Higher-Dimensional Rotary Position Embedding

Anonymous ACL submission

Abstract

Transformers rely on position embedding mechanisms in long context modeling in most cases. Rotary Position Embedding (RoPE) embeds positional information with independent 2D rotations, forming relative position terms in self-attention. However, its pairwise, block-based, and decoupled structure limits deep mixing and robustness across channels. We propose HD-RoPE, which extends RoPE from independent 2D rotations to higher-dimensional rotations and introduces a Paley-I orthogonal basis to obtain balanced, isotropic, and dense phase mixing within each rotation subspace. This significantly enhances channel coupling and rotational degrees of freedom while maintaining orthogonal stability and the relative position closure property. Furthermore, HD-RoPE is easily optimized for engineering efficiency without introducing additional trainable parameters. We have conducted extensive evaluation results demonstrating that HD-RoPE achieves significant performance improvements over standard RoPE across various popular benchmarks and in both long and short contexts.

1 Introduction

Transformer architecture has established itself as the dominant backbone for large language models (LLMs) across a wide range of domains. They heavily rely on efficient positional encoding mechanisms in long sequence modeling. Rotary Position Embedding (RoPE; Su et al., 2024), because it encodes absolute position as rotation and explicitly forms relative position signals in self-attention, has become one of the default choices for current mainstream language models.

When models need to extrapolate from the pre-training context length to longer contexts, RoPE exhibits typical degradation. Hence, a large body of effort on RoPE has focused on modifying the frequency schedule or angular velocity, such as rescaling the RoPE base, interpolating positions, or selectively weakening positional bias (Chen et al., 2023; Peng et al., 2023; Ding et al., 2024; Yang et al., 2025). These methods improve how phases evolve across positions.

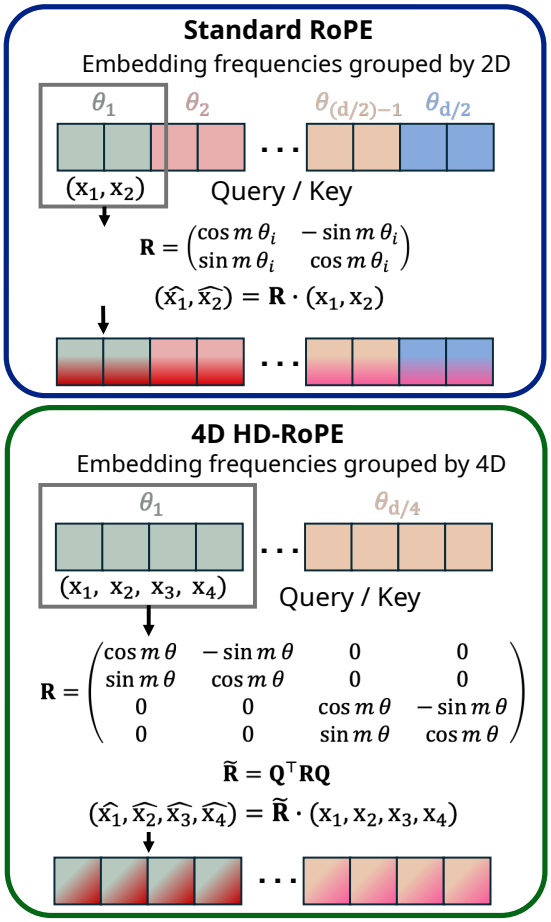


Figure 1: The illustration of our HD-RoPE. Q denotes the selected orthogonal matrix in Section 3, enabling more isotropic, balanced, and dense phase interaction.

However, a more fundamental assumption of RoPE that position is injected only through several mutually independent two-dimensional rotations has largely remained untouched. RoPE can

be roughly understood as: projecting the representation vector of each token onto several two-dimensional planes, and rotating it within each plane by a position-dependent angle (phase). The physical meaning of two-dimensional rotation is very intuitive and sound, while it also indicates that standard RoPE binds each positional frequency to an isolated two-dimensional channel pair. In this case, each relative phase can interact with only two fixed-dimension Q/K features, making the phase-sensitive attention term sparse and pairwise.

Based on the above intuition, this paper proposes **HD-RoPE**, a method for injecting positional information that generalizes the two-dimensional RoPE system to an higher N-dimensional rotation space. HD-RoPE groups channels into sets of N, applying rotations within N-dimensional subspaces. This changes the rotation unit of RoPE from an isolated 2D pair to a higher-dimensional phase subspace, allowing the same positional phase to be shared and mixed across more Q/K channels. Within these subspaces, we instantiate the mixing basis using Paley-I conference matrices (Paley, 1933) to obtain a ***more isotropic, balanced, and dense phase interaction***, enabling positional phases to couple more sufficiently with Q/K channels under the basic invariants of RoPE, which is effective, efficient and parameter-free.

This significantly enhances channel mixing and rotational degrees of freedom while maintaining the stability and norm preservation of orthogonal transformations, further enhancing the relative position encoding capability of RoPE. The main contributions of HD-RoPE include:

1. We propose a higher-dimensional rotational perspective, providing a general framework for high-dimensional scaling, wherein 4D represents the sweet spot for balancing performance and efficiency.

2. We adopt the conference matrix as a fixed and structurally simple orthogonal mixing transformation (balanced, low-coherence, and parameter-free) to enhance better cross-channel coupling without introducing additional trainable parameters.

3. We conducted extensive experiments and found that HD-RoPE significantly outperforms the original RoPE in various tasks with various settings, while maintaining almost no additional computational overhead.

2 Preliminary

Self-Attention Given an input sequence of length N, its embedding without positional information is denoted as $\mathbf{E_N} = \{\mathbf{x}_i\}_{i=1}^N$, where $\mathbf{x_i} \in \mathbb{R}^d$. The self-attention of Transformer maps the representation of each position to:

$$\mathbf{q}_m = f_q(\mathbf{x}_m, m),\ \mathbf{k}_n = f_k(\mathbf{x}_n, n). \tag{1}$$

The output is then calculated as:

$$\begin{aligned} \mathbf{a}_{m,n} &= \frac{\exp\left(\mathbf{q}_m^\top \mathbf{k}_n / \sqrt{d}\right)}{\sum_{j=1}^N \exp\left(\mathbf{q}_m^\top \mathbf{k}_j / \sqrt{d}\right)}, \\ \mathbf{o}_m &= \sum_{n=1}^N a_{m,n} \mathbf{v}_n. \end{aligned} \tag{2}$$

The essential function of positional encoding is to design f_q, f_k, f_v or their equivalent implementations so that the inner product $\mathbf{q}_m^\top \mathbf{k}_n$ explicitly carries the positional signal in the attention weights.

Rotary Position Embedding (RoPE) RoPE starts with the relative position objective and imposes structured constraints on the attention inner product. It aims to find a function g such that:

$$\langle f_q(\mathbf{x}_m, m),\ f_k(\mathbf{x}_n, n)\rangle = g(\mathbf{x}_m, \mathbf{x}_n, m-n). \tag{3}$$

In this form, position is injected into the attention only in the relative form $(m-n)$. RoPE provides a multiplicative implementation: the linearly projected vector is rotated in terms of position in several two-dimensional subspaces, thereby injecting a controllable relative phase difference into the inner product without changing the vector norm.

