GROUP REPRESENTATIONAL POSITION ENCODING

Anonymous authors
Paper under double-blind review

ABSTRACT

We present **GRAPE** (**G**roup **R**epresent**A**tional **P**osition **E**ncoding), a unified framework for positional encoding based on group actions. GRAPE brings together two families of mechanisms: (i) *multiplicative* rotations (Multiplicative GRAPE) in $\mathrm{SO}(d)$ and (ii) *additive* logit biases (Additive GRAPE) arising from unipotent actions in the general linear group GL. In Mul-GRAPE, a position $n \in \mathbb{Z}$ (or $t \in \mathbb{R}$) acts as $\mathbf{G}(n) = \exp(n\,\omega\,\mathbf{L})$ with a rank-2 skew generator $\mathbf{L} = \mathbf{a}\mathbf{b}^\top - \mathbf{b}\mathbf{a}^\top \in \mathfrak{so}(d)$, yielding a relative, compositional, norm-preserving map with a closed-form matrix exponential. RoPE is recovered exactly when the $d/2$ planes are the canonical coordinate pairs with log-uniform spectrum. Learned commuting subspaces and compact non-commuting mixtures strictly extend this geometry at $O(d)$ and $O(rd)$ cost per head, respectively. In Additive GRAPE, additive logits arise as rank-1 (or low-rank) unipotent actions, recovering ALiBi and the Forgetting Transformer (FoX) as exact special cases while preserving an exact relative law and streaming cacheability. Altogether, GRAPE supplies a principled design space for positional geometry in long-context models, subsuming RoPE and ALiBi as special cases.

1 INTRODUCTION

Positional information is essential for sequence modeling with Transformers (Vaswani et al., 2017), whose self-attention is otherwise permutation-invariant. Early work injected absolute positional codes (sinusoidal or learned) into token representations (Vaswani et al., 2017). Later, relative encodings depending on offsets (Shaw et al., 2018) and linear logit biases such as ALiBi (Press et al., 2021) were introduced, the latter offering strong length extrapolation with negligible overhead.

Rotary Position Embedding (RoPE) (Su et al., 2021) realizes relative positions as orthogonal planar rotations of queries and keys, preserving norms and yielding exact origin invariance of attention scores. Despite its appeal, RoPE fixes coordinate planes and typically a log-uniform spectrum, limiting cross-subspace coupling and contextual warping of phase. More broadly, absolute codes break translation equivariance; table-based relatives add window-dependent overhead. These observations motivate a unified formulation that (i) preserves RoPE's orthogonality and exact relativity when desired, (ii) *also* covers additive/forgetting mechanisms such as ALiBi (Press et al., 2021) and Forgetting Transformer (FoX) (Lin et al., 2025), and (iii) admits learned and contextual generalizations with clean streaming.

We therefore propose **G**roup **R**epresent**A**tional **P**osition **E**ncoding (**GRAPE**), a group-theoretic framework that unifies two complementary families of positional mechanisms. The multiplicative family (Multiplicative GRAPE) models positions as norm-preserving rotations in $\mathrm{SO}(d)$ acting on $(\mathbf{q}, \mathbf{k})$; the additive family (Additive GRAPE/Path-Integral Additive GRAPE) models positions as unipotent actions in the general linear group GL that yield linear-in-offset logit biases (including content-gated and path-integral forms). This perspective recovers RoPE and ALiBi as exact special cases, proves that FoX is an exact instance of Additive-GRAPE, and supplies principled, streaming-friendly contextual extensions on both sides.

Concretely: *(a)* Multiplicative GRAPE encodes $n \in \mathbb{Z}$ (or $t \in \mathbb{R}$) as an element of $\mathrm{SO}(d)$ via a rank-2 skew generator; and *(b)* Additive GRAPE (and Path-Integral Additive GRAPE) lifts to the general linear group GL using homogeneous coordinates to produce linear-in-offset logit biases (recovering ALiBi and FoX).

For Multiplicative GRAPE, positions are mapped as

$$\mathbf{G}(n) = \exp\big(n\,\omega\,\mathbf{L}\big) \in \mathrm{SO}(d), \qquad \mathbf{L} = \mathbf{a}\mathbf{b}^\top - \mathbf{b}\mathbf{a}^\top \in \mathfrak{so}(d),$$

where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ define a rank-2 skew generator $\mathbf{L}$ and $\omega > 0$ is a frequency. The action is an isometry, and $\mathbf{G}(n+m) = \mathbf{G}(n)\mathbf{G}(m)$ guarantees exact origin invariance of attention logits. We derive a closed-form Rodrigues-type formula (Rodrigues, 1840; Hall, 2013), enabling fast linear-time application with stable derivatives and no explicit matrix materialization. RoPE is recovered when $d/2$ commuting rank-2 generators act on disjoint coordinate planes with prescribed frequencies.

For Additive GRAPE, positions are mapped via the matrix exponential $\mathbf{G}_{\mathrm{add}}(n) = \exp(n\omega\mathbf{A}) = \mathbf{I} + n\omega\mathbf{A}$ in a lifted homogeneous space. Here, the generator $\mathbf{A} \in \mathfrak{gl}(d+1)$ is a nilpotent matrix of rank one. While this additive transformation is not an isometry, it preserves the exact relative law, ensuring attention scores depend only on position offsets. This formulation provides a rigorous group-theoretic foundation for additive biases, recovering ALiBi and FoX as exact instances.

Our contributions are highlighted as follows:

1. We propose **GRAPE** as a unified group-theoretic view that subsumes *multiplicative* orthogonal rotations in $\mathrm{SO}(d)$ and *additive* unipotent (all eigenvalues equal to 1) mechanisms in general linear group GL, recovering RoPE and ALiBi as exact special cases and proving FoX is an exact instance (Appendix A).
2. **Multiplicative GRAPE.** We derive a closed-form rank-2 matrix exponential with fast application and stable differentiation; we show RoPE equals commuting MS-GRAPE in a possibly learned orthogonal basis.
3. **Additive GRAPE.** We show that linear-in-offset logit biases arise from rank-1 (or low-rank) unipotent actions in the general linear group GL with an exact relative law and streaming cacheability. This includes query- or key-gated slopes, a commuting dictionary of additive components, and exact recoveries of ALiBi and FoX in closed form (Sections 5, 5.2, Appendix A). We also formalize path-integral additive biases that remain causal and support efficient training. (Section 6).

2 Multiplicative Group Representational Position Encoding

We propose the **Multiplicative GRAPE**, as a Lie-group positional map with a closed-form rank-2 matrix exponential, an exact relative law, and a streaming/cache methodology. The core intuition is to encode position as a norm-preserving rotation in the special orthogonal group $\mathrm{SO}(d)$ (Hall, 2013). A single skew-symmetric generator $\mathbf{L} \in \mathfrak{so}(d)$ produces the entire family of rotations via the matrix exponential. We begin with notation and the rank-2 generator.

2.1 Preliminaries and Rank-2 Generator

The generator $\mathbf{L}$ is formally defined as an element of the corresponding Lie algebra, $\mathfrak{so}(d)$. Let $\mathfrak{so}(d) = \{\mathbf{L} \in \mathbb{R}^{d\times d} : \mathbf{L}^\top = -\mathbf{L}\}$ denote the Lie algebra of $\mathrm{SO}(d)$. The simplest non-trivial generator defines a rotation within a single 2D plane. We construct such a rank-2 generator from two vectors, $\mathbf{a}$ and $\mathbf{b}$, that span this plane of action. For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, define the rank-2 generator $\mathbf{L} \equiv \mathbf{L}(\mathbf{a}, \mathbf{b})$ as

$$\mathbf{L}(\mathbf{a}, \mathbf{b}) = \mathbf{a}\mathbf{b}^\top - \mathbf{b}\mathbf{a}^\top, \alpha = \|\mathbf{a}\|^2,\ \beta = \|\mathbf{b}\|^2,\ \gamma = \mathbf{a}^\top\mathbf{b}, \Delta = \alpha\beta - \gamma^2 \geq 0,\ s = \sqrt{\Delta}. \quad (2.1)$$

Rank-2 structure. Let $\mathcal{U} = \mathrm{span}\{\mathbf{a}, \mathbf{b}\}$. The rank-2 generator $\mathbf{L}$ has a useful geometric property: applying it twice projects onto the action plane $\mathcal{U}$ and scales. A direct calculation shows

$$\mathbf{L}^2 = -\,s^2\,\mathbf{P}_{\mathcal{U}},$$

where $\mathbf{P}_{\mathcal{U}}$ is the orthogonal projector to the space $\mathcal{U}$. Hence spectrum of $\mathbf{L}$ (the set of its eigenvalues), denoted $\sigma(\mathbf{L})$, is $\{\pm is, 0, \ldots, 0\}$ and the minimal polynomial is $\lambda(\lambda^2 + s^2)$. A detailed derivation is given in Appendix D.

Gauge symmetries and initialization. Write $\mathbf{A} \triangleq [\mathbf{a}\ \mathbf{b}] \in \mathbb{R}^{d\times 2}$ and $\mathbf{J} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ so that $\mathbf{L} = \mathbf{A}\mathbf{J}\mathbf{A}^\top$. For any $\mathbf{M} \in \mathrm{SL}(2)$, $\mathbf{M}\mathbf{J}\mathbf{M}^\top = \mathbf{J}$ and thus $\mathbf{A} \mapsto \mathbf{A}\mathbf{M}$ leaves $\mathbf{L}$ invariant; for general

$\mathbf{M} \in \mathrm{GL}(2)$, $\mathbf{L}$ scales by $\det(\mathbf{M})$. Therefore the oriented plane $\mathcal{U} = \mathrm{span}\{\mathbf{a}, \mathbf{b}\}$ and the scalar $s = \sqrt{\alpha\beta - \gamma^2}$ determine the action. We fix a gauge at initialization by $\|\mathbf{a}\| = \|\mathbf{b}\| = 1$ and $\mathbf{a}^\top \mathbf{b} = 0$ (absorbing scale into ω) and optionally enforce it softly via penalties during training.

Canonical $90°$ rotation operator. Fix a block-diagonal complex structure $\mathcal{J} \in \mathfrak{so}(d)$ with $\mathcal{J}^\top = -\mathcal{J}$ and $\mathcal{J}^2 = -\mathbf{I}$ (for odd d, act on the top-left $2\lfloor d/2 \rfloor$ coordinates and leave the final coordinate unchanged). Concretely, $\mathcal{J} = \bigoplus_{i=1}^{\lfloor d/2 \rfloor} \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. For any $\mathbf{a} \in \mathbb{R}^d$, write $\mathbf{a}_\perp := \mathcal{J}\mathbf{a}$, which equals "$\mathbf{a}$ rotated by $90°$" within the canonical 2D blocks and satisfies $\mathbf{a}^\top \mathbf{a}_\perp = 0$ and $\|\mathbf{a}_\perp\| = \|\mathbf{a}\|$.

