GRAPE: Group Representational Position Encoding

Yifan Zhang[1], Zixiang Chen[*2], Yifeng Liu[*2], Zhen Qin[*], Huizhuo Yuan[2], Kangping Xu[3], Quanquan Gu[†2], Andrew Chi-Chih Yao[†3]

[1]Princeton University [2]University of California, Los Angeles [3]IIIS, Tsinghua University [*]Core contribution [†]Corresponding authors

ICLR 2026

Code & Paper

Motivation & Overview

Transformers lack inherent positional awareness; existing methods like RoPE and ALiBi inject position information but lack principled justification. **GRAPE** provides a unified *group-theoretic* framework: position n acts via a one-parameter subgroup $\mathbf{G}(n) = \exp(n \cdot \omega \cdot \mathbf{L})$, yielding two natural families—**multiplicative** (rotations in $\mathrm{SO}(d)$) and **additive** (translations in GL).

▶ Why Not RoPE / ALiBi Alone?

- **RoPE** fixes coordinate planes and a log-uniform spectrum, limiting cross-subspace coupling and contextual warping of phase.
- **ALiBi** provides length extrapolation but its linear bias is content-independent and lacks a compositional group structure.
- **Table-based relatives** add window-dependent overhead.
- Current methods *isolate* the essential properties of stability, monotonic distance penalty, and expressivity.

In short: RoPE can only rotate pairs of dimensions independently; ALiBi applies a fixed, content-blind decay. GRAPE unifies and generalizes both.

GRAPE

Group Representational Position Encoding

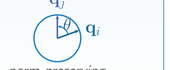
$$\mathbf{G}(n) = \exp(n\omega\mathbf{L}) \ , \quad \mathbf{G}(t-s) = \mathbf{G}(s)^{-1}\mathbf{G}(t)$$

Multiplicative

Rotation

in $\mathrm{SO}(d)$

$$\exp(n\omega\mathbf{L}) = \mathbf{I} + \tfrac{\sin\theta}{s}\mathbf{L} + \tfrac{1-\cos\theta}{s^2}\mathbf{L}^2$$



norm-preserving

Recovers RoPE

Additive

Translation

in $\mathrm{GL}(d+k)$

$$\exp(n\omega\mathbf{A}) = \mathbf{I} + n\omega\mathbf{A}$$

(exact!)

additive logit bias

Recovers ALiBi / FoX

▶ Design Desiderata

A principled PE framework should:

1. Preserve RoPE's orthogonality and exact relativity when desired
2. Also cover additive/forgetting mechanisms (ALiBi, FoX)
3. Admit learned and contextual generalizations with clean streaming

▶ Contributions

1. Unified group-theoretic view subsuming **RoPE**, **ALiBi**, **FoX** as exact special cases
2. **GRAPE-M**: closed-form rank-2 matrix exponential with fast $O(d)$ application
3. **GRAPE-A / GRAPE-AP**: additive biases from unipotent GL actions with streaming cacheability

Multiplicative GRAPE

Position n acts on query/key vectors via a rotation $\mathbf{G}(n) = \exp(n\,\omega\,\mathbf{L}) \in \mathrm{SO}(d)$, yielding a **norm-preserving, exactly relative** positional encoding with no matrix materialization.

▶ Rank-2 Skew Generator

From two learned vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, form a rank-2 skew-symmetric generator $\mathbf{L} = \mathbf{a}\mathbf{b}^\top - \mathbf{b}\mathbf{a}^\top$. Its square $\mathbf{L}^2 = -s^2\mathbf{P}_{\mathcal{U}}$ projects onto $\mathrm{span}\{\mathbf{a}, \mathbf{b}\}$ with $s = \sqrt{\|\mathbf{a}\|^2\|\mathbf{b}\|^2 - (\mathbf{a}^\top\mathbf{b})^2}$.

▶ Closed-Form (Rodrigues)

The matrix exponential reduces to a **quadratic** in $\mathbf{L}$:

$$\exp(n\omega\mathbf{L}) = \mathbf{I} + \tfrac{\sin(n\omega s)}{s}\mathbf{L} + \tfrac{1-\cos(n\omega s)}{s^2}\mathbf{L}^2$$

Rotation by angle $n\omega s$ in $\mathrm{span}\{\mathbf{a}, \mathbf{b}\}$; identity on the complement.

▶ Exact Relative Law

One-parameter subgroup structure guarantees $\mathbf{G}(t-s) = \mathbf{G}(s)^\top\mathbf{G}(t)$ and $\mathbf{G}(n)^\top\mathbf{G}(n) = \mathbf{I}$. Attention scores depend *only* on offset $j-i$, not absolute positions.