Specifically, when d is even, $\mathbb{R}^d$ is divided into $d/2$ two-dimensional subspaces and a block diagonal rotation matrix is defined:

$$\begin{aligned} \mathbf{R}^d_{\Theta,m} &= \operatorname{diag}\left(\mathbf{R}^{(2)}_{\theta_1,m}, \mathbf{R}^{(2)}_{\theta_2,m}, \ldots, \mathbf{R}^{(2)}_{\theta_{d/2},m}\right), \\ \mathbf{R}^{(2)}_{\theta_i,m} &= \begin{pmatrix} \cos m\theta_i & -\sin m\theta_i \\ \sin m\theta_i & \cos m\theta_i \end{pmatrix}. \end{aligned} \tag{4}$$

Moreover, f is defined as:

$$f_{\{q,k\}}(\mathbf{x}_m, m) = \mathbf{R}^d_{\Theta,m} \mathbf{W}_{\{q,k\}} \mathbf{x}_m. \tag{5}$$

Orthogonality $\mathbf{R}^\top_{\Theta,m}\mathbf{R}_{\Theta,m} = \mathbf{I}$ guarantees the numerical stability and norm preservation of the encoding process: for any vector $\mathbf{z}$, $\|\mathbf{R}_{\Theta,m}\mathbf{z}\|_2 = \|\mathbf{z}\|_2$. Rotation does not change the magnitude of

Algorithm 1 RoPE (2D Rotation)

Input: embedding $\mathbf{t}$, rotation frequencies $\mathbf{freqs}$
$\cos \leftarrow \cos(\text{freqs})$; $\sin \leftarrow \sin(\text{freqs})$
Split $\mathbf{t}$: $(\mathbf{x}_1, \mathbf{x}_2) \leftarrow \text{Chunk}(\mathbf{t}, 2)$
$\text{RotateHalf}(\mathbf{t}) \leftarrow \text{Concat}(-\mathbf{x}_2,\ \mathbf{x}_1)$
$\mathbf{t} \leftarrow \mathbf{t} \odot \cos\ + \text{RotateHalf}(\mathbf{t}) \odot \sin$
Output: updated embedding $\mathbf{t}$

Algorithm 2 HD-RoPE (4D Rotation; Paley-I)

Input: embedding $\mathbf{t}$, rotation frequencies $\mathbf{freqs}$
$\cos \leftarrow \cos(\text{freqs})$; $\sin \leftarrow \sin(\text{freqs})$
Split $\mathbf{t}$: $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \leftarrow \text{Chunk}(\mathbf{t}, 4)$
$\mathbf{t}_1 \leftarrow \text{Concat}(-\mathbf{x}_2,\ \mathbf{x}_1,\ \mathbf{x}_4,\ -\mathbf{x}_3)$
$\mathbf{t}_2 \leftarrow \text{Concat}(\mathbf{x}_3,\ \mathbf{x}_4,\ \mathbf{x}_2,\ -\mathbf{x}_1)$
$\mathbf{t}_3 \leftarrow \text{Concat}(\mathbf{x}_4,\ -\mathbf{x}_3,\ -\mathbf{x}_1,\ -\mathbf{x}_2)$
$\mathbf{t} \leftarrow \mathbf{t} \odot \cos\ + \mathbf{t}_1 \odot \frac{1}{3}\sin\ + \mathbf{t}_2 \odot \frac{2}{3}\sin\ + \mathbf{t}_3 \odot \frac{2}{3}\sin$
Output: updated embedding $\mathbf{t}$

the features, thus possessing numerical stability. More importantly, it produces an explicit relative positional form in the attention inner product:

$$\begin{aligned}\mathbf{q}_m^\top \mathbf{k}_n &= (\mathbf{R}^d_{\Theta,m}\mathbf{W}_q\mathbf{x}_m)^\top(\mathbf{R}^d_{\Theta,n}\mathbf{W}_k\mathbf{x}_n) \\ &= \mathbf{x}_m^\top \mathbf{W}_q^\top \mathbf{R}^d_{\Theta,n-m}\mathbf{W}_k\mathbf{x}_n,\end{aligned} \tag{6}$$

where $\mathbf{R}^d_{\Theta,n-m} = (\mathbf{R}^d_{\Theta,m})^\top \mathbf{R}^d_{\Theta,n}$ depends only on the relative displacement $(n-m)$. In practice, a set of multi-scale frequencies $\Theta = \{\theta_i\}$ is often chosen (e.g., covering different frequency bands according to an exponential law) so that the model can simultaneously characterize short-range and long-range relative position signals.

3 Method

3.1 HD-RoPE's Orthogonal conjugation preserves RoPE invariants

HD-RoPE aims to introduce higher-dimensional rotation for positional modeling. Specifically, we employ orthogonal conjugation to reparameterize the rotation matrix by orientation to introduce cross-dimensional mixing. Given any orthogonal matrix $\mathbf{Q}$ with $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$, we define:

$$\widetilde{\mathbf{R}}_{\Theta,m} \triangleq \mathbf{Q}^\top \mathbf{R}_{\Theta,m}\mathbf{Q}. \tag{7}$$

Based on orthogonality and conjugate structure:

$$\begin{aligned}\widetilde{\mathbf{R}}^\top_{\Theta,m}\widetilde{\mathbf{R}}_{\Theta,n} &= \mathbf{Q}^\top \mathbf{R}^\top_{\Theta,m}\mathbf{R}_{\Theta,n}\mathbf{Q} \\ &= \mathbf{Q}^\top \mathbf{R}_{\Theta,n-m}\mathbf{Q}.\end{aligned} \tag{8}$$

Then $\widetilde{\mathbf{R}}_{\Theta,m}$ remains orthogonal, retaining the two critical properties: (a). Norm Preservation: $\|\widetilde{\mathbf{R}}_{\Theta,m}\mathbf{z}\|_2 = \|\mathbf{z}\|_2$; (b). Distance Preservation: $\widetilde{\mathbf{R}}_{\Theta,n-m} = \widetilde{\mathbf{R}}^\top_{\Theta,m}\widetilde{\mathbf{R}}_{\Theta,n}$. Therefore, while ***preserving relative displacement encoding and stable rotation of the original RoPE***, $\mathbf{Q}$ allows us to change the coordinate system in which the rotation is applied, ***increasing higher-cross-dimensional mixing and rotational degrees of freedom***.

3.2 Better Orthogonal Transformation Q Selection for HD-RoPE

Conference matrices as structured orthogonal family. In the classical lineage of orthogonal matrix research, many constructions (such as Haar random orthogonality, Gram-Schmidt/QR generation, etc.) provide good isotropy, but often rely on randomness or numerical decomposition, leading to low reproducibility and additional overhead. We focus on a family of orthogonal matrices that are ***scalable, numerically stable and reproducible, and provide sufficient cross-dimensional mixing***. Under these conditions, the available orthogonal families are limited: Hadamard/Schur type, discrete cosine/sine transforms, and sparse orthogonal parameterization based on givens rotations are all insufficient to meet the requirements. More details are in Appendix A.5.