2.2 Exact relative law

For a fixed $\mathbf{L} \in \mathfrak{so}(d)$, define $\mathbf{G}(n) = \exp(n\mathbf{L}) \in \mathrm{SO}(d)$, which forms a one-parameter subgroup. The exact relative law property for positional encoding implies:

$$\mathbf{G}(t-s) = \mathbf{G}(s)^\top \mathbf{G}(t), \qquad \mathbf{G}(n)^\top \mathbf{G}(n) = \mathbf{I}.$$

This algebraic property underpins relative positional encoding: interactions depend only on offsets. In Multiplicative GRAPE, we introduce a frequency ω to scale the generator. The resulting operator obeys the following position law:

$$\mathbf{G}(n) = \exp(n\omega\mathbf{L}), \quad \mathbf{G}(n+m) = \mathbf{G}(n)\mathbf{G}(m), \quad \mathbf{G}(0) = \mathbf{I}, \quad \text{and} \quad \mathbf{G}(-n) = \mathbf{G}(n)^\top.$$

2.3 Closed-form matrix exponential

Based on the minimal polynomial mentioned in Section 2.1, the exponential map $\exp(\mathbf{L})$ for a rank-2 generator can be expressed as a quadratic in $\mathbf{L}$. This yields a convenient closed-form solution, often referred to as a Rodrigues-type formula (Rodrigues, 1840; Hall, 2013):

$$\exp(\mathbf{L}) = \mathbf{I} + \frac{\sin s}{s}\,\mathbf{L} + \frac{1 - \cos s}{s^2}\,\mathbf{L}^2.$$

Geometrically, the formula is best understood via $\mathbf{L}^2$ as a projector onto $\mathcal{U}$. Since $\mathbf{L}^2 = -s^2 \mathbf{P}_\mathcal{U}$, the exponential can be written as

$$\exp(\mathbf{L}) = \mathbf{I} - (1 - \cos s)\,\mathbf{P}_\mathcal{U} + \frac{\sin s}{s}\,\mathbf{L},$$

which reveals its action explicitly: it is a rotation by angle s within the plane $\mathcal{U} = \mathrm{span}\{\mathbf{a}, \mathbf{b}\}$ and the identity on the orthogonal complement $\mathcal{U}^\perp$. The vectors $\mathbf{a}$ and $\mathbf{b}$ thus define the plane of action for the positional rotation.

Cost of application. For a single rank-2 plane, computing $\mathbf{y} = \mathbf{G}(n)\mathbf{x}$ requires two inner products $u = \langle \mathbf{a}, \mathbf{x} \rangle$, $v = \langle \mathbf{b}, \mathbf{x} \rangle$, followed by $\mathbf{y} = \mathbf{x} + f_1(n)(\mathbf{a}v - \mathbf{b}u) + f_2(n)[\gamma(\mathbf{a}v + \mathbf{b}u) - \beta\mathbf{a}u - \alpha\mathbf{b}v]$, where (α, β, γ) are plane scalars and $f_{1,2}$ are trigonometric scalars (with series guards as $s \to 0$). This is $O(d)$ flops with a small constant and no materialization of $\mathbf{G}(n)$; derivative expressions are in Appendix D.

2.4 The $\mathbf{b} = \mathcal{J}\mathbf{a}$ constraint

We now consider an important special case by setting $\mathbf{b} = \mathcal{J}\mathbf{a}$. This constraint, which makes the plane vectors $\mathbf{a}$ and $\mathbf{b}$ orthogonal and equal in norm, significantly simplifies the generator's structure and reveals a direct connection to the canonical RoPE formulation. With this constraint, the scalars simplify: $\gamma = \mathbf{a}^\top \mathbf{b} = \mathbf{a}^\top \mathcal{J}\mathbf{a} = 0$, $\beta = \|\mathbf{b}\|^2 = \|\mathbf{a}\|^2 = \alpha$, and hence $s = \sqrt{\alpha\beta - \gamma^2} = \alpha$. Moreover, on the 2D subspace $\mathcal{U} = \mathrm{span}\{\mathbf{a}, \mathcal{J}\mathbf{a}\}$ one has

$$\mathbf{L}(\mathbf{a}, \mathcal{J}\mathbf{a})\mathbf{a} = -(\mathcal{J}\mathbf{a})\alpha, \qquad \mathbf{L}(\mathbf{a}, \mathcal{J}\mathbf{a})\,\mathcal{J}\mathbf{a} = \alpha\,\mathbf{a},$$

so $\mathbf{L}(\mathbf{a}, \mathcal{J}\mathbf{a})|_\mathcal{U} = -\alpha\,\mathcal{J}|_\mathcal{U}$ and $\mathbf{L}(\mathbf{a}, \mathcal{J}\mathbf{a})|_{\mathcal{U}^\perp} = 0$. Therefore

$$\exp\big(n\omega\mathbf{L}(\mathbf{a}, \mathcal{J}\mathbf{a})\big) = \mathbf{I} - \big(1 - \cos(n\omega\alpha)\big)\mathbf{P}_\mathcal{U} - \sin(n\omega\alpha)\,\mathcal{J}\mathbf{P}_\mathcal{U},$$

which is a pure planar rotation by angle $n\omega\alpha$ on $\mathcal{U}$ and the identity on $\mathcal{U}^\perp$.

Corollary 2.1 (Frequency–norm coupling)**.** If $\|\mathbf{a}\| = 1$, the rotation angle reduces to $n\omega$. Without normalization, the effective frequency is $\omega_{\text{eff}} = \omega\|\mathbf{a}\|^2$, so the scale of a can be absorbed into ω.

2.5 APPLICATION TO RELATIVE ENCODING AND EQUIVARIANCE

We now demonstrate how the **Mul-GRAPE** operator $\mathbf{G}(n)$ is applied in practice. As established in Section 2.2, the operator's group structure guarantees the exact relative law. We first transform the query and key vectors, $\mathbf{q}_i$ and $\mathbf{k}_j$, into position-aware representations, $\widetilde{\mathbf{q}}_i$ and $\widetilde{\mathbf{k}}_j$:

$$\widetilde{\mathbf{q}}_i := \mathbf{G}(i)\mathbf{q}_i, \qquad \widetilde{\mathbf{k}}_j := \mathbf{G}(j)\mathbf{k}_j.$$

It follows from the exact relative law established in Section 2.2 that the attention score between these position-aware vectors simplifies to:

$$\widetilde{\mathbf{q}}_i^\top \widetilde{\mathbf{k}}_j = \mathbf{q}_i^\top \mathbf{G}(i)^\top \mathbf{G}(j)\mathbf{k}_j = \mathbf{q}_i^\top \mathbf{G}(j-i)\mathbf{k}_j.$$

Hence, the attention score depends solely on the relative offset $j - i$, not on the absolute positions.

Streaming and caching. At inference, cache $\mathbf{k}_j^\star = \mathbf{G}(j)\mathbf{k}_j$ once when token j arrives. At step t, form $\widetilde{\mathbf{q}}_t = \mathbf{G}(t)\mathbf{q}_t$ and compute logits $\widetilde{\mathbf{q}}_t^\top \mathbf{k}_j^\star$. No cache rotation is needed when t increments; complexity matches RoPE. A full integration into multi-head attention (per-head formulation, logits, and streaming) is detailed in Section 4.

3 MULTI-SUBSPACE MULTIPLICATIVE GRAPE

A single rank-2 generator acts on a 2D subspace, leaving the rest of the d-dimensional space untouched. To encode position across the entire hidden dimension, we can combine multiple generators. This leads to the Multi-Subspace (MS) **Mul-GRAPE** model, which forms the basis for both RoPE and more expressive types. Detailed rank-2 algebra appears in Appendix D.

3.1 COMMUTING MS-MUL-GRAPE AND ROPE AS A SPECIAL CASE

The simplest way to combine generators is to ensure they act on mutually orthogonal subspaces, which guarantees they commute. Let d be even. For $i = 1, \ldots, d/2$, we can define a set of rank-2 generators $\{\mathbf{L}_i\}$, each acting on a distinct 2D plane. RoPE is the canonical example of this construction.

Let the 2×2 canonical skew matrix be $\mathbf{J} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and the coordinate selector be $\mathbf{U}_i = [\mathbf{e}_{2i-1}\ \mathbf{e}_{2i}] \in \mathbb{R}^{d\times 2}$. We set the rank-2 generators as $\mathbf{L}_i = \mathbf{U}_i\mathbf{J}\mathbf{U}_i^\top = \mathbf{L}(\mathbf{e}_{2i-1}, \mathbf{e}_{2i})$ and assign per-plane frequencies $\theta_i > 0$. The total generator is the commuting sum:

$$\mathbf{L}_{\mathrm{RoPE}} = \sum_{i=1}^{d/2} \theta_i \mathbf{L}_i \qquad \text{with} \qquad [\mathbf{L}_i, \mathbf{L}_j] = 0 \ \text{ for } i \neq j.$$

Then

$$\mathbf{G}(n) = \exp\left(n\mathbf{L}_{\mathrm{RoPE}}\right) = \prod_{i=1}^{d/2} \exp(n\theta_i \mathbf{L}_i) = \mathrm{blockdiag}\left(\mathbf{R}_2(n\theta_1), \ldots, \mathbf{R}_2(n\theta_{d/2})\right), \tag{3.1}$$

where $\mathbf{R}_2(\cdot)$ is a standard 2×2 rotation matrix, and the last equality holds because each term $\exp(n\theta_i\mathbf{L}_i)$ is identity except for a single 2×2 rotation block on its diagonal. Eq. (3.1) is precisely the RoPE mapping: a block-diagonal product of planar rotations with per-subspace angles $n\theta_i$.

Equality holds when the planes $\{\mathbf{U}_i\}$ are the coordinate 2D blocks and $\{\theta_i\}$ follow the canonical log-uniform spectrum.

Proposition 3.1 (RoPE via $\mathcal{J}$-paired planes)**.** Choose $d/2$ mutually orthogonal vectors $\{\mathbf{a}_i\}$ and set $\mathbf{b}_i = \mathcal{J}\mathbf{a}_i$ with per-plane angles θ_i. Then the commuting MS-GRAPE $\mathbf{G}(n) = \prod_{i=1}^{d/2} \exp(n\theta_i \mathbf{L}(\mathbf{a}_i, \mathcal{J}\mathbf{a}_i))$ equals the standard RoPE map in a (possibly learned) orthogonal basis. If the planes are the canonical coordinate pairs and $\{\theta_i\}$ follow the log-uniform spectrum, we recover the canonical RoPE exactly.

Spectral parameterization. Classical RoPE chooses θ_i on a log-uniform grid across i. In GRAPE, θ_i can be learned or shared/tied across heads or layers. The MS-GRAPE view also allows replacing the coordinate selectors $\mathbf{U}_i$ by a learned orthogonal basis $\mathbf{B} \in \mathrm{SO}(d)$ so that $\mathbf{L} = \sum_i \theta_i \mathbf{B}\mathbf{U}_i\mathbf{J}\mathbf{U}_i^\top\mathbf{B}^\top$, preserving commutativity while learning subspaces.

Theorem 3.2 (RoPE is commuting MS-Mul-GRAPE)**.** Let $\mathbf{L}_{\mathrm{RoPE}} = \sum_{i=1}^{d/2} \theta_i \mathbf{U}_i \mathbf{J} \mathbf{U}_i^\top$ with mutually orthogonal planes. Then for any $n \in \mathbb{Z}$, $\exp\left(n\mathbf{L}_{\mathrm{RoPE}}\right) = \bigoplus_{i=1}^{d/2} \mathbf{R}_2(n\theta_i)$.