▶ RoPE as a Special Case

When rotation planes are canonical coordinate pairs with log-uniform frequencies:

$$\mathbf{G}(n) = \mathrm{blockdiag}\big(\mathbf{R}_2(n\theta_1), \ldots, \mathbf{R}_2(n\theta_{\frac{d}{2}})\big)$$

This is **exactly RoPE**. Learned commuting subspaces and non-commuting mixtures strictly extend this geometry.

▶ Streaming KV-Cache

At token j: cache $\widetilde{\mathbf{k}}_{j,h} = \mathbf{G}_{h,j}\mathbf{k}_{j,h}$. At step t: compute $\widetilde{\mathbf{q}}_{t,h} = \mathbf{G}_{h,t}\mathbf{q}_{t,h}$. No cache rewrite as t grows.

▶ Multi-Head Attention

Each head h applies its own $\mathbf{G}_{h,t} \in \mathrm{SO}(d)$: rotate $\widetilde{\mathbf{q}}_{t,h} = \mathbf{G}_{h,t}\mathbf{q}_{t,h}$, $\widetilde{\mathbf{k}}_{t,h} = \mathbf{G}_{h,t}\mathbf{k}_{t,h}$, then score $\ell_{t,j,h} = \mathbf{q}_{t,h}^\top(\mathbf{G}_{h,t}^\top\mathbf{G}_{h,j})\mathbf{k}_{j,h}/\sqrt{d}$. Values are **not rotated**; each head learns its own rotation plane and frequency.

▶ Non-Commuting Variants

Thin compression $\mathbf{L} = \mathbf{E}\mathbf{L}_r\mathbf{E}^\top$ ($\mathbf{E} \in \mathbb{R}^{d\times r}$ orthonormal) enables **cross-subspace feature coupling** at $O(rd)$ cost per head, generalizing beyond RoPE's fixed coordinate-pair structure.

▶ Spectral Properties

- All eigenvalues lie on the **unit circle** — spectral radius $\rho = 1$ always
- All **Lyapunov exponents = 0**: no exponential growth/decay at any sequence length

Cost: $O(d)$ per token — two inner products, no $d\times d$ matrix needed.

Additive GRAPE

Additive logit biases arise from **unipotent actions** (linear maps with all eigenvalues equal to 1) in $\mathrm{GL}(d+k)$ via a homogeneous lift, giving an exact relative law and streaming compatibility analogous to Multiplicative GRAPE.

▶ Nilpotent Generator

Fix a nilpotent $\mathbf{A}$ with $\mathbf{A}^2 = \mathbf{0}$. The exponential is **exact and linear**:

$$\mathbf{G}_{\mathrm{add}}(n) := \exp(n\,\omega\,\mathbf{A}) = \mathbf{I} + n\,\omega\,\mathbf{A}$$

No truncation — this is the *complete* series, not an approximation.

▶ Homogeneous Lift

Augment coordinates $\widehat{\mathbf{x}} = [\mathbf{x}; 1]$. The unipotent action becomes a translation: $\mathbf{G}_{\mathrm{add}}(n)\widehat{\mathbf{x}} = [\mathbf{x} + n\omega\,\mathbf{u}_{\mathrm{shift}}; 1]$. The relative law $\widehat{\mathbf{q}}_i^\top\mathbf{G}_{\mathrm{add}}(j-i)^{-\top}\widehat{\mathbf{k}}_j$ depends only on offset $j-i$.

▶ ALiBi Recovery

Rank-1 nilpotent in $\mathrm{GL}(d+2)$ with $\widehat{\mathbf{q}} = [\mathbf{q}; 1; 0]$, $\widehat{\mathbf{k}} = [\mathbf{k}; 0; 1]$:

$$\widehat{\mathbf{q}}_i^\top\mathbf{G}_{\mathrm{add}}(j-i)^{-\top}\widehat{\mathbf{k}}_j = \mathbf{q}_i^\top\mathbf{k}_j + (j-i)\beta_h$$

This is **exactly ALiBi**. Similarly, FoX is recovered as an exact special case.

▶ Content-Gated (GRAPE-A-QK)

Non-negative gates $\lambda_q(\mathbf{q}) = \mathrm{softplus}(\mathbf{v}^\top\mathbf{q}/\sqrt{d})$, $\lambda_k(\mathbf{k}) = \mathrm{softplus}(\mathbf{u}^\top\mathbf{k}/\sqrt{d})$ yield content-dependent biases: $\widetilde{\mathbf{q}}_i^\top\widetilde{\mathbf{k}}_j = \mathbf{q}_i^\top\mathbf{k}_j + (j-i)\omega[\lambda_q(\mathbf{q}_i) + \lambda_k(\mathbf{k}_j)]$.