The **conference matrices** offer a compromise between theoretical properties and engineering friendliness: their discrete structures with elements of $\{0, \pm 1\}$ make their implementations close to sparse addition and naturally exhibit strong cross-dimensional coupling. Therefore, we adopt the conference matrix as the primary source of construction for the orthogonal hybrid transform $\mathbf{Q}_N$ in HD-RoPE. The N-order conference matrix $\mathbf{C}_N \in \{0, \pm 1\}^{N\times N}$ satisfies:

$$\text{diag}(\mathbf{C}_N) = \mathbf{0}, \quad \mathbf{C}_N\mathbf{C}_N^\top = (N-1)\mathbf{I}. \tag{9}$$

The corresponding N-order orthogonal matrix can be obtained by:

$$\mathbf{Q}_N \triangleq \frac{1}{\sqrt{N-1}}\mathbf{C}_N, \quad \mathbf{Q}_N^\top\mathbf{Q}_N = \mathbf{I}. \tag{10}$$

This construction has two practical advantages: (i) The element values of $\mathbf{Q}_N$ are ***structured*** (only $0, \pm 1/\sqrt{N-1}$), making it easy to store and implement quickly (essentially a sign flip and sparse

summation); and (ii) Given N and the existence of a conference matrix of that order, $\mathbf{Q}_N$ can be directly generated by a deterministic construction algorithm, thus providing a ***consistent depth-mixing coordinate transformation*** for HD-RoPE without the need for additional learning parameters.

Paley-I. Paley-I (Paley, 1933) is a special conference matrices with good properties as $\mathbf{Q}$. The special advantage of Paley-I lies in its balanced discrete structure. Its entries are restricted to $\{0, \pm 1\}$ before normalization, and each coordinate is mixed with all other coordinates through equal-magnitude coefficients after normalization. This yields a low-coordinate-bias basis: no single original channel dominates the transformed subspace, and the positional phase is distributed more uniformly across the N-dimensional block. Therefore, it provides a ***deterministic, balanced, isotropic, and parameter-free basis*** in which higher-dimensional rotary phases can be applied ***more uniformly and more sufficiently across query/key channels***. Our HD-RoPE adopts Paley-I as $\mathbf{Q}$ in Eq. (7).

Dimension handling. It is important to emphasize that the Paley-I construction for conference matrix does not exist for any dimension N; its existence is strongly constrained by combinatorial design and algebraic constraints. For example, the available dimension options below 32 include only: 2, 4, 8, and 32, but it is already sufficient to meet our main needs. Details are in the Appendix A.6.

3.3 Our Higher-Dimensional RoPE

3.3.1 Rotation Matrix

We attempt to construct a position-dependent rotation operator that (i) is orthogonal and norm-preserving, (ii) preserves RoPE's relative-position structure, and (iii) induces stronger cross-channel coupling by rotating in mixed directions within an N-dimensional subspace.

Let the hidden dimension be d and choose an even N. We partition $\mathbb{R}^d$ into $B = d/N$ disjoint N-dimensional blocks and apply the same construction to each block. Within an N-dimensional block, we first define a *pairwise* (RoPE-isomorphic) rotation as:

$$\mathbf{R}_{\text{pair}}(p) = \operatorname{diag}\big(\mathrm{R}(\omega_1 p), \ldots, \mathrm{R}(\omega_{N/2} p)\big), \quad (11)$$

$$\mathrm{R}(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}. \quad (12)$$

To obtain a dense N-dimensional rotation with enhanced coupling, we perform an orthogonal change of basis as follows:

$$\mathbf{R}_N(p) \triangleq \mathbf{Q}_N^\top \mathbf{R}_{\text{pair}}(p)\, \mathbf{Q}_N, \quad \mathbf{Q}_N \in SO(N). \quad (13)$$

Therefore, attention scores continue to depend on the relative displacement $(s-p)$, while the rotation directions are no longer confined to fixed 2D coordinate planes, yielding richer directional couplings inside each N-dimensional subspace.

3.3.2 A 4D Case

Similar to RoPE, we first consider the simple case with dimension $d = 4$. The goal is to construct a function g such that for any word vector $x_m, x_n \in \mathbb{R}^4$ and position $m, n \in \mathbb{Z}$. We have:

$$\langle f_q(x_m, m), f_k(x_n, n)\rangle = g(x_m, x_n, m-n). \quad (14)$$

Specifically, we further express $f\{q,k\}$ in the form of a multiplication matrix:

$$f_{\{q,k\}}(\boldsymbol{x}_m, n) = \mathbf{Q}^\top \mathbf{R}\mathbf{Q} \cdot \mathbf{W}\mathbf{x}, \quad (15)$$

where

$$\mathbf{R} = \begin{pmatrix} \cos m\theta & -\sin m\theta & 0 & 0 \\ \sin m\theta & \cos m\theta & 0 & 0 \\ 0 & 0 & \cos m\theta & -\sin m\theta \\ 0 & 0 & \sin m\theta & \cos m\theta \end{pmatrix}, \quad (16)$$

$$\mathbf{W} = \begin{pmatrix} W_{\{q,k\}}^{(11)} & \cdots & W_{\{q,k\}}^{(14)} \\ \vdots & \ddots & \vdots \\ W_{\{q,k\}}^{(41)} & \cdots & W_{\{q,k\}}^{(44)} \end{pmatrix}, \quad (17)$$

$$\mathbf{x} = \begin{pmatrix} x_m^{(1)} & x_m^{(2)} & x_m^{(3)} & x_m^{(4)} \end{pmatrix}^\top. \quad (18)$$

The orthogonal matrix $\mathbf{Q}_4$ is derived from the Paley-I conference matrix, noted as:

$$\mathbf{C}_4 = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & -1 & 1 & 0 \end{bmatrix}, \quad \mathbf{Q}_4 = \frac{1}{\sqrt{3}}\mathbf{C}_4. \quad (19)$$

The conjugated rotation $\mathbf{Q}_4^\top \mathbf{R}\mathbf{Q}_4$ preserves the norm and the relative-position closure with the desired characteristic stated in Section 3.2, replacing the original RoPE's pairwise phase interaction with a dense 4D phase interaction.

#Params	Rotation	θ	HellaSwag	ARC-E	ARC-C	WinoGrande	BoolQ	AVG
650M	RoPE	10^4	26.69	48.87	26.58	51.29	52.09	41.10
		5×10^5	28.00	51.52	28.28	51.04	54.53	<u>42.68</u>
	3D-RPE	10^4	26.21	48.36	26.14	50.76	52.70	40.83
		5×10^5	27.41	50.91	27.63	50.52	54.10	42.11
	FoPE	10^4	26.56	48.86	26.58	51.12	53.04	41.23
		5×10^5	27.86	51.51	28.15	51.23	54.64	42.67
	HD-RoPE	10^4	27.39	50.53	28.25	51.39	54.76	42.46
		5×10^5	28.93	51.62	27.18	52.91	54.98	**43.12**
1.3B	RoPE	10^4	36.37	61.36	39.52	62.32	63.22	52.56
		5×10^5	37.20	63.21	40.62	63.11	62.73	53.37
	3D-RPE	10^4	35.76	60.88	38.94	61.84	62.56	52.00
		5×10^5	36.54	62.63	39.97	62.54	62.10	52.76
	FoPE	10^4	36.47	61.46	39.58	62.60	63.12	52.65
		5×10^5	37.16	63.02	40.67	63.27	62.78	53.38
	HD-RoPE	10^4	37.68	62.57	40.95	63.58	63.01	<u>53.56</u>
		5×10^5	41.43	63.45	42.54	64.59	66.47	**55.70**

Table 1: Main results on general benchmarks with different model sizes and mainstream θ selections.