Proof. Orthogonality implies $[\mathbf{L}_i, \mathbf{L}_j] = 0$ for $i \neq j$. Hence $\exp(n \sum_i \theta_i \mathbf{L}_i) = \prod_i \exp(n\theta_i \mathbf{L}_i)$; each factor is a 2×2 rotation on its plane. □

4 Application in Multi-Head Attention

Building upon the algebraic foundation for relative encoding established in Section 2.5, this section details the concrete integration of the rotational map $\mathbf{G}(n)$ into the full Multi-Head Attention (MHA) architecture, covering the per-head formulation, streaming policy, and implementation complexity.

Per-head formulation. Let H be the number of heads and d the per-head width. For head $h \in [H]$, let $(\mathbf{q}_{t,h}, \mathbf{k}_{t,h}, \mathbf{v}_{t,h}) \in \mathbb{R}^d$ denote the query/key/value at position t. A **Mul-GRAPE** position map is realized as an orthogonal operator $\mathbf{G}_{h,t} \in \mathrm{SO}(d)$ applied to $(\mathbf{q}_{t,h}, \mathbf{k}_{t,h})$:

$$\widetilde{\mathbf{q}}_{t,h} = \mathbf{G}_{h,t}\, \mathbf{q}_{t,h}, \qquad \widetilde{\mathbf{k}}_{t,h} = \mathbf{G}_{h,t}\, \mathbf{k}_{t,h}, \qquad \widetilde{\mathbf{v}}_{t,h} = \mathbf{v}_{t,h} \ \ \text{(unchanged)}.$$

The headwise attention logits and outputs are then

$$\ell_{t,j,h} = \frac{\widetilde{\mathbf{q}}_{t,h}^\top \widetilde{\mathbf{k}}_{j,h}}{\sqrt{d}} = \frac{\mathbf{q}_{t,h}^\top \big(\mathbf{G}_{h,t}^\top \mathbf{G}_{h,j}\big) \mathbf{k}_{j,h}}{\sqrt{d}}, \qquad \mathbf{y}_{t,h} = \sum_{j \le t} \mathrm{softmax}\big(\ell_{t,\cdot,h}\big)_j \, \widetilde{\mathbf{v}}_{j,h}, \tag{4.1}$$

with the usual output projection applied after concatenation across heads.

Exact relative law. If $\mathbf{G}_{h,t}$ arises from a one-parameter subgroup $\mathbf{G}_h(n) = \exp(n\,\mathbf{L}_h)$ (commuting MS-Mul-GRAPE, including RoPE and learned commuting bases), then

$$\mathbf{G}_{h,t}^\top \mathbf{G}_{h,j} = \mathbf{G}_h(j{-}t) \qquad \Longrightarrow \qquad \ell_{t,j,h} = \frac{q_{t,h}^\top \mathbf{G}_h(j{-}t)\, \mathbf{k}_{j,h}}{\sqrt{d}},$$

so logits depend only on the offset $j{-}t$ (exact origin invariance).

Streaming cache policy. Applying the rotational map $\mathbf{G}(t)$ independently to each query and key vector is the core property that enables an efficient streaming cache policy. For any Type where $\mathbf{G}_t$ is known at token arrival (non-contextual and phase-modulated), cache $\widetilde{\mathbf{k}}_j = \mathbf{G}_j \mathbf{k}_j$ once and never rewrite it; at step t, compute $\widetilde{\mathbf{q}}_t = \mathbf{G}_t q_t$ and use logits $\ell_{t,j,h} = \widetilde{\mathbf{q}}_t^\top \widetilde{\mathbf{k}}_j / \sqrt{d}$.

5 Additive Group Representational Position Encoding

This section shows that additive positional mechanisms (absolute shifts of features and additive logit biases, including ALiBi (Press et al., 2021)) also admit a group-theoretic formulation as GRAPE. The key is a homogeneous lift to an augmented space and a one-parameter subgroup of the general linear group GL that acts by unipotent (all eigenvalues equal to 1) transformations. This yields an exact relative law and streaming/cache rules analogous to Section 2.5.

5.1 Homogeneous lift and a unipotent action

To produce additive biases from a multiplicative group action, we employ the homogeneous lift. This is a standard method in linear algebra for representing affine transformations (such as translations) as linear transformations in a higher-dimensional space. Let $\widehat{\mathbf{x}} := [\mathbf{x}; 1] \in \mathbb{R}^{d+1}$ denote a homogeneous augmentation of $\mathbf{x} \in \mathbb{R}^d$. We now work within the general linear group $\mathrm{GL}(d+1)$ and its corresponding Lie algebra $\mathfrak{gl}(d+1)$, which is the set of all $(d+1) \times (d+1)$ real matrices. Fix a generator

$$\mathbf{A} \;=\; \begin{bmatrix} \mathbf{0}_{d\times d} & \mathbf{u} \\ \mathbf{0}_{1\times d} & 0 \end{bmatrix} \in \mathfrak{gl}(d{+}1), \qquad \mathbf{A}^2 = \mathbf{0}, \tag{5.1}$$

where $\mathbf{u} \in \mathbb{R}^d$. Its exponential is unipotent:

$$\mathbf{G}_{\mathrm{add}}(n) := \exp(n\,\omega\,\mathbf{A}) = \mathbf{I}_{d+1} + n\,\omega\,\mathbf{A} = \begin{bmatrix} \mathbf{I}_d & n\,\omega\,\mathbf{u} \\ \mathbf{0}^\top & 1 \end{bmatrix} \in \mathrm{GL}(d{+}1),$$

$$\mathbf{G}_{\mathrm{add}}(n{+}m) = \mathbf{G}_{\mathrm{add}}(n)\mathbf{G}_{\mathrm{add}}(m).$$

Application and exact relative law in GL**.** For queries/keys augmented as $\widehat{\mathbf{q}}_i = [\mathbf{q}_i; 1]$ and $\widehat{\mathbf{k}}_j = [\mathbf{k}_j; 1]$, define

$$\widetilde{\mathbf{q}}_i := \mathbf{G}_{\mathrm{add}}(i)\,\widehat{\mathbf{q}}_i, \qquad \widetilde{\mathbf{k}}_j := \mathbf{G}_{\mathrm{add}}(j)^{-\top}\,\widehat{\mathbf{k}}_j, \tag{5.2}$$

and score with the standard inner product on $\mathbb{R}^{d+1}$. The key is transformed using the inverse transpose ($\mathbf{G}_{\mathrm{add}}(j)^{-\top}$). This is necessary because for a general linear group GL, the simple transpose is no longer the inverse (unlike in $\mathrm{SO}(d)$), and the inverse transpose is required to recover the exact relative law: $\mathbf{G}_{\mathrm{add}}(i)^{\top}\mathbf{G}_{\mathrm{add}}(j)^{-\top} = \mathbf{G}_{\mathrm{add}}(j{-}i)^{-\top}$ for any one-parameter subgroup in GL. This composition results in the final form:

$$\widetilde{\mathbf{q}}_i^{\top}\widetilde{\mathbf{k}}_j = \widehat{\mathbf{q}}_i^{\top}\mathbf{G}_{\mathrm{add}}(j{-}i)^{-\top}\widehat{\mathbf{k}}_j, \quad \text{depending only on } j{-}i. \tag{5.3}$$

Streaming matches Section 2.5: cache $\widehat{\mathbf{k}}_j^{\star} = \mathbf{G}_{\mathrm{add}}(j)^{-\top}\widehat{\mathbf{k}}_j$ once; at step t form $\widetilde{\mathbf{q}}_t = \mathbf{G}_{\mathrm{add}}(t)\widehat{\mathbf{q}}_t$ and compute $\widetilde{\mathbf{q}}_t^{\top}\widehat{\mathbf{k}}_j^{\star}$.

Closed form and content-gated additive term. Since $\mathbf{A}^{\top} = \left(\begin{smallmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{u}^{\top} & 0 \end{smallmatrix}\right)$ and $(\mathbf{A}^{\top})^2 = \mathbf{0}$,

$$\mathbf{G}_{\mathrm{add}}(m)^{-\top} = \mathbf{I}_{d+1} - m\,\omega\,\mathbf{A}^{\top} = \begin{bmatrix} \mathbf{I}_d & \mathbf{0} \\ -\,m\,\omega\,\mathbf{u}^{\top} & 1 \end{bmatrix}, \qquad m = j{-}i, \tag{5.4}$$

whence

$$\widetilde{\mathbf{q}}_i^{\top}\widetilde{\mathbf{k}}_j = \mathbf{q}_i^{\top}\mathbf{k}_j \;+\; 1 \;-\; (j{-}i)\,\omega\,\mathbf{u}^{\top}\mathbf{k}_j. \tag{5.5}$$

The constant "+1" is softmax-shift invariant; the final term is an additive, linear-in-offset bias whose slope is key-gated by $\mathbf{u}^{\top}\mathbf{k}_j$. A symmetric generator for the query, $\mathbf{A}_{\mathrm{qry}} = \left(\begin{smallmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{v}^{\top} & 0 \end{smallmatrix}\right)$ applied analogously produces a query-gated slope $(j{-}i)\,\omega\,\mathbf{v}^{\top}\mathbf{q}_i$. Using both the key-gated and query-gated components yields a combined bias of the form $(j{-}i)\,\omega\,(\mathbf{v}^{\top}\mathbf{q}_i - \mathbf{u}^{\top}\mathbf{k}_j)$, still obeying the exact relative law Eq. (5.3).

5.2 Exact ALiBi as a Rank-1 unipotent in GL($d{+}2$)

ALiBi adds a head-specific scalar slope $\beta_h(j{-}i)$ to the logits that is independent of content. This is captured exactly by augmenting with two constant coordinates:

$$\widehat{\mathbf{q}}_i = [\mathbf{q}_i;\ 1;\ 0] \in \mathbb{R}^{d+2}, \qquad \widehat{\mathbf{k}}_j = [\mathbf{k}_j;\ 0;\ 1] \in \mathbb{R}^{d+2},$$

and choosing the rank-1 nilpotent generator

$$\mathbf{A}_h^{\top} \;=\; \beta_h\,\mathbf{e}_{d+1}\,\mathbf{e}_{d+2}^{\top} \quad\Longleftrightarrow\quad \mathbf{A}_h \;=\; \beta_h\,\mathbf{e}_{d+2}\,\mathbf{e}_{d+1}^{\top}, \qquad (\mathbf{A}_h^{\top})^2 = \mathbf{0}. \tag{5.6}$$

Then $\mathbf{G}_{\mathrm{add},h}(m)^{-\top} = \mathbf{I} - m\,\mathbf{A}_h^{\top}$ and

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

i.e., the ALiBi term emerges as a unipotent GL($d{+}2$) action with exact relative composition.