▶ Path-Integral (GRAPE-AP)

Adaptive biases via causal prefix sums with edge potentials $\psi_h(t, \ell) \le 0$:

$$b_h(t, j) = \sum_{\ell=j+1}^{t} \psi_h(t, \ell) \le 0$$

The path product of unipotent factors collapses to $\mathbf{I} - b_h(t, j)\mathbf{E}$. Subsumes ALiBi and FoX; enables $O(t)$ streaming per step. GRAPE-AP combines content-adaptive additive biases with the multiplicative rotation of GRAPE-M, achieving the **best empirical performance**.

▶ FoX Recovery

With endpoint-independent edges $\psi_h(t, \ell) \equiv \log f_{\ell,h} \le 0$, we get $b_h(t, j) = \sum_{\ell=j+1}^{t} \log f_{\ell,h}$ — **exactly** the FoX forgetting bias. Setting $f_{\ell,h} \equiv e^{-\beta_h}$ further recovers ALiBi: $D_{ij,h} = -\beta_h(i-j)$.

Streaming compatible: all additive variants support incremental $O(1)$ bias updates per new token — no KV-cache rewrite needed.

Summary. By varying edge potentials ψ_h, GRAPE-A smoothly interpolates between ALiBi (fixed decay), FoX (learned forgetting), and GRAPE-AP (content-adaptive biases).

Experiments

Setup. FineWeb-Edu 100B corpus (50B training tokens), Llama architecture (nanoGPT), context length 4096.

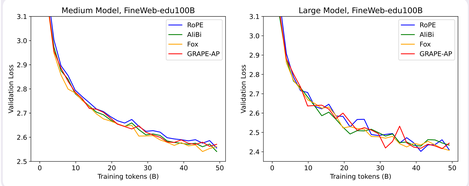
- **Medium**: 350M params, 24 layers, 8 heads, d=1024
- **Large**: 770M params, 36 layers, 10 heads, d=1280

Evaluated 0-shot on ARC-E, ARC-C, HellaSwag, OBQA, PIQA, WinoGrande, SciQ.

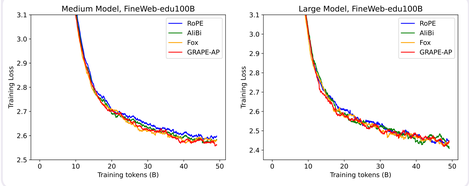
▶ Average 0-Shot Accuracy

Method	350M	770M
RoPE	51.73	55.76
ALiBi	52.87	56.44
FoX	52.96	56.30
GRAPE-AP	**53.25**	**56.91**

▶ Validation Loss



▶ Training Loss



GRAPE-AP consistently achieves the **lowest validation and training loss** across both model scales, demonstrating stable optimization throughout the entire training trajectory.

Note: All methods use identical architecture and hyperparameters; only the position encoding differs.

Key Results & Takeaways

▶ Unified Framework

First principled PE framework based on matrix Lie groups—recovers **RoPE**, **ALiBi**, and **FoX** as *exact* special cases via choice of group $\mathcal{G}$ and representation ρ.

▶ GRAPE-AP Outperforms All Baselines

Best average 0-shot accuracy at both model scales:

- **+0.29** over FoX at 350M (53.25 vs. 52.96)
- **+0.47** over ALiBi at 770M (56.91 vs. 56.44)

Combines multiplicative rotation with path-integral additive biases for richer position encoding.

▶ Training Stability

GRAPE variants exhibit steadier convergence throughout training. RoPE shows notable instability in loss curves, while GRAPE-AP maintains smooth optimization.

▶ $O(d)$ Efficiency

Closed-form rank-2 exponential—no $d\times d$ matrix materialization required. Fully compatible with KV-cache and streaming inference. Negligible overhead vs. standard attention.

▶ Per-Benchmark Scores (350M)

Task	RoPE	FoX	GRAPE-AP
ARC-E	56.36	58.38	**59.26**
ARC-C	30.38	30.89	31.31
HellaSwag	44.65	**45.80**	45.42
OBQA	35.20	35.00	35.60
PIQA	68.77	69.37	68.17
WinoGr.	52.33	52.88	53.28
SciQ	74.40	78.40	**79.70**
Avg.	51.73	52.96	**53.25**

▶ With KV-Shift (350M)

Method	Avg.
RoPE + KV-shift	52.89
ALiBi + KV-shift	53.18
FoX + KV-shift	53.32
GRAPE-AP + KV-shift	**53.46**

KV-shift (a key–value centering technique from FoX) is orthogonal to GRAPE. Gains are **complementary**.

▶ 2D/3D Vision Extensions

GRAPE extends naturally beyond 1D sequences: compose one exponential per spatial axis, e.g. $\mathbf{G}_{2D}(u, v) = \exp(u\omega_x\mathbf{L}^{(x)})\exp(v\omega_y\mathbf{L}^{(y)})$. Per-token cost stays $O(d)$. Disjoint planes reduce to separable 2D/3D RoPE; learned planes give the full GRAPE-M variant.