3.3.3 General Form

Let $\mathbf{Q}_4 \in \mathbb{R}^{4\times 4}$ be an orthogonal matrix in a 4-dimensional subspace ($\mathbf{Q}_4^\top \mathbf{Q}_4 = \mathbf{I}_4$). Extending this to a block diagonal form in d dimensions:

$$\mathbf{Q}_4^d \triangleq \mathrm{diag}(\mathbf{Q}_4, \ldots, \mathbf{Q}_4) \in \mathbb{R}^{d\times d}. \tag{20}$$

The rotation matrix after the transformation corresponding to Section 3.3.2 is:

$$\begin{aligned}\widetilde{\mathbf{R}}_{4,\Theta,m}^d &\triangleq (\mathbf{Q}_4^d)^\top \mathbf{R}_{\Theta,m}^d \mathbf{Q}_4^d \\ &= \mathrm{diag}(\mathbf{R}_{\theta_1,m}^4, \ldots, R_{\theta_{d/4},m}^4) \in \mathbb{R}^{d\times d}.\end{aligned} \tag{21}$$

We divide the d-dimensional space into d/4 subspaces and combine them using the linear property of the inner product, transforming $f\{q,k\}$ into:

$$\begin{aligned}f_{q,k}(\boldsymbol{x}_m, m) &= \widetilde{\mathbf{R}}_{4,\Theta,m}^d \mathbf{W}_{\{q,k\}} \boldsymbol{x}_m \\ &= (\mathbf{Q}_4^d)^\top \mathbf{R}_{4,\Theta,m}^d \mathbf{Q}_4^d \mathbf{W}_{\{q,k\}} \boldsymbol{x}_m.\end{aligned} \tag{22}$$

Similarly, we can extend $\mathbf{Q}_N$ to d dimensions by block-diagonal replication as $\mathbf{Q}_N^d$. The resulting d-D rotation takes form as:

$$\begin{aligned}\widetilde{\mathbf{R}}_{N,\Theta,m}^d &\triangleq (\mathbf{Q}_N^d)^\top \mathbf{R}_{N,\Theta,m}^d \mathbf{Q}_N^d \\ &= \mathrm{diag}(\mathbf{R}_{\theta_1,m}^N, \ldots, R_{\theta_{d/N},m}^N) \in \mathbb{R}^{d\times d}.\end{aligned} \tag{23}$$

3.4 Implementation and Efficiency

Since $\mathbf{Q}$ is constant and orthogonal, each entry of $\widetilde{\mathbf{R}}(m)$ remains a fixed linear combination of $\sin(\cdot)$ and $\cos(\cdot)$ terms, while the use of structured conference matrices yields a *dense* mixed rotation (with very few zero entries).

This structure enables the same engineering optimization as RoPE: we precompute the required trigonometric terms and apply the rotation through linear operations that emulate the matrix product, substantially reducing the overhead of HD-RoPE. The full procedure is summarized in Algorithm 2. We provide a more in-depth analysis of the effectiveness of HD-RoPE in the appendix.

4 Experiment

4.1 Experiment Setup

Model setting and training configuration. We trained from scratch on a LLAMA-3 (Grattafiori et al., 2024) style decoder-only Transformer architecture, training two models of different sizes: 650M and 1.3B. All models in Table 1 were trained with 50B tokens. All comparative experiment settings maintained strict consistency, and the details are presented in Appendix A.3.

Pre-training dataset. The pre-training data used the open-source SlimPajama (Shen et al., 2023) corpus. We used the Meta-rater (Zhuang et al., 2025) quality scoring to screen SlimPajama, and selected the Readability-30B, Professionalism-30B, and Reasoning-30B subsets for deduplication and merging to form the final training set.

#Params	Rotation	θ	Passkey-Retrieval	Stack-Select	TextSort	LongBench-v2	AVG
650M	2D	10^4	42.49	39.43	29.58	32.85	36.09
		5×10^5	41.87	38.96	30.94	31.29	35.77
	4D	10^4	41.37	38.49	30.42	34.13	36.10
		5×10^5	40.04	41.13	32.07	35.72	**37.24**
1.3B	2D	10^4	58.84	49.79	46.87	43.66	49.79
		5×10^5	60.50	51.29	48.31	45.22	51.33
	4D	10^4	60.53	50.67	47.76	44.94	50.97
		5×10^5	62.10	52.12	49.06	46.52	**52.45**

Table 2: Evaluation results on long-context benchmarks with different model sizes and θ selections.

Rotation	Orthogonal Matrix ($\mathbf{Q}$)	θ	Short-Text AVG
2D	Vanilla	10^4	51.70
		5×10^5	52.49
4D	Identity ($\mathbf{Q} = \mathbf{I}$)	10^4	51.04
		5×10^5	50.91
	Random	10^4	35.12
		5×10^5	37.92
	Conference (Paley-I)	10^4	51.95
		5×10^5	53.74

Table 3: Ablation Study of Rotation for HD-RoPE.

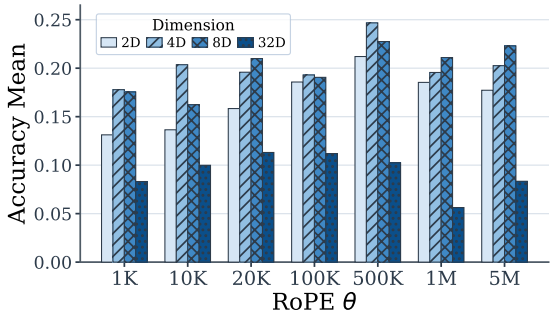


Figure 2: Rotation dimension–θ sweep. The mean accuracy is reported for 2D, 4D, 8D, and 32D rotations under different RoPE base values.

Baselines and evaluations. The major baseline is the standard 2D RoPE (Su et al., 2024). We systematically scanned the RoPE base θ in the range of $[1\text{k}, 5\text{M}]$. In the main comparison, we report representative and best-performing settings, including $\theta = 10\text{k}$ and $\theta = 500\text{k}$, which are widely used in existing LLMs. We also compare with 3D-RPE (Ma et al., 2025) and FoPE (Hua et al., 2024). We evaluate models on five short-context benchmarks and three long-context benchmarks. More details are in Appendix A.2.