FoX as Add-GRAPE. Let $f_t \in (0, 1]$ be per-token forget scalars and set $\omega_t := \log f_t$. Using the rank-1 generator of Section 5.2, the resulting additive bias is $b(t, j) = \sum_{\ell=j+1}^{t} \omega_\ell$, which coincides with FoX's forgetting bias D_{ij}. A full derivation and the unipotent path product are given in Appendix A.

6 Path-Integral Additive GRAPE

Additive GRAPE (Add-GRAPE) (Section 5) realizes exactly relative additive logits via a one-parameter unipotent action in the general linear group GL; the bias depends only on an offset $m = j{-}i$ (or a contextual phase difference $\Phi_j{-}\Phi_i$ when using cumulative phases). In practice, we sometimes want the amount of additive encouragement/suppression between a key at j and a

query at t to depend on the endpoint t (e.g., the current syntactic or semantic needs of the query token), while preserving causality, boundedness, and clean composition with the orthogonal GRAPE acting on $(\mathbf{q}, \mathbf{k})$. We formalize this by a rigorously defined path-integral sum, deriving conditions under which the exact relative law of Additive GRAPE is recovered.

Definition (Path-integral bias). Fix a head h and per-head scale $\alpha_h > 0$. For each time u, let $\mathbf{p}_{u,h} \in \mathbb{R}^d$ be a "positional probe" obtained from token-local features (a linear projection followed by RMS normalization in our implementation). Let $\mathcal{J}$ be the canonical block-diagonal 90° operator (Section 2.4), and define $\mathbf{R}_\ell := \exp(\ell\,\mathcal{J})$ (a fixed commuting rotation). For a link function $g : \mathbb{R} \to (-\infty, 0)$ that is monotone increasing and 1-Lipschitz[1], define the *edge potential*

$$\psi_h(t,\ell) := \alpha_h\, g\!\left(\frac{1}{d}\,\langle \mathbf{p}_{t,h},\, \mathbf{R}_\ell\, \mathbf{p}_{\ell,h}\rangle\right) \;\le\; 0, \qquad \ell < t. \tag{6.1}$$

The *path-integral additive bias* from key position j to query position t is the causal sum

$$b_h(t,j) \;:=\; \sum_{\ell=j+1}^{t} \psi_h(t,\ell) \;\le\; 0. \tag{6.2}$$

The attention logit combines this additive term with either the raw or orthogonally-rotary bilinear part:

$$\ell_{t,j,h} \;=\; \frac{1}{\sqrt{d}}\,\mathbf{q}_{t,h}^\top \mathbf{k}_{j,h} \;+\; b_h(t,j) \quad \text{or} \quad \ell_{t,j,h} \;=\; \frac{1}{\sqrt{d}}\,\mathbf{q}_{t,h}^\top \mathbf{G}_h(j-t)\mathbf{k}_{j,h} \;+\; b_h(t,j). \tag{6.3}$$

Group-theoretic formalization and path composition. Let $\mathbf{E} \in \mathbb{R}^{(d+2)\times(d+2)}$ be a fixed rank-1 nilpotent with $\mathbf{E}^2 = \mathbf{0}$ (e.g., $\mathbf{E} = \mathbf{e}_{d+2}\mathbf{e}_{d+1}^\top$ as in Section 5.2). For each fixed endpoint t, define endpoint-indexed unipotent factors

$$\mathbf{H}_h^{(t)}(\ell) \;:=\; \mathbf{I} + \psi_h(t,\ell)\,\mathbf{E}.$$

Since $\mathbf{E}^2 = 0$, the path product along $(j,t]$ collapses additively:

$$\prod_{\ell=j+1}^{t} \mathbf{H}_h^{(t)}(\ell) \;=\; \mathbf{I} + \Big(\sum_{\ell=j+1}^{t} \psi_h(t,\ell)\Big)\,\mathbf{E} \;=\; \mathbf{I} + b_h(t,j)\,\mathbf{E}. \tag{6.4}$$

Scoring in homogeneous coordinates as in Section 5 with the paired inverse-transpose removes multiplicative anisotropy and yields exactly the additive term $b_h(t,j)$, cf. Eq. (5.3). The *rowwise* semigroup law is preserved (Eq. (6.4)), while the t-dependence of the factors intentionally relaxes the global one-parameter group law.

Relation to Add-GRAPE. PI-Add-GRAPE strictly contains Add-GRAPE as the special case in which edge potentials do not depend on the endpoint:

$$\psi_h(t,\ell) \equiv \theta_h\, a_\ell \;\Longrightarrow\; b_h(t,j) = \theta_h \sum_{\ell=j+1}^{t} a_\ell = \theta_h\big(A_t - A_j\big), \quad A_u := \sum_{\ell<u} a_\ell.$$

Two important instances follow directly:

- **Exact ALiBi.** $a_\ell \equiv 1$ gives $b_h(t,j) = \theta_h(t-j)$; this is exactly the ALiBi term recovered via the rank-1 unipotent lift in Section 5.2.
- **Phase-modulated Additive GRAPE.** If $a_\ell = \omega_\ell$ with $\omega_\ell = g(x_\ell) \ge 0$, then $b_h(t,j) = \theta_h(\Phi_t - \Phi_j)$ with $\Phi_u = \sum_{\ell<u} \omega_\ell$.

In both cases, $b_h(t,j)$ depends only on a (possibly contextual) phase difference and thus obeys the exact relative law with the same streaming/cache policy as Section 5. Outside these endpoint-independent regimes, PI-Add-GRAPE provides strictly more expressive, path-integral biases while preserving row-wise path composition (Eq. (6.4)).

[1] Our experiments take $g(z) = \log(\text{Sigmoid}(z))$; then $g'(z) = 1 - \text{Sigmoid}(z) \in (0,1)$, ensuring 1-Lipschitzness.

Computation and streaming. For each head h and decoding step t, compute the row $\{\psi_h(t,\ell)\}_{\ell\le t}$ by a single similarity sweep $\ell \mapsto \langle \mathbf{p}_{t,h}, \mathbf{R}_\ell \mathbf{p}_{\ell,h}\rangle$ (the rotated probes $\mathbf{R}_\ell \mathbf{p}_{\ell,h}$ can be cached on arrival), apply the link g, and take a prefix sum to obtain $j \mapsto b_h(t,j)$. This yields $O(t)$ per-step overhead with $O(1)$ recomputation per cached key; memory is $O(L)$ per head for the cached probes (or $O(d)$ if the per-ℓ rotations are recomputed on the fly).

Spectral and stability. Each factor $\mathbf{H}_h^{(t)}(\ell) = \mathbf{I} + \psi_h(t,\ell)\mathbf{E}$ is unipotent with all eigenvalues 1 and at most two singular values deviating from 1; the full path product equals $\mathbf{I} + b_h(t,j)\mathbf{E}$ (Eq. (6.4)). As in Appendix E.3, the paired inverse-transpose used for scoring cancels multiplicative distortions and delivers exactly the additive bias $b_h(t,j)$; operator norms remain controlled linearly in $|b_h(t,j)|$.

A more extensive spectral analysis, including eigenvalue structure and singular-value behavior across GRAPE variants, is provided in Appendix E. There, we also give an explicit comparison to PaTH Attention (Yang et al., 2025), which is shown to be contractive and near singular. These properties may impair PaTH's effectiveness in long-context modeling.

7 Experiments

In this section, we will evaluate the performance of GRAPE on the language modeling task in comparison with baseline positional encoding mechanisms, including RoPE (Su et al., 2021), ALiBi (Press et al., 2021) as well as Forgetting Transformer (FoX) (Lin et al., 2025).

7.1 Implementation Details

Based on nanoGPT code framework (Karpathy, 2022), our experiment are implemented on Llama model (Touvron et al., 2023a). We only change the positional encoding mechanism and keep the rest of the model architecture same with Llama. We choose FineWeb-Edu 100B dataset (Lozhkov et al., 2024), which contains 100 billion training tokens and 0.1 billion validation tokens, and we randomly choose 50B tokens for training. Our models are with 36 layers and 10 heads, with a hidden size of 1280 and head dimension of 128. The context length is set to 4,096 and the batch size is 480. All the models are optimized by AdamW optimizer (Loshchilov & Hutter, 2019), with a maximum learning rate of 2×10^{-4}, $(\beta_1, \beta_2) = 0.9, 0.95$, and a weight decay of 0.1. We use a cosine learning rate scheduler with 2,000 warm-up iterations and the minimum learning is 1×10^{-5}. We also clip the gradient to 1.0 for stabler training. The frequency of RoPE is set to 10,000. Moreover, for fair comparison, we do not use FoX-Pro and disabled the KV-shift module within it.

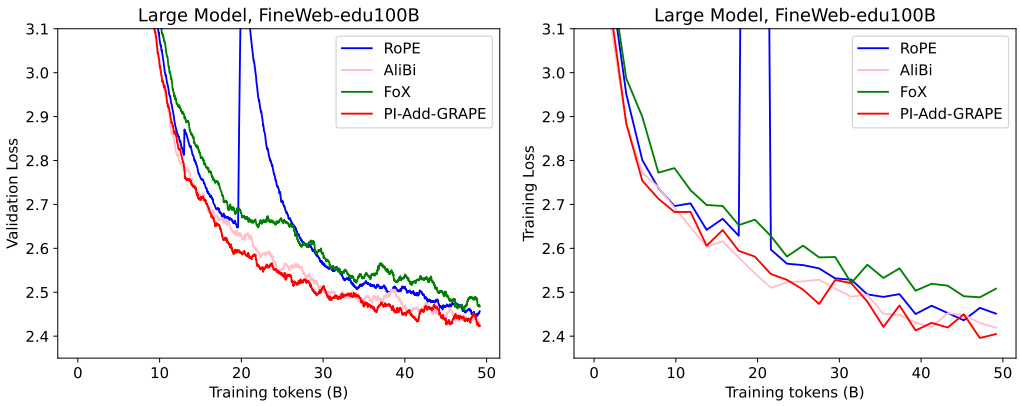
7.2 Result Analysis

The curves for training and validation loss of models with variant positional encoding mechanism are displayed in Figure 1. It can be observed that GRAPE can keep a persistent edge over other mechanisms, including RoPE and FoX. Moreover, model with RoPE suffers from a great spike while the model with GRAPE embedding steadily improves during the training process.

8 Related Work

Positional information in Transformers mainly can be categorized into these classes: (a) absolute encodings (sinusoidal or learned) (Vaswani et al., 2017; Devlin et al., 2019; Neishi & Yoshinaga, 2019; Kiyono et al., 2021; Likhomanenko et al., 2021; Wang et al., 2020; Liu et al., 2020; Wang et al., 2021; Sinha et al., 2022; Wennberg & Henter, 2021; Ke et al., 2020); (b) relative encodings that depend on offsets (Shaw et al., 2018; Dai et al., 2019; Raffel et al., 2020; He et al., 2020); and (c) linear logit biases with strong length extrapolation (Press et al., 2021; Chi et al., 2022a;b; Li et al., 2023; Ruoss et al., 2023), all shaping recency/extrapolation behavior (Haviv et al., 2022; Kazemnejad et al., 2023).

