qquad R_k^{(b)} = \frac{\min_{r \in \mathcal{U}_k} \chi_r^{(b)}}{\max(\max_{r \in \mathcal{L}_k} \chi_r^{(b)}, \varepsilon)}.$$

$$S_k = \frac{1}{B} \sum_b \mathbf{1}[R_k^{(b)} \geq \tau], \qquad \text{keep if } \rho_k \geq \tau \text{ and } S_k \geq \gamma.$$

- 40 prompts, $B = 1000$ bootstrap replicates; $\tau = 1.5$.
- Qwen: $(\alpha, \gamma) = (0.25, 0.9)$; Llama: $(0.50, 0.5)$.
- Require at least 2 groups. Fallback cuts prioritize (S_k, ρ_k), even if natural thresholds fail. No explicit maximum group count.

Paper, App. B, Eqs. (16)–(19), PDF pp. 18–19. Bootstrap resamples prompts, retaining the original coordinate ordering.

Backup: coefficient fitting uses disjoint data splits

WikiText-2 blocks	Role	Number of sequences
0–15	Frozen evaluation	16
16–23	Coefficient training	8
24–25	Coefficient validation	2

- Adam, learning rate 0.02, FP32, clipping at 10; up to 1000 updates per initialization; 32 causal query positions/update.
- Three starts: learned-global embedding, Taylor, perturbed global.
- Validate every 100 updates; stop after 400 updates without improvement; choose the best checkpoint per head.
- Learned Global uses the same fitting data and optimization budget.

Paper, App. B, PDF pp. 19–20. Each sequence has 4096 tokens; short prompts used for grouping are separate.

Backup: the baselines use adapted causal interfaces

- **Performer:** 250 fixed positive orthogonal features/head; only LoRA is trained; FP64 features and prefix accumulation.
- **SpectraFormer:** OPRF–FastFoodL, 65 features; three trainable 64D vectors/layer, shared across heads; FP64 features and accumulation.
- **DiJiang:** DCT projection, feature softmax, 64 learned positive weights/head; FP32. Positional decay, GroupNorm, and output gate omitted.
- **TaylorShift:** normalized Q/K, fixed quadratic, learned temperature/head; causal aggregation; no extra output-amplitude scaling.

Paper, App. F.1, PDF pp. 23–25. Comparisons preserve pretrained Q/K/V/O, RoPE, and GQA rather than original full-model recipes.

Backup: positive weights do not imply a Mercer kernel

Teaching counterexample, not a paper experiment

For one-dimensional Q/K, take $\kappa(q, k) = \epsilon + (1 - qk)^2$, $\epsilon > 0$.

This fits the OAL form with

$$M = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \succeq 0.$$

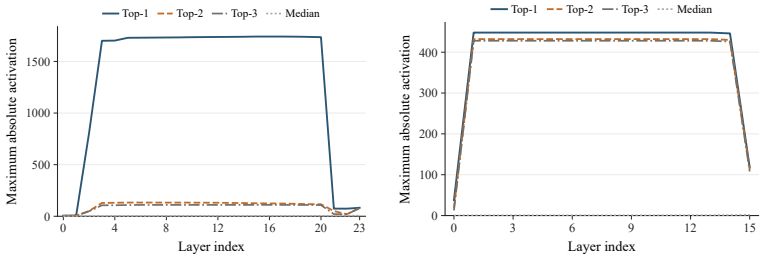
For the two inputs $q_1 = 1$ and $q_2 = -1$, the Gram matrix is

$$K = \begin{bmatrix} \epsilon & \epsilon + 4 \\ \epsilon + 4 & \epsilon \end{bmatrix}, \qquad \lambda(K) = \{2\epsilon + 4, -4\}.$$

Every weight is positive, but the Gram matrix is indefinite. HD evaluation requires separable algebraic terms, not Mercer positivity.

Paper, App. D, PDF p. 21 states the distinction. Counterexample derived for this discussion.

Backup: layer selection and task scoring affect interpretation



(a) Qwen2.5-0.5B (b) Llama-3.2-1B

- Layer profiles: per-dimension maximum absolute residual-stream output over 40 prompts; Top-1/2/3 versus median.
- Zero-shot candidate scores: mean token log-likelihood for PIQA/HellaSwag; character-normalized for ARC-E.

Paper, Fig. 7, PDF p. 23; App. F.3, p. 25. No chat template or added BOS/EOS tokens.