4.2 Main Results on General Benchmarks

HD-RoPE outperforms RoPE. Table 1 shows the main results on the short-context benchmarks. HD-RoPE outperformed 2D RoPE on the vast majority of specific tasks and average results. At the same time, HD-RoPE also surpasses other baselines, demonstrating consistency across different tasks, model sizes, and typical RoPE θ.

Stable gains with various θ. This performance improvement is not limited solely to the two specific θ values mentioned above. Under comparable θ settings, our 4D HD-RoPE variant consistently improves the average score over 2D RoPE. Details are shown in Section 4.5.

Efficiency and usability. As in Section 3.4, our implementation references the engineering optimizations of RoPE, performing similar matrix multiplications for linear complexity. HD-RoPE can be applied plug-and-play to various models.

4.3 Long-Context Evaluations

Table 2 reports the long-context evaluation results. The results confirm that HD-RoPE improves long-context performance in both model scales, suggesting that our high-dimensional rotation is beneficial for tasks requiring long-range dependency modeling and long-context reasoning.

To further evaluate HD-RoPE's long-context model ability, we implement our model and baseline with popular long-context extrapolation methods, i.e., YaRN (Peng et al., 2023) and LongRoPE (Ding et al., 2024). The results in Table 4 clearly indicate that HD-RoPE achieves better overall performance compared to the original 2D RoPE on long-context benchmarks, implying better potential of HD-RoPE in practice.

#Params	Rotation	Method	Passkey-Retrieval	Stack-Select	TextSort	LongBench-v2	AVG
650M	2D	YaRN	39.72	37.84	29.91	30.44	34.48
		LongRoPE	40.58	38.21	30.15	30.88	34.96
	4D	YaRN	39.25	39.88	31.22	34.21	36.14
		LongRoPE	39.61	40.42	31.58	34.76	**36.59**
1.3B	2D	YaRN	58.73	50.14	47.23	44.16	50.07
		LongRoPE	59.31	50.63	47.81	44.58	50.58
	4D	YaRN	60.24	50.95	48.06	45.14	51.10
		LongRoPE	60.91	51.38	48.54	45.61	**51.61**

Table 4: Evaluation results with long-context extrapolation methods (YaRN and LongRoPE), where $\theta = 5 \times 10^5$.

#Params	Rotation	Passkey-Retrieval	Stack-Select	TextSort	LongBench-v2	AVG
1.3B	2D	61.27	54.79	50.13	45.53	52.93
	4D	64.12	55.29	50.51	46.91	**54.21**

Table 5: Evaluation results for the 1.3B model with long-context continue pre-training, $\theta = 5 \times 10^5$.

4.4 Ablation Study

To isolate the contribution of our HD-RoPE, we conduct an ablation study during the transition from standard 2D RoPE to 4D HD-RoPE, while keeping all other training and evaluation configurations identical. All models in the ablation study were trained with 30B tokens. Specifically, we compare the following three settings: (i) ***4D grouping with pairwise RoPE (no mixing).*** We still apply the standard 2D pairwise RoPE rotation, but assign frequencies and organize channels in 4D groups. This corresponds to a degenerate 4D formulation with an identity orthogonal transform, i.e., $\mathbf{Q} = \mathbf{I}$; (ii) ***4D + random orthogonal mixing.*** $\mathbf{Q}$ is obtained by random orthogonalization to test whether orthogonality alone is sufficient to yield improvements; and (iii) ***4D + conference mixing (Paley-I; our final version).***

The results are reported in Table 3. Transitioning from standard 2D RoPE to *4D grouping with pairwise RoPE* yields only limited gains, indicating that merely modifying the frequency indexing without introducing cross-dimensional mixing is insufficient to obtain substantial improvements. Moreover, *4D + random orthogonal mixing* does not consistently improve over baseline and can even degrade performance, suggesting that *orthogonality alone* is not the key factor: unstructured mixing may disrupt the alignment between frequency assignment and channel geometry. The gains of HD-RoPE primarily arise from *structured, deterministic, and strongly mixing* orthogonal conjugation, rather than regrouping or unstructured orthogonal transformations.

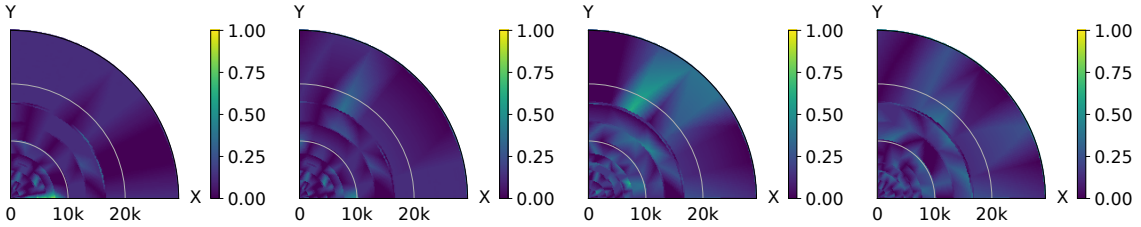
4.5 Effect of Different RoPE Base θ Across Various Rotation Dimensions

We further study how the RoPE base θ interacts with the rotation dimension. Figure 2 aggregates the results across different θ values. The 4D setting achieves the best overall performance, with 8D being close but less stable across θ sweep. In contrast, 32D performs substantially worse, suggesting that simply increasing the rotation dimension is not sufficient. Excessively high-dimensional rotation may over-mix the channel structure, or make the effective frequency allocation less suitable for the model. These results support 4D HD-RoPE as an efficient sweet spot: it introduces sufficient cross-channel coupling while avoiding the degradation observed in overly large rotation groups.

Moreover, we also observe that 4D HD-RoPE consistently outperforms 2D RoPE with different θ, reconfirming the effectiveness of our method.

4.6 Effect of HD-RoPE with Long-Context Continue Pre-training

Based on the model used in our main experiments, we conducted long-context continued pre-training to verify the effectiveness of HD-RoPE, utilizing a sequence length of 32,768 and a total of 13.4B



(a) 2D RoPE. (b) 4D HD-RoPE. (c) 8D HD-RoPE. (d) 32D HD-RoPE.

Figure 3: Passkey-position profile of Passkey Retrieval accuracy under fixed RoPE base $\theta = 1\text{k}$, visualized as polar fan heatmaps.

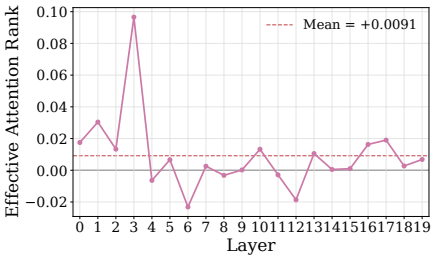


Figure 4: Layer-wise effective attention-rank difference between HD-RoPE and 2D RoPE.

training tokens. As shown in Table 5, HD-RoPE maintained its performance advantage.

4.7 Mechanistic Evidence for HD-RoPE

To better understand why HD-RoPE improves long-context behavior, we analyze both the internal attention structure and the external retrieval pattern. Figure 3 provides a complementary task-level view. Under fixed RoPE base $\theta = 1\text{k}$, the radial coordinate represents sequence length and the polar angle denotes the relative position of the answer. Compared with 2D RoPE, higher-dimensional rotations expand the high-accuracy region toward the middle part of the sequence, where retrieval is typically more difficult. Figure 4 shows the layer-wise effective attention-rank difference between 4D HD-RoPE and 2D RoPE. The mean difference is positive, suggesting that HD-RoPE tends to produce less collapsed and more diverse attention patterns.