Multiplicative position encoding. RoPE realizes offsets as block-diagonal planar rotations of queries/keys, preserving norms and exact origin invariance; it is widely deployed across LLMs and modalities (Su et al., 2021; Touvron et al., 2023a;b; Heo et al., 2024). Angle/spectrum designs improve long-context fidelity (e.g., xPos) (Sun et al., 2022); LRPE formalizes separable relative



(a) Training Loss (b) Validation Loss

Figure 1: The training and validation loss of large-size models (770M), with different positional encoding mechanisms on the FineWeb-Edu 100B dataset.

transforms for linear attention models (Qin et al., 2023); mechanistic work analyzes frequency usage (Barbero et al., 2025). These methods are also compatible with sparse/linear attentions (Beltagy et al., 2020; Zaheer et al., 2020; Katharopoulos et al., 2020; Choromanski et al., 2020) and with context-scaling procedures (Xiong et al., 2023; Chen et al., 2023; Peng et al., 2023; Zhu et al., 2023; Jin et al., 2024). **Mul-GRAPE** identifies RoPE as commuting rank-2 exponentials in $\mathrm{SO}(d)$ and extends it to learned subspaces and compact non-commuting mixtures in closed form.

Additive position encoding and forgetting mechanisms. Additive schemes such as ALiBi (Press et al., 2021) and related kernelized/randomized forms (Chi et al., 2022a;b; Li et al., 2023; Ruoss et al., 2023) are captured exactly by Add-GRAPE as unipotent actions inthe general linear group GL that preserve the same relative law and streaming cacheability. Importantly, *forgetting mechanisms are additive*: the Forgetting Transformer (FoX) implements a learnable per-head exponential decay in the attention logits and is a specific Add-GRAPE / PI-Add-GRAPE instance imposing distance-dependent attenuation (Lin et al., 2025). FoX's data-dependent forget gates yield a path-additive bias $\mathbf{D}$ that we show is exactly the endpoint-independent PI-Add-GRAPE case; see Appendix A for a constructive equivalence and its streaming implementation (Lin et al., 2025).

Contextual position encoding. Content-adaptive position modulates effective phase or distance via token features through gating/scaling and algebraic parameterizations (Wu et al., 2020; Zheng et al., 2024; Kogkalidis et al., 2024), and contextual counting (CoPE) (Golovneva et al., 2024). GRAPE introduces phase-modulated and dictionary-based contextual variants that replace a linear phase with cumulative token-adaptive phases (single or multi-subspace) while retaining exact headwise relativity and streaming caches. Finally, models can length-generalize without explicit encodings ("NoPE") under suitable training (Wang et al., 2024), which corresponds to the trivial generator $\mathbf{L} = 0$ in our view.

9 Conclusion

GRAPE provides a general framework for positional encoding based on group actions, unifying *multiplicative* and *additive* mechanisms. Multiplicative GRAPE offers a closed-form, rank-2 exponential that is relative, compositional, and norm-preserving; it recovers RoPE and yields learned-basis and non-commuting extensions at controlled cost. Additive GRAPE realizes ALiBi and FoX exactly via unipotent general linear group GL lifts with the same streaming/cache policy. The GRAPE framework integrates seamlessly with existing Transformer models and offers a principled, extensible design space for future architectures.

REFERENCES

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličkovic. Round and round we go! what makes rotary positional encodings useful? In *International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://arxiv.org/abs/2410.06205. Also arXiv:2410.06205.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via position interpolation. *arXiv preprint arXiv:2306.15595*, 2023. URL https://arxiv.org/abs/2306.15595.

Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. *Advances in Neural Information Processing Systems*, 35:8386–8399, 2022a.

Ta-Chung Chi, Ting-Han Fan, Alexander I Rudnicky, and Peter J Ramadge. Dissecting transformer length extrapolation via the lens of receptive field analysis. *arXiv preprint arXiv:2212.10356*, 2022b.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Zihang Dai, Zhilin Yang, Yiming Yang, William Cohen, Ruslan Salakhutdinov, and Jaime Carbonell. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of ACL*, pp. 2978–2988, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Olga Golovneva, Jackie C Lou, Daniel Holtmann-Rice, Aditya Kusupati, Chengeng Cai, Zijian Hu, Prateek Vijay Kumar, Tim Dettmers, Pratyusha Sharma, Behnam Neyshabur, Jason D. Lee, and Mohammad Bavarian. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*, 2024. URL https://arxiv.org/abs/2405.18719.

Brian C Hall. Lie groups, lie algebras, and representations. In *Quantum Theory for Mathematicians*, pp. 333–366. Springer, 2013.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. *arXiv preprint arXiv:2203.16634*, 2022.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235, pp. 22099–22114. PMLR, 2024.

Andrej Karpathy. NanoGPT. https://github.com/karpathy/nanoGPT, 2022.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*, 2020.

Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui. SHAPE: Shifted absolute position embedding for transformers. In *Proceedings of EMNLP*, pp. 3309–3321, 2021.

Konstantinos Kogkalidis, Jean-Philippe Bernardy, and Vikas Garg. Algebraic positional encodings. *Advances in Neural Information Processing Systems*, 37:34824–34845, 2024.

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. *arXiv preprint arXiv:2310.04418*, 2023.

Tatiana Likhomanenko, Qiantong Xu, Gabriel Synnaeve, Ronan Collobert, and Alex Rogozhnikov. CAPE: Encoding relative positions with continuous augmented positional embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 16079–16092, 2021.

Mengtian Lin, Ji Lin, Wei-Ming Chen, and Yonglong Tian. Forgetting transformer: Softmax attention with a forget gate. *arXiv preprint arXiv:2503.02130*, 2025. URL https://arxiv.org/abs/2503.02130.

Xuanqing Liu, Hsiang-Fu Yu, Inderjit Dhillon, and Cho-Jui Hsieh. Learning to encode position for transformer with continuous dynamical model. In *International conference on machine learning*, pp. 6327–6335. PMLR, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

Masato Neishi and Naoki Yoshinaga. On the relation between position information and sentence length in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 328–338, 2019.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Zhen Qin, Weixuan Sun, Kaiyue Lu, Hui Deng, Dongxu Li, Xiaodong Han, Yuchao Dai, Lingpeng Kong, and Yiran Zhong. Linearized relative positional encoding. *arXiv preprint arXiv:2307.09270*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Olinde Rodrigues. Des lois géométriques qui régissent les déplacemens d'un système solide dans l'espace, et de la variation des coordonnées provenant de ces déplacemens considérés indépendamment des causes qui peuvent les produire. *Journal de Mathématiques Pures et Appliquées*, 5:380–440, 1840.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. *arXiv preprint arXiv:2305.16843*, 2023.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Koustuv Sinha, Amirhossein Kazemnejad, Siva Reddy, Joelle Pineau, Dieuwke Hupkes, and Adina Williams. The curious case of absolute position embeddings. *arXiv preprint arXiv:2210.12574*, 2022.

Jianlin Su, Yuancheng Zhang, Shengfeng Pan, Shengyu Ge, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Benyou Wang, Donghao Zhao, Christina Lioma, Qiuchi Li, Peng Zhang, and Jakob Grue Simonsen. Encoding word order in complex embeddings. In *Proceedings of ICLR*, 2020.

Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen. On position embeddings in BERT. In *Proceedings of ICLR*, 2021.

Jie Wang, Tao Ji, Yuanbin Wu, Hang Yan, Tao Gui, Qi Zhang, Xuanjing Huang, and Xiaoling Wang. Length generalization of causal transformers without position encoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14024–14040, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.834. URL https://aclanthology.org/2024.findings-acl.834/.

Ulme Wennberg and Gustav Eje Henter. The case for translation-invariant self-attention in transformer-based language models. *arXiv preprint arXiv:2106.01950*, 2021.

Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. Da-transformer: Distance-aware transformer. *arXiv preprint arXiv:2010.06925*, 2020.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*, 2023.

Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, and Yoon Kim. Path attention: Position encoding via accumulating householder transformations. *arXiv preprint arXiv:2505.16381*, 2025.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, et al. Dape: Data-adaptive positional encoding for length extrapolation. *Advances in Neural Information Processing Systems*, 37:26659–26700, 2024.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Pose: Efficient context window extension of llms via positional skip-wise training. *arXiv preprint arXiv:2309.10400*, 2023.

Appendix

A Forgetting Transformer as a Special Additive GRAPE **14**

B Composition of Additive GRAPE and Multiplicative GRAPE **15**

C Algorithmic Details and Pseudo Code **15**

D Differentiation and Fast Application of Rank-2 Matrix Exponential **16**

E Spectral Analysis of GRAPE and Other Methods **17**

E.1 Rank-2 Plane: Exact Spectrum and Geometric Interpretation 17

E.2 Commuting Multi-subspace Mul-GRAPE and RoPE 17

E.3 Add-GRAPE and Path-Integral Additive (PI-Add-GRAPE) 18

E.4 Comparison to PaTH Attention 19

A FORGETTING TRANSFORMER AS A SPECIAL ADDITIVE GRAPE

The Forgetting Transformer (FoX) introduces a scalar forget gate $f_t \in (0, 1]$ per head and timestep and adds the cumulative log-gate as an additive bias in the attention logits. Concretely, for a head h,

$$f_{t,h} = \sigma(\mathbf{w}_{f,h}^\top \mathbf{x}_t + b_{f,h}), \qquad F_{ij,h} = \prod_{\ell=j+1}^{i} f_{\ell,h}, \qquad D_{ij,h} = \log F_{ij,h} = \sum_{\ell=j+1}^{i} \log f_{\ell,h},$$

and the attention is

$$\mathbf{O}_h = \mathrm{softmax}\Big(\tfrac{1}{\sqrt{d}}\mathbf{Q}\mathbf{K}^\top + \mathbf{D}_h\Big)\mathbf{V}. \tag{FoX}$$

We now show that Eq. (FoX) is exactly realized by our Add-GRAPE framework using the endpoint-independent path-additive specialization of Section 6.

FoX as PI-Add-GRAPE with endpoint-independent edges. In PI-Add-GRAPE (Section 6), a head-wise additive logit $b_h(t, j)$ arises as a causal path sum

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

If the edge potentials do not depend on the endpoint, i.e. $\psi_h(t, \ell) \equiv a_{\ell,h}$, then $b_h(t, j)$ reduces to a difference of per-time potentials:

$$b_h(t, j) \;=\; \sum_{\ell=j+1}^{t} a_{\ell,h} \;=\; U_{t,h} - U_{j,h}, \qquad U_{u,h} := \sum_{\ell<u} a_{\ell,h}.$$

FoX corresponds to the choice $a_{\ell,h} = \log f_{\ell,h} \le 0$, yielding

$$b_h(t, j) \equiv D_{ij,h} = \sum_{\ell=j+1}^{t} \log f_{\ell,h}.$$

Thus the FoX forgetting bias $\mathbf{D}_h$ is precisely the PI-Add-GRAPE path-integral additive bias with endpoint-independent edges.