5 Related Works

Position modeling in Transformers is generally categorized into three types: absolute position encoding, relative position encoding, and explicit biases. Classical Transformers employ sinusoidal absolute method (Vaswani et al., 2017); subsequent work shifted towards direct forms, such as introducing relative distance representations (Shaw et al., 2018) or sharing bias parameters across different distance (Raffel et al., 2020). ALiBi (Press et al., 2021) demonstrates length extrapolation capabilities through negative biases.

Rotary Position Embedding (RoPE; Su et al., 2024) makes the attention inner product explicitly dependent on the relative displacement. Recent works is mainly *recalibrating the position index or frequency*. Position Interpolation (PI) (Chen et al., 2023) reduces the phase curse by linearly compressing the position index. YaRN (Peng et al., 2023) further proposes a more efficient RoPE windowing scheme. LongRoPE combines interpolation search, extending the available context(Ding et al., 2024). Other works have analyzed the extrapolation patterns of RoPE from periodicity and scale (Liu et al., 2023), or provided a explanation (Zhong et al., 2025) of the mechanism of RoPE extension methods.

Recent works rethinks positional encoding based on representation morphology. FoPE (Hua et al., 2024) interprets RoPE as an implicit discrete Fourier transform. PoPE (Gopalakrishnan et al., 2025) explicitly separates content and position in polar coordinates. DroPE (Gelberg et al., 2025) removes positional embeddings after pre-training.

6 Conclusion

We propose HD-RoPE as a novel position embedding, focusing on the rotational degrees of freedom itself. By extending the two-dimensional rotation of RoPE to higher-dimensional subspaces and using structured orthogonal transformations to achieve cross-dimensional mixing, the expressiveness of position representation is greatly improved while preserving the relative-position properties.

Limitations

HD-RoPE focuses on enhancing RoPE by replacing isolated 2D rotations with higher-dimensional rotation subspaces. This design increases the channel capacity and mixing density of each positional phase, but it also reduces the number of independent frequency groups under a fixed hidden dimension. Therefore, HD-RoPE introduces a trade-off between fine-grained frequency diversity and richer within-frequency channel interaction. In addition, our current implementation uses fixed Paley-I conference matrices, whose construction only exists for specific dimensions and may not be universally optimal. Future work may investigate more flexible orthogonal bases, adaptive mixing strategies, and a deeper theoretical understanding of the optimal rotation dimension.

References

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2025. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Ting-Rui Chiang and Dani Yogatama. 2025. The rotary position embedding may cause dimension inefficiency in attention heads for long-distance retrieval. *arXiv preprint arXiv:2502.11276*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pages 2924–2936.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

Yoav Gelberg, Koshi Eguchi, Takuya Akiba, and Edoardo Cetin. 2025. Extending the context of pretrained llms by dropping their positional embeddings. *arXiv preprint arXiv:2512.12167*.

Anand Gopalakrishnan, Robert Csordás, Jürgen Schmidhuber, and Michael C Mozer. 2025. Decoupling the" what" and" where" with polar coordinate positional embeddings. *arXiv preprint arXiv:2509.10534*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Youbang Sun, Yuchen Fan, Xuekai Zhu, Biqing Qi, Ning Ding, and Bowen Zhou. 2024. Fourier position embedding: Enhancing attention's periodic extension for length generalization. *arXiv preprint arXiv:2412.17739*.

Haiping Liu and Hongpeng Zhou. 2025. Rethinking rope: A mathematical blueprint for n-dimensional positional encoding. *arXiv preprint arXiv:2504.06308*.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. 2023. Scaling laws of rope-based extrapolation. *arXiv preprint arXiv:2310.05209*.

Xindian Ma, Wenyuan Liu, Peng Zhang, and Nan Xu. 2025. 3d-rpe: Enhancing long-context modeling through 3d rotary position encoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24804–24811.

Sophie Ostmeier, Brian Axelrod, Michael E Moseley, Akshay Chaudhari, and Curtis Langlotz. 2024. Liere: Generalizing rotary position encodings. *arXiv preprint arXiv:2406.10322*, 2(4).

R. Paley. 1933. On orthogonal matrices. *Journal of Mathematics and Physics*, 12:311–320.

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

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

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

Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, and 1 others. 2023. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

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

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. 2024a. Ada-leval: Evaluating long-context llms with length-adaptable benchmarks. *arXiv preprint arXiv:2404.06480*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Bowen Yang, Bharat Venkitesh, Dwarak Talupuru, Hangyu Lin, David Cairuz, Phil Blunsom, and Acyr Locatelli. 2025. Rope to nope and back again: A new hybrid attention strategy. *arXiv preprint arXiv:2501.18795*.

Yupu Yao and Bowen Yang. 2025. Geope: A unified geometric positional embedding for structured tensors. *arXiv preprint arXiv:2512.04963*.

Hao Yu, Tangyu Jiang, Shuning Jia, Shannan Yan, Shunning Liu, Haolong Qian, Guanghao Li, Shuting Dong, and Chun Yuan. 2025. Comrope: Scalable and robust rotary position embedding parameterized by trainable commuting angle matrices. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4508–4517.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800.

Meizhi Zhong, Chen Zhang, Yikun Lei, Xikai Liu, Yan Gao, Yao Hu, Kehai Chen, and Min Zhang. 2025. Understanding the rope extensions of long-context llms: An attention perspective. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8955–8962.

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, and Conghui He. 2025. Meta-rater: A multi-dimensional data selection method for pre-training language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10856–10896.

A Appendix

A.1 Supplement to Related Works

Beyond the windowing approach based on frequency recalibration, recent research has begun to reconstruct the feasible domain and generalizability of RoPE from an algebraic perspective. LieRE (Ostmeier et al., 2024) extends rotation encoding to a more general n-dimensional input domain and validates the effectiveness of relative position rotation in non-textual modalities in 2D/3D vision tasks. (Liu and Zhou, 2025) further presents a unified mathematical blueprint for RoPE, emphasizing constraints such as relativity on high-dimensional constructions and pointing out that cross-dimensional interactions can be characterized by learning/selecting appropriate orthogonal basis transformations. Complementary to this, (Yu et al., 2025) starts from the RoPE equations, parameterizing rotations into trainable angle matrices and emphasizing the crucial role of commutativity in maintaining translation consistency and scalability. These works collectively demonstrate that the generalization of RoPE is not merely an engineering technique, but a problem strongly constrained by algebraic structures. The HD-RoPE in this paper employs deterministic structured orthogonal hybridization to achieve high-dimensional deep rotation, enhancing cross-dimensional coupling while maintaining the relative displacement closure.

Regarding the specific forms of high-dimensional rotation, 3D-RPE (Ma et al., 2025) elevates rotation from a two-dimensional plane to a three-dimensional spherical parameterization, improving long-range attenuation control and position resolution. Meanwhile, GeoPE (Yao and Yang, 2025) roposes a geometrically coupled rotation based on quaternions and Lie algebra means for structured tensors (2D/3D), emphasizing cross-axis joint geometric consistency. In terms of vision and multimodal applications, (Heo et al., 2024) systematically evaluates the implementation

details and resolution extrapolation performance of RoPE in ViT. Qwen2-VL (Wang et al., 2024b) introduces M-RoPE to unify the fusion of position signals from text, images, and videos, supporting dynamic resolution visual tokenization. Work (Chiang and Yogatama, 2025) indicates that large-angle rotations of RoPEs in long-distance scenarios may lead to a decrease in the utilization of some dimensions, suggesting that the structure and hybridization of the rotated subspace itself affects long-range retrieval capabilities. These results further support the motivation of this paper: to improve available dimensions and cross-dimensional information interaction through higher-dimensional rotated spaces and structured orthogonal hybridization while maintaining the relative constraints of RoPEs.

A.2 Detailed Baselines and evaluations.

The major baseline is the standard 2D RoPE (Su et al., 2024). We systematically scanned the RoPE base θ in the range of $[1\text{k}, 5\text{M}]$. In the main comparison, we report representative and best-performing settings, including $\theta = 10\text{k}$ and $\theta = 500\text{k}$, which are widely used in existing LLMs. We also compare with 3D-RPE (Ma et al., 2025) and FoPE (Hua et al., 2024).

We evaluate models on four short-context benchmarks and three long-context benchmarks. First, for the main short-context evaluation, we report standard benchmark results: (1) Hellaswag (Zellers et al., 2019); (2) AI2 Reasoning Challenge (Clark et al., 2018); (3) WinoGrande (Sakaguchi et al., 2021); (4) BoolQ (Clark et al., 2019). Second, for long-context evaluation: (1) Passkey-Retrieval, measuring precise retrieval capabilities in long sequences, following the specifications of the LongRoPE (Ding et al., 2024); (2) Ada-LEval (Wang et al., 2024a), covering task types more sensitive to long dependencies, including two subtasks: stack-select and textsort; (3) LongBench-v2 (Bai et al., 2025), emphasizing long context understanding and reasoning in real-world multi-task scenarios. Note that the scores of LongBench-v2 are computed from logits, i.e. LOGIT SCORES, not accuracy. Best in bold; second best underlined.

A.3 Model and Training Settings

This appendix provides the detailed model and training configurations used in our experiments. We will open-source the training and data framework in a future version.

A.3.1 Model Architecture

We train two Llama-style decoder-only Transformer models from scratch, with sizes of 650M and 1.3B parameters. The architectural configurations are summarized in Table 6. Both models use RMSNorm, SwiGLU feed-forward networks, and Gaussian weight initialization with standard deviation 0.02.

Configuration	650M	1.3B
#Params	650M	1.3B
Hidden Dim.	1536	2048
Layers	16	20
Heads	24	32
KV Heads	6	8
Head Dim.	64	64
Normalization	RMSNorm	RMSNorm
FFN Type	SwiGLU	SwiGLU
Weight Initialization	$\mathcal{N}(0, 0.02)$	$\mathcal{N}(0, 0.02)$

Table 6: Model architecture configurations.

A.3.2 Training Configuration

The training data construction follows Section 4.1. All models are trained with the same training budget and optimization configuration for fair comparison. We use a global batch size of 64 and train each model on approximately 50B tokens. Training is conducted with bf16 mixed precision and DeepSpeed ZeRO-2.

For optimization, we use AdamW with the WarmupDecayLR scheduler implemented in DeepSpeed. The learning rate is warmed up from 1×10^{-6} to 1×10^{-4}, followed by decay according to the scheduler. All model variants share the same optimizer, scheduler, precision, batch size, and training-token budget.

All models and components are configured with a unified random seed of 42 to ensure fair comparison. Specifically, the seeds for `random`, `numpy`, and `torch` are all fixed to 42.

A.3.3 Attention Implementation

The attention computation uses PyTorch SDPA. RoPE or HD-RoPE is applied before the query and key tensors enter the attention kernel, and is therefore decoupled from the kernel implementation:

$$\begin{aligned} \mathbf{Q}, \mathbf{K} &= \mathrm{ApplyRotaryEmb}(\mathbf{Q}, \mathbf{K}), \\ \mathbf{O} &= \mathrm{SDPA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}). \end{aligned} \tag{24}$$

This implementation allows HD-RoPE to be integrated into the standard attention pipeline without

modifying the SDPA kernel.

A.3.4 Computational Overhead

We report the FLOPs overhead using the 1.3B model with sequence length 8192 and micro-batch size 1 as an example. The total model FLOPs for a single forward pass, excluding RoPE, is approximately 21.7T. Compared with the main model computation, the additional FLOPs introduced by RoPE and HD-RoPE are negligible.

Setting	Total FLOPs	Proportion vs. Model
2D RoPE	1.26G	0.0058%
4D HD-RoPE	2.94G	0.0135%

Table 7: Additional FLOPs introduced by positional rotation.

We further measure GPU memory usage under 8 GPUs, sequence length 8192, and DeepSpeed ZeRO-2. As shown in Table 8, HD-RoPE introduces almost no additional memory overhead compared with standard 2D RoPE.

Setting	650M / GB	1.3B / GB
w/o PE	19.69	29.83
2D RoPE	19.71	29.87
4D HD-RoPE	19.72	29.87
2D–4D Interleave	19.71	29.87

Table 8: GPU memory usage under sequence length 8192 and ZeRO-2.

For runtime efficiency, we report the relative change in step time and throughput for the 1.3B model. HD-RoPE only introduces minor latency overhead, while maintaining practical training efficiency.

Machine Setting	Step Time	Throughput
1 Machine / 8 GPUs	+2.58%	-2.54%
4 Machines / 32 GPUs	+1.23%	-1.10%

Table 9: Runtime overhead of HD-RoPE on the 1.3B model.

Overall, HD-RoPE introduces negligible FLOPs and memory overhead, and only minor runtime latency, making it practical for standard large-scale pre-training settings.

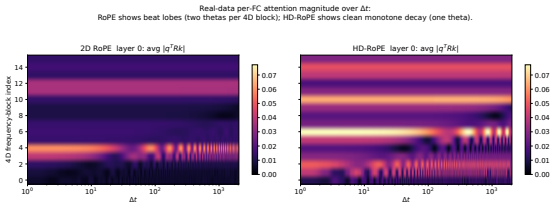


Figure 5: FC Attention Magnitude of RoPE and HD-RoPE, layer 0.

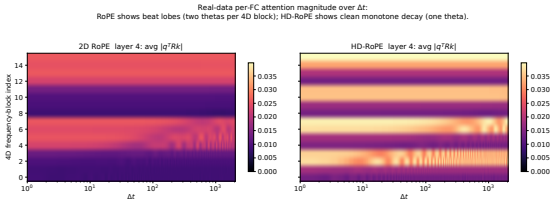


Figure 6: FC Attention Magnitude of RoPE and HD-RoPE, layer 4.