Unipotent GL lift (Add-GRAPE view). Let $E := \mathbf{e}_{d+2}\mathbf{e}_{d+1}^\top$ be the rank-1 nilpotent used in Section 5.2. For a fixed head h and endpoint t, define per-link unipotent factors

$$\mathbf{H}_h^{(t)}(\ell) \;=\; \mathbf{I} + \psi_h(t, \ell)\, E, \qquad \psi_h(t, \ell) = \log f_{\ell,h}.$$

Since $E^2 = 0$, the path product collapses:

$$\prod_{\ell=j+1}^{t} \mathbf{H}_h^{(t)}(\ell) \;=\; \mathbf{I} + \Big(\sum_{\ell=j+1}^{t} \log f_{\ell,h}\Big) E \;=\; \mathbf{I} + D_{ij,h}\, E.$$

Scoring in homogeneous coordinates as in Section 5 with the paired inverse-transpose,

$$\widetilde{\mathbf{q}}_{t,h}^\top \widetilde{\mathbf{k}}_{j,h} \;=\; \widehat{\mathbf{q}}_{t,h}^\top \Big(\mathbf{I} + D_{ij,h}\, E\Big)^{-\top} \widehat{\mathbf{k}}_{j,h} \;=\; \mathbf{q}_{t,h}^\top \mathbf{k}_{j,h} \;+\; D_{ij,h},$$

recovers Eq. (FoX) exactly (up to the standard $1/\sqrt{d}$ factor we include throughout). Hence *FoX is an exact Add-GRAPE / PI-Add-GRAPE instance* realized by a rank-1 unipotent path with endpoint-independent edges.

Streaming and complexity. Compute prefix sums $U_{t,h} = \sum_{\ell<t} \log f_{\ell,h}$ once per step; then $D_{ij,h} = U_{i,h} - U_{j,h}$ is obtained by subtraction, preserving the $O(L)$ rowwise cost and the streaming cache policy from Section 5–Section 6. The headwise gates $f_{t,h}$ add $O(1)$ parameters and negligible computation.

Special cases and composition. If $f_{t,h} \equiv e^{-\beta_h}$ (constant per head), then $D_{ij,h} = -\beta_h(i-j)$ and FoX reduces to exact ALiBi (Section 5.2). More generally, FoX composes additively with the multiplicative (orthogonal) GRAPE acting on $(\mathbf{q}, \mathbf{k})$ as in Eq. (6.3), preserving norm-preservation of the rotational part while adding bounded, non-positive, content-adaptive path biases.

B Composition of Additive GRAPE and Multiplicative GRAPE

For the unipotent forms of Additive GRAPE, applying $\mathbf{G}_{\mathrm{add}}(m)^{-\top}$ requires one inner product and one axpy per active component (no trigonometry; gradients are polynomial in m). Thus, the per-head overhead is $O(d)$ and typically negligible relative to attention matmuls. Multiplicative GRAPE (Section 3) and GRAPE-Additive (GRAPE-Add) compose naturally, either additively at the logit level

$$\ell_{t,j,h} = \tfrac{1}{\sqrt{d}}\, \mathbf{q}_{t,h}^{\top} \mathbf{G}_h(j-t)\mathbf{k}_{j,h} \;+\; \Big[\widehat{\mathbf{q}}_{t,h}^{\top} \mathbf{G}_{\mathrm{add},h}(j-t)^{-\top} \widehat{\mathbf{k}}_{j,h} - \mathbf{q}_{t,h}^{\top}\mathbf{k}_{j,h}\Big],$$

or as a single block-upper-triangular GL action in homogeneous coordinates (semidirect product/semidirect *sum* view). Concretely, define the joint lift

$$\widehat{\mathbf{q}} = [\mathbf{q}; 1], \quad \widehat{\mathbf{k}} = [\mathbf{k}; 1], \qquad \widehat{\mathbf{G}}(m) \;=\; \begin{bmatrix} \exp(m\,\mathbf{L}) & m\,\omega\,\mathbf{u} \\ \mathbf{0}^{\top} & 1 \end{bmatrix} \in \mathrm{GL}(d{+}1),$$

which combines the orthogonal rotation $\exp(m\mathbf{L})$ on features with a unipotent "translation" along the homogeneous axis. Scoring with the paired inverse-transpose as in (5.2) yields

$$\widehat{\mathbf{q}}^{\top}\, \widehat{\mathbf{G}}(m)^{-\top} \widehat{\mathbf{k}}{+} \;=\; \mathbf{q}^{\top} \exp(m\mathbf{L})\mathbf{k} \;-\; m\,\omega\,\mathbf{u}^{\top}\mathbf{k} \;+\; \text{const},$$

exactly reproducing the sum of multiplicative (rotary) and additive (bias) components up to a softmax-invariant constant. In both formulations, exact relativity and streaming caches are retained.

C Algorithmic Details and Pseudo Code

This appendix contains the detailed pseudocode.

Algorithm 1 Commuting Multi-Subspace Mul-GRAPE

Require: $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{B\times L\times H\times d}$, orthogonal $\mathbf{E} \in \mathbb{R}^{d\times d}$, frequencies $\{\omega_{h,j}\}_{j=1}^{d/2}$, positions $n \in \mathbb{Z}^L$

for $h = 1$ **to** H **do**
  $\mathbf{Q}'[:,:,h,:] \leftarrow \mathbf{Q}[:,:,h,:]\,\mathbf{E}$; $\mathbf{K}'[:,:,h,:] \leftarrow \mathbf{K}[:,:,h,:]\,\mathbf{E}$
  for $\ell = 0$ **to** $L-1$ **do**
    for $j = 1$ **to** $d/2$ **do**
      $\theta \leftarrow n_\ell\, \omega_{h,j}$; apply 2×2 rotation $\mathbf{G}_2(\theta)$ to coords $(2j{-}1, 2j)$ of $\mathbf{Q}'[:,\ell,h,:]$ and $\mathbf{K}'[:,\ell,h,:]$
    end for
  end for
  $\widetilde{\mathbf{Q}}[:,:,h,:] \leftarrow \mathbf{Q}'\,\mathbf{E}^{\top}$; $\widetilde{\mathbf{K}}[:,:,h,:] \leftarrow \mathbf{K}'\,\mathbf{E}^{\top}$
end for
return $(\widetilde{\mathbf{Q}}, \widetilde{\mathbf{K}})$

Algorithm 2 Fast On-contextual Non-commuting MS-Mul-GRAPE via Schur-Mode Rotation

Require: $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{B\times L\times H\times d}$; planes $\{(\mathbf{a}_{h,j}, \mathbf{b}_{h,j}, \omega_{h,j})\}_{j=1}^{m}$; positions n

for $h = 1$ **to** H **do** ▷ one-time per head
  Build $\mathbf{U}_h = \mathrm{span}\{\mathbf{a}_{h,j}, \mathbf{b}_{h,j}\}$; orthonormalize $\mathbf{b}_h \in \mathbb{R}^{d\times r_h}$
  $\mathbf{L}_{\mathbf{U},h} \leftarrow \mathbf{b}_h^\top \Big(\sum_{j=1}^{m} \omega_{h,j} \mathbf{L}(\mathbf{a}_{h,j}, \mathbf{b}_{h,j}) \Big) \mathbf{b}_h \in \mathfrak{so}(r_h)$
  Orthogonally Schur-decompose: $\mathbf{L}_{\mathbf{U},h} = \mathbf{T}_h \left(\bigoplus_{t=1}^{r_h/2} \theta_{h,t} \mathbf{J} \right) \mathbf{T}_h^\top$
  $\mathbf{E}_h \leftarrow \mathbf{b}_h \mathbf{T}_h \in \mathbb{R}^{d\times r_h}$; precompute $(c_{h,t}, s_{h,t}) = (\cos\theta_{h,t}, \sin\theta_{h,t})$
end for
for $\ell = 0$ **to** $L-1$ **do** ▷ token loop
  for $h = 1$ **to** H **do**
    $y_{\mathbf{Q}} \leftarrow \mathbf{E}_h^\top \mathbf{Q}[:,\ell,h,:]$; $y_{\mathbf{K}} \leftarrow \mathbf{E}_h^\top \mathbf{K}[:,\ell,h,:]$
    for $t = 1$ **to** $r_h/2$ **do**
      $(\mathbf{C}_{h,t}, \mathbf{S}_{h,t}) \leftarrow \textsc{PhaseTo}(n_\ell;\, c_{h,t}, s_{h,t})$ ▷ $(\mathbf{C}, \mathbf{S})$ from $(\cos\theta, \sin\theta)$ via angle-addition or binary exponentiation
      Apply $\left(\begin{smallmatrix} \mathbf{C}_{h,t} & -\mathbf{S}_{h,t} \\ \mathbf{S}_{h,t} & \mathbf{C}_{h,t} \end{smallmatrix}\right)$ to coordinates $(2t-1, 2t)$ of y_Q, y_K
    end for
    $\widetilde{\mathbf{Q}}[:,\ell,h,:] \leftarrow \mathbf{Q}[:,\ell,h,:] + \mathbf{E}_h(y_Q - \mathbf{E}_h^\top \mathbf{Q}[:,\ell,h,:])$
    $\widetilde{\mathbf{K}}[:,\ell,h,:] \leftarrow \mathbf{K}[:,\ell,h,:] + \mathbf{E}_h(y_K - \mathbf{E}_h^\top \mathbf{K}[:,\ell,h,:])$
  end for
end for
return $(\widetilde{\mathbf{Q}}, \widetilde{\mathbf{K}})$

Algorithm 3 GRAPE-Additive (GRAPE-Add) with streaming cache

Require: $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{B\times L\times H\times d}$; per-head additive generators $\{\mathbf{A}_h\}$ with $\mathbf{A}_h^2 = \mathbf{0}$; positions $n \in \mathbb{Z}^L$

Augment: $\widehat{\mathbf{Q}} \leftarrow [\mathbf{Q}; \mathbf{1}; \mathbf{0}]$, $\widehat{\mathbf{K}} \leftarrow [\mathbf{K}; \mathbf{0}; \mathbf{1}]$ as needed (Section 5.2)
for $j = 0$ **to** $L-1$ **do** ▷ cache once on arrival
  for $h = 1$ **to** H **do**
    $\widehat{\mathbf{K}}^\star[:,j,h,:] \leftarrow \left(\mathbf{I} - n_j \mathbf{A}_h^\top\right) \widehat{\mathbf{K}}[:,j,h,:]$
  end for
end for
for $t = 0$ **to** $L-1$ **do**
  for $h = 1$ **to** H **do**
    $\widetilde{\mathbf{Q}}[:,t,h,:] \leftarrow \left(\mathbf{I} + n_t \mathbf{A}_h\right) \widehat{\mathbf{Q}}[:,t,h,:]$
    Compute additive logits: $\lambda_{t,j,h} \leftarrow \widetilde{\mathbf{Q}}[:,t,h,:]^\top \widehat{\mathbf{K}}^\star[:,j,h,:]$
  end for
end for
return $\{\lambda_{t,j,h}\}$ (to be added to orthogonal GRAPE/RoPE logits)

D Differentiation and Fast Application of Rank-2 Matrix Exponential

Differentiation and stability. Let $f_1(z) = \frac{\sin z}{z}$ and $f_2(z) = \frac{1-\cos z}{z^2}$ with $z = n\omega s$. Then

$$\exp(n\omega\mathbf{L}) = \mathbf{I} + f_1(z)\mathbf{L} + f_2(z)\mathbf{L}^2.$$

For any scalar parameter $\theta \in \{\omega\} \cup \{\text{entries of } a, b\}$,

$$\partial_\theta \exp(n\omega\mathbf{L}) = f_1(z)\,\partial_\theta\mathbf{L} + f_2(z)\,(\mathbf{L}\,\partial_\theta\mathbf{L} + \partial_\theta\mathbf{L}\,\mathbf{L}) + \partial_\theta z\,\left(f_1'(z)\mathbf{L} + f_2'(z)\mathbf{L}^2\right),$$
$$\partial_\theta z = n\omega\,\partial_\theta s + ns\,\partial_\theta\omega, \qquad \partial_\theta s = \tfrac{1}{2}s^{-1}\partial_\theta(\alpha\beta - \gamma^2).$$

Use series for $|z| < \varepsilon$: $f_1(z) = 1 - \frac{z^2}{6} + O(z^4)$ and $f_2(z) = \frac{1}{2} - \frac{z^2}{24} + O(z^4)$. These formulas enable mixed-precision backprop with small-s guards.

Fast application. For any $x \in \mathbb{R}^d$,

$$\mathbf{L}\mathbf{x} = \mathbf{a}\langle \mathbf{b}, \mathbf{x}\rangle - \mathbf{b}\langle \mathbf{a}, \mathbf{x}\rangle, \qquad \mathbf{L}^2\mathbf{x} = \gamma(\mathbf{a}\langle \mathbf{b}, \mathbf{x}\rangle + \mathbf{b}\langle \mathbf{a}, \mathbf{x}\rangle) - \beta\, \mathbf{a}\langle \mathbf{a}, \mathbf{x}\rangle - \alpha\, \mathbf{b}\langle \mathbf{b}, \mathbf{x}\rangle.$$

Thus $\mathbf{G}(n)\mathbf{x} = \mathbf{x} + f_1\mathbf{L}\mathbf{x} + f_2\mathbf{L}^2\mathbf{x}$ with $f_1 = \frac{\sin(n\omega s)}{s}$ and $f_2 = \frac{1-\cos(n\omega s)}{s^2}$, which is evaluable in $O(d)$ time via a few inner products. By the minimal polynomial $\lambda(\lambda^2+s^2)$, $\mathbf{L}^3 = -s^2\mathbf{L}$; expanding $\exp(\eta\mathbf{L})$ and regrouping yields the rank-2 update form used throughout

E Spectral Analysis of GRAPE and Other Methods

In this section, we discuss eigenvalue-level results for Mul-GRAPE generators/exponentials and summarize the unipotent spectra of Add-GRAPE/PI-Add-GRAPE. Throughout, $\mathbf{L}(\mathbf{a}, \mathbf{b}) = \mathbf{a}\mathbf{b}^\top - \mathbf{b}\mathbf{a}^\top \in \mathfrak{so}(d)$, and $\alpha = \|\mathbf{a}\|^2$, $\beta = \|\mathbf{b}\|^2$, $\gamma = \mathbf{a}^\top\mathbf{b}$, $\Delta = \alpha\beta - \gamma^2$, $s = \sqrt{\Delta}$ as in Section 2.

E.1 Rank-2 Plane: Exact Spectrum and Geometric Interpretation

Lemma E.1 (Rank-2 spectrum)**.** For $\mathbf{L} = \mathbf{L}(\mathbf{a}, \mathbf{b})$, the eigenvalues are $\{\pm is\} \cup \{0\}^{d-2}$, and there exists $\mathbf{B} \in \mathrm{SO}(d)$ such that

$$\mathbf{B}^\top\mathbf{L}\mathbf{B} = \begin{bmatrix} s\mathbf{J} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{d-2} \end{bmatrix}, \qquad \mathbf{J} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right).$$

Moreover, $s = \|\mathbf{a}\|\|\mathbf{b}\| \sin\phi$, where $\phi \in [0, \pi]$ is the angle between a and b.

Proof. From Section 2, $\mathbf{L}^2 = -s^2\mathbf{P}_{\mathcal{U}}$ with $\mathcal{U} = \mathrm{span}\{\mathbf{a}, \mathbf{b}\}$, whence the minimal polynomial is $\lambda(\lambda^2 + s^2)$ and $\sigma(\mathbf{U}) = \{\pm is, 0\}$. Choosing an orthonormal basis aligned with $\mathcal{U} \oplus \mathcal{U}^\perp$ yields the claimed form. Finally, $\Delta = \alpha\beta - \gamma^2 = \|\mathbf{a}\|^2\|\mathbf{b}\|^2(1 - \cos^2\phi) = (\|\mathbf{a}\|\|\mathbf{b}\| \sin\phi)^2$. □

Corollary E.2 (Phase bounds and orthogonality)**.** The per-step rotation angle of $\exp(\eta\mathbf{K})$ on $\mathcal{U}$ equals $\theta = \eta s$ and satisfies $0 \le \theta \le \eta\|\mathbf{a}\|\|\mathbf{b}\|$, with equality when $\mathbf{a} \perp \mathbf{b}$. If $\mathbf{b} = \mathcal{J}\mathbf{a}$ (Section 2.4) and $\|\mathbf{a}\| = 1$, then $s = 1$ and $\theta = \eta$.

Exponential spectrum. For any $n \in \mathbb{Z}$,

$$\sigma\big(\exp(n\mathbf{L})\big) = \{e^{\pm ins}\} \cup \{1\}^{d-2}.$$

Hence $\rho(\exp(n\mathbf{L})) = 1$, the map is unitary (orthogonal), and all Lyapunov exponents are zero. Periodicity holds with fundamental period $T = 2\pi/s$ when $s/\pi \in \mathbb{Q}$; otherwise, the trajectory is quasi-periodic on the unit circle.

E.2 Commuting Multi-subspace Mul-GRAPE and RoPE

Let $\mathbf{L} = \sum_{j=1}^m \theta_j\mathbf{L}_j$ with mutually orthogonal planes (hence $[\mathbf{L}_i, \mathbf{L}_j] = 0$ for $i \ne j$) and $\mathbf{L}_j = \mathbf{U}_j\mathbf{J}\mathbf{U}_j^\top$. Then

$$\mathbf{B}^\top\mathbf{L}\mathbf{B} = \bigoplus_{j=1}^m \theta_j\mathbf{J} \oplus \mathbf{0}_{d-2m}, \qquad \sigma(\mathbf{L}) = \{\pm i\theta_j\}_{j=1}^m \cup \{0\}^{d-2m},$$

for some $\mathbf{Q} \in \mathrm{SO}(d)$. Consequently,

$$\sigma\big(\exp(n\mathbf{L})\big) = \{e^{\pm in\theta_j}\}_{j=1}^m \cup \{1\}^{d-2m}.$$

This recovers RoPE when the planes are the coordinate pairs and $\{\theta_j\}$ follow the canonical log-uniform spectrum (Theorem 3.2).

E.3 ADD-GRAPE AND PATH-INTEGRAL ADDITIVE (PI-ADD-GRAPE)

We now analyze the spectral properties of the additive lifts in GL introduced in Sections 5 and 6. The key structural fact is unipotency: all per-step factors are identity plus a rank-1 (or few-rank) nilpotent update of index 2.

Setup. Let $\mathbf{A} \in \mathfrak{gl}(d{+}1)$ (or $\mathfrak{gl}(d{+}2)$ for ALiBi) satisfy $\mathbf{A}^2 = \mathbf{0}$ as in (5.1) and (5.6). For a scalar path parameter $s \in \mathbb{R}$, define the unipotent factor

$$\mathbf{H}(s) \;:=\; \exp(s\mathbf{A}) \;=\; \mathbf{I} + s\,\mathbf{A}, \qquad \mathbf{H}(s)^{-1} = \mathbf{I} - s\,\mathbf{A}, \qquad \det \mathbf{H}(s) = 1.$$

For GRAPE-Additive (GRAPE-Add) with offset $m = j{-}i$, $s = m\,\omega$; for GRAPE-PA, $s = s_h(t,j) := \sum_{\ell=j+1}^{t} \psi_h(t,\ell)$ from (6.2).

Proposition E.3 (Eigenvalues and Jordan structure of additive lifts)**.** Let $\mathbf{A} \in \mathfrak{gl}(D)$ satisfy $\mathbf{A}^2 = \mathbf{0}$ and $\mathbf{A} \neq \mathbf{0}$. Then for every $s \neq 0$,

$$\sigma\big(\mathbf{H}(s)\big) = \{1\}^D, \qquad (\mathbf{H}(s) - \mathbf{I})^2 = \mathbf{0}, \qquad \det \mathbf{H}(s) = 1, \qquad \rho(\mathbf{H}(s)) = 1.$$

Hence, the minimal polynomial of $\mathbf{H}(s)$ is $(\lambda - 1)^2$, and the Jordan form consists of size-2 Jordan blocks for the 1-eigenspace, with the number of nontrivial blocks equal to $\mathrm{rank}(\mathbf{A})$.

Proof. Since $\mathbf{A}^2 = \mathbf{0}$, $\exp(s\mathbf{A}) = \mathbf{I} + s\mathbf{A}$ and $(\mathbf{H}(s) - \mathbf{I})^2 = s^2\mathbf{A}^2 = \mathbf{0}$. The characteristic polynomial is $(\lambda - 1)^D$ for $\mathbf{H}(s)$, so all eigenvalues equal 1. The determinant equals the product of eigenvalues, hence 1; the spectral radius is therefore 1. □

Dictionary closure. If $\{\mathbf{A}_r\}_{r=1}^R$ satisfy $\mathbf{A}_r^2 = \mathbf{0}$ and $\mathbf{A}_r\mathbf{A}_s = \mathbf{0}$ for all r, s, then

$$\Big(\sum_r \theta_r \mathbf{A}_r\Big)^2 = \sum_r \theta_r^2 \mathbf{A}_r^2 + \sum_{r \neq s} \theta_r\theta_s \mathbf{A}_r\mathbf{A}_s = \mathbf{0},$$

so the combined generator is also index-2 nilpotent and yields the same unipotent spectrum.

Singular values. Although $\mathbf{H}(s)$ is not orthogonal, its deviation from $\mathbf{I}$ is rank-limited and exactly analyzable. We first give a sharp, explicit formula for the canonical rank-1 case (ALiBi block), then a general bound.

Lemma E.4 (Exact singular-value pair for a canonical rank-1 unipotent)**.** Let $E := \mathbf{e}_p\,\mathbf{e}_q^\top$ with $p \neq q$ and define $H(s) := \mathbf{I} + sE \in \mathbb{R}^{D\times D}$. Then $D-2$ singular values equal 1, and the remaining two are

$$\sigma_\pm(H(s)) \;=\; \sqrt{1 + \tfrac{s^2}{2} \;\pm\; |s|\sqrt{1 + \tfrac{s^2}{4}}}\,, \qquad \sigma_+(H(s))\,\sigma_-(H(s)) = 1. \tag{E.1}$$

In particular, $\kappa_2(H(s)) = \sigma_+(H(s))/\sigma_-(H(s)) = 1 + 2|s| + O(s^2)$ as $s \to 0$.