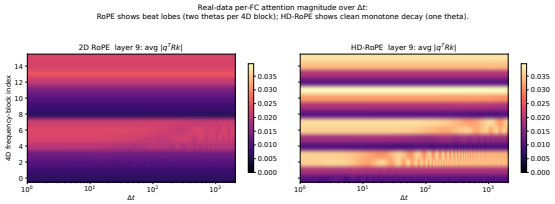


Figure 7: FC Attention Magnitude of RoPE and HD-RoPE, layer 9.

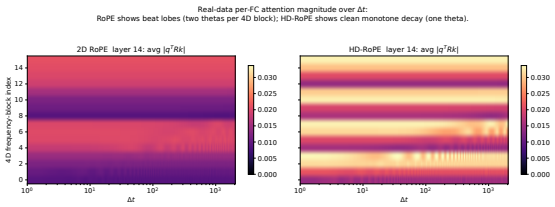


Figure 8: FC Attention Magnitude of RoPE and HD-RoPE, layer 14.

A.4 FC Attention Magnitude of RoPE and HD-RoPE

We also attempted to analyze how HD-RoPE influences attention. Shown below is the amplitude plot following Fourier decomposition.

A.5 Structured Orthogonal Families for Reproducible Deep Mixing

A central component of HD-RoPE is a fixed orthogonal mixing transform that re-parameterizes

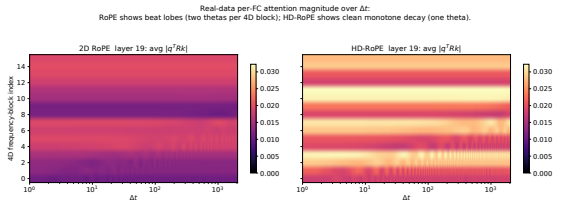


Figure 9: FC Attention Magnitude of RoPE and HD-RoPE, layer 19.

the rotation subspace without altering the norm or the relative-displacement form inherited from RoPE. While the classical literature offers many orthogonal constructions (e.g., Haar-random orthogonal matrices or QR/Gram–Schmidt generation), such approaches are typically either *random* or *numerically decomposed*, which complicates strict reproducibility across platforms and introduces non-negligible overhead when the transform must be instantiated repeatedly (e.g., across layers/heads or for multiple target dimensions).

Design requirements. We restrict attention to a class of orthogonal transforms $\mathbf{Q} \in \mathbb{R}^{N\times N}$ that satisfy, simultaneously:

1. **Determinism and reproducibility:** $\mathbf{Q}$ is generated by a deterministic rule given $(N, \text{seed-free})$, avoiding randomness and floating-point decompositions whenever possible.

2. **Numerical stability:** $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$ (exactly or up to a controlled machine-precision error), ensuring $\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for all $\mathbf{x}$.

3. **Scalability:** the construction extends to many target orders N (ideally an infinite family), or can be composed to cover arbitrary N with minimal artifacts.

4. **Low overhead:** applying $\mathbf{Q}$ admits either a fast algorithm (sub-quadratic) or a highly structured implementation (e.g., sign flips and additions) with small constant factors.

5. **Strong cross-dimensional mixing:** $\mathbf{Q}$ should distribute each coordinate across many dimensions, rather than acting as a near-permutation. A convenient quantitative proxy is the *incoherence*

$$\mu(\mathbf{Q}) \triangleq \sqrt{N}\max_{i,j}|\mathbf{Q}_{ij}|, \tag{25}$$

where smaller $\mu(\mathbf{Q})$ indicates more uniform energy spreading from standard basis vectors to mixed coordinates.

Why common orthogonal constructions are insufficient under these constraints. Under the above combined constraints, the space of viable orthogonal families becomes notably limited.

(1) Haar-random orthogonal and QR/Gram–Schmidt generation. Haar-random orthogonal matrices achieve near-optimal isotropy and typically small incoherence with high probability. However, they rely on randomness and floating-point factorization (e.g., QR), which (i) weakens strict determinism across implementations, and (ii) incurs overhead for generation (typically $\mathcal{O}(N^3)$) that is undesirable when N varies across experimental settings or must be instantiated repeatedly. Moreover, numerical decomposition can introduce subtle platform-dependent discrepancies, which is undesirable for reproducible ablation and deployment.

(2) Hadamard/Schur-type matrices. Hadamard matrices (after normalization) are deterministic and perfectly orthogonal, with entries $\pm 1/\sqrt{N}$ and thus $\mu(\mathbf{Q}) = 1$, which is near-optimal. They also admit fast transforms (Walsh–Hadamard) with $\mathcal{O}(N \log N)$ complexity. Nevertheless, their existence is restricted to specific orders (classically $N = 1, 2$ or $N \equiv 0 \pmod 4$, with nontrivial existence gaps in general), limiting their direct applicability when one requires a construction at a prescribed N. In addition, many Hadamard-based transforms are "too regular" for certain mixing patterns we target (e.g., enforcing a zero diagonal or other structural constraints), which motivates considering broader design families.

(3) Discrete cosine/sine transforms (DCT/DST). DCT/DST are deterministic, orthogonal, and have mature fast algorithms. However, these transforms are intrinsically *harmonic bases* with strong frequency ordering and boundary-condition structure: the mixing they induce is highly anisotropic and biased toward sinusoidal modes. Consequently, while they are excellent for frequency-domain reparameterization, they do not necessarily realize the kind of near-uniform cross-dimensional coupling desirable for "deep rotation" in positional encoding, especially when we aim to avoid injecting additional inductive biases unrelated to position.

(4) Sparse orthogonal parameterizations via Givens rotations. Products of Givens rotations yield a flexible class of orthogonal transforms. Yet,

achieving strong mixing (e.g., small incoherence and broad support) typically requires a sufficiently deep rotation network, increasing both implementation complexity and runtime cost. If the rotation angles or rotation pairs are randomized, the determinism requirement is violated; if they are fixed deterministically and kept shallow for efficiency, the resulting transform may remain too sparse/structured and fail to provide the desired mixing strength.

Implication. The above discussion indicates that, when we simultaneously require *(i)* deterministic, stable, reproducible generation at prescribed dimensions, *(ii)* low overhead, and *(iii)* strong cross-dimensional mixing, the set of suitable structured orthogonal families is comparatively narrow. This motivates our choice of a highly structured design family: conference-matrix-derived orthogonal transforms provide a deterministic, discrete, and strongly mixing construction, and are particularly convenient when used in small block sizes (e.g., $N = 4, 8, 32$), where the $\mathcal{O}(N^2)$ mixing cost is negligible relative to the attention/MLP computation while still yielding substantial representational gains.

A.6 Additional Background on Conference Matrices

A.6.1 Definition and basic constraints

A (real) conference matrix of order N is a matrix $\mathbf{C}_N \in \{0, \pm 1\}^{N \times N}$ with $\mathrm{diag}(\mathbf{C}_N) = \mathbf{0}$ and

$$\mathbf{C}_N^\top \mathbf{C}_N = (N-1)\mathbf{I}. \tag{26}$$

Equation (26) implies that the N rows of $\mathbf{