Proof. The action of $H(s)^\top H(s)$ is identity on $\mathrm{span}\{\mathbf{e}_p, \mathbf{e}_q\}^\perp$. In the basis $\{\mathbf{e}_q, \mathbf{e}_p\}$ it equals $\left(\begin{smallmatrix} 1+s^2 & s \\ s & 1 \end{smallmatrix}\right)$, whose eigenvalues are $1 + \frac{s^2}{2} \pm |s|\sqrt{1 + \frac{s^2}{4}}$. Taking square roots yields (E.1). The product equals $\sqrt{\det(H^\top H)} = |\det H| = 1$. □

Corollary E.5 (ALiBi and GRAPE-Additive (GRAPE-Add) conditioning numbers)**.** For the exact ALiBi generator in (5.6), $E = \mathbf{e}_{d+2}\mathbf{e}_{d+1}^\top$ and $s = m\,\beta_h$, so the only nontrivial singular values of $\mathbf{G}_{\mathrm{add},h}(m) = \mathbf{I} + s\mathbf{E}$ are given by (E.1). For the single-vector additive lift (5.1) with $\mathbf{A} = \left(\begin{smallmatrix} \mathbf{0} & \mathbf{u} \\ \mathbf{0}^\top & 0 \end{smallmatrix}\right)$ and $\|\mathbf{u}\| = 1$, the same formula holds with $\mathbf{E}$ replaced by an orthogonally similar rank-1 update and $s = m\,\omega$.

Lemma E.6 (General operator-norm bounds for index-2 unipotents)**.** For any $\mathbf{A}$ with $\mathbf{A}^2 = \mathbf{0}$ and any $s \in \mathbb{R}$,

$$1 - |s|\,\|\mathbf{A}\|_2 \;\le\; \sigma_{\min}(\mathbf{I} + s\mathbf{A}) \;\le\; \sigma_{\max}(\mathbf{I} + s\mathbf{A}) \;\le\; 1 + |s|\,\|\mathbf{A}\|_2.$$

When $\mathrm{rank}(\mathbf{A}) = 1$ and $\|\mathbf{A}\|_2 = 1$, these bounds are tight and coincide with Lemma E.4 at first order in $|s|$.

Proof. Use the triangle inequality $\|(\mathbf{I}+s\mathbf{A})\mathbf{x}\|_2 \le \|\mathbf{x}\|_2 + |s|\,\|\mathbf{A}\|_2\|\mathbf{x}\|_2$ and its reverse form applied to $(\mathbf{I}+s\mathbf{A})^{-1} = \mathbf{I} - s\mathbf{A}$; see also Weyl inequalities for singular values under rank-1 perturbations. □

Cancellation in the relative logit. While $\mathbf{H}(s)$ can be anisotropic (Lemma E.4), the GRAPE-Additive (GRAPE-Add) scoring uses a paired inverse-transpose (Eq. (5.2)), which cancels all multiplicative distortions and yields a pure additive term:

$$\widetilde{\mathbf{q}}_i^\top \widetilde{\mathbf{k}}_j = \widehat{\mathbf{q}}_i^\top \big(\mathbf{I} + i\,\mathbf{A}\big)^\top \big(\mathbf{I} - j\,\mathbf{A}^\top\big)\,\widehat{\mathbf{k}}_j \;=\; \widehat{\mathbf{q}}_i^\top \big(\mathbf{I} + (i-j)\,\mathbf{A}^\top\big)\widehat{\mathbf{k}}_j \;=\; \widehat{\mathbf{q}}_i^\top \mathbf{G}_{\mathrm{add}}(j-i)^{-\top}\widehat{\mathbf{k}}_j,$$

since $(\mathbf{A}^\top)^2 = \mathbf{0}$. This reproduces the exact relative law (5.3) and the closed form (5.4) (e.g. Eq. (5.5)), independently of $\sigma_\pm(H(s))$.

PI-Add-GRAPE as a path-integral unipotent. Fix a head h and endpoint t. The per-row path product in Section 6 is

$$\prod_{\ell=j+1}^{t} \big(\mathbf{I} + \psi_h(t,\ell)\,E\big) \;=\; \mathbf{I} + \Big(\sum_{\ell=j+1}^{t} \psi_h(t,\ell)\Big)E \;=\; \mathbf{I} + s_h(t,j)\,E,$$

because $E^2 = \mathbf{0}$. Thus PI-Add-GRAPE inherits the unipotent spectrum of Prop. E.3 with row-dependent $s = s_h(t,j) \le 0$ (since $\psi_h \le 0$ by construction). Its only two nontrivial singular values are exactly (E.1) with $s \mapsto s_h(t,j)$; the rest equal 1. Consequently,

$$\kappa_2\big(\text{PA factor}\big) \;=\; \frac{\sigma_+\big(s_h(t,j)\big)}{\sigma_-\big(s_h(t,j)\big)} \;=\; 1 + 2\,|s_h(t,j)| + O\big(s_h(t,j)^2\big),$$

while the determinant remains 1 and eigenvalues are all 1. As in GRAPE-Additive (GRAPE-Add), the paired inverse-transpose used in the bilinear scoring removes any multiplicative anisotropy, leaving the bounded additive term $b_h(t,j)$ in Eq. (6.2).

Implications. Now we summarize the implications of previous results. For all s, $\mathbf{H}(s)$ is invertible with $\mathbf{H}(s)^{-1} = \mathbf{I} - s\mathbf{A}$; eigenvalues do not grow with offset length (spectral radius $= 1$). The operator norm grows at most linearly in $|s|$ (Lemma E.6) and is exactly characterized in the rank-1 canonical cases (Lemma E.4).

Secondly, $\det \mathbf{H}(s) = 1$ implies no net volume change; any expansion along one direction is exactly balanced by contraction along its paired direction (product $\sigma_+\sigma_- = 1$). Despite anisotropy, the Add-GRAPE and PI-Add-GRAPE logits remain exactly relative because the key transform uses $\mathbf{H}(s)^{-\top}$, algebraically eliminating multiplicative distortion and yielding the closed-form additive bias (Eqs. (5.3), (5.4), (6.2)).

E.4 Comparison to PaTH Attention

PaTH Attention (Yang et al., 2025) proposes a contextual multiplicative position map given by a cumulative product of identity-plus-rank-one matrices

$$\mathbf{H}_t = \mathbf{I} - \beta_t\,\mathbf{w}_t\mathbf{w}_t^\top, \qquad \|\mathbf{w}_t\|_2 = 1, \quad \beta_t \in (0,2),$$

applied along the path between key position j and query position i as $\prod_{s=j+1}^{i} \mathbf{H}_s$ (see Section 2 of the PaTH paper). In contrast to **Mul-GRAPE** factors, each $\mathbf{H}_t$ is *not* orthogonal unless $\beta_t \in \{0,2\}$. This has immediate spectral consequences.

Per-step spectrum. Since $\mathbf{H}_t$ is symmetric rank-1 perturbation of the identity with projector $\mathbf{P}_t := \mathbf{w}_t\mathbf{w}_t^\top$,

$$\sigma(\mathbf{H}_t) = \{\,1-\beta_t,\, \underbrace{1,\ldots,1}_{d-1}\,\}, \qquad \det(\mathbf{H}_t) = 1-\beta_t, \qquad \|\mathbf{H}_t\|_2 = \max\{1, |1-\beta_t|\} = 1.$$

Thus $\mathbf{H}_t$ is norm nonexpansive (operator norm 1) but *not norm-preserving* unless $\beta_t \in \{0,2\}$. Singular values equal the absolute eigenvalues because $\mathbf{H}_t$ is symmetric; the component along $\mathbf{w}_t$ is scaled by $|1-\beta_t| < 1$ for any $\beta_t \in (0,2)\setminus\{0,2\}$, and flips sign when $\beta_t > 1$ (a design choice in PaTH to allow negative eigenvalues for state-tracking).

Path product is contractive and near-singular. Let $\mathbf{P}_{j\to i} = \prod_{s=j+1}^{i} \mathbf{H}_s$. Submultiplicativity of singular values gives

$$\sigma_{\max}(\mathbf{P}_{j\to i}) \le \prod_{s=j+1}^{i} \|\mathbf{H}_s\|_2 = 1, \qquad \sigma_{\min}(\mathbf{P}_{j\to i}) \ge \prod_{s=j+1}^{i} \sigma_{\min}(\mathbf{H}_s) = \prod_{s=j+1}^{i} |1-\beta_s|.$$

Hence $\mathbf{P}_{j\to i}$ is (at best) nonexpansive, with a worst-case exponential lower bound on the smallest singular value governed by the path-length product of $|1-\beta_s|$. Whenever some β_s is close to 1, $\mathbf{H}_s$ is nearly singular (and exactly singular if $\beta_s = 1$), driving $\sigma_{\min}(\mathbf{P}_{j\to i})$ toward zero. Volume contraction is quantified by

$$\det(\mathbf{P}_{j\to i}) = \prod_{s=j+1}^{i} (1-\beta_s),$$

which typically decays exponentially in $i-j$ unless β_s concentrates at the orthogonal endpoints $\{0,2\}$.

Aligned-plane special case. If the directions are time-invariant, $\mathbf{w}_s \equiv \mathbf{w}$, then $\mathbf{P}_t = \mathbf{w}\mathbf{w}^\top$ is an idempotent projector and the factors commute:

$$\prod_{s=j+1}^{i} \mathbf{H}_s = \prod_{s=j+1}^{i} \big(\mathbf{I} - \beta_s \mathbf{P}\big) = \mathbf{I} - \Big(1 - \prod_{s=j+1}^{i} (1-\beta_s)\Big)\mathbf{P},$$

so the eigenvalue along w is exactly $\prod_{s=j+1}^{i}(1-\beta_s)$, making the contraction along w explicit and exponential in path length unless $\beta_s \in \{0,2\}$.

Implications for long-context modeling. Because the PaTH transport multiplies the Q/K bilinear by $\mathbf{P}_{j\to i}$, any persistent deviation of β_t from $\{0,2\}$ yields cumulative energy loss along a moving one-dimensional subspace. This concentrates mass in progressively fewer directions and can flatten or attenuate long-range logits $\mathbf{q}_i^\top \mathbf{P}_{j\to i}\mathbf{k}_j$ as $i-j$ grows, unless additional renormalizations or forget-gates are introduced. In contrast, **Mul-GRAPE** maps lie in $\mathrm{SO}(d)$, so for both non-contextual and contextual types, all singular values are 1; volumes and norms are preserved, and Lyapunov exponents are 0, avoiding contraction-induced degradation of long-range interactions.

Lemma E.7 (Orthogonality condition for PaTH factors)**.** For $\mathbf{H}_t = \mathbf{I} - \beta_t \mathbf{w}_t \mathbf{w}_t^\top$ with $\|\mathbf{w}_t\| = 1$, $\mathbf{H}_t$ is orthogonal iff $\beta_t \in \{0,2\}$. For $\beta_t \in (0,2) \setminus \{0,2\}$, $\mathbf{H}_t$ is symmetric, diagonalizable with eigenvalues in $(-1,1] \cup \{1\}$, and strictly contractive on $\mathrm{span}\{\mathbf{w}_t\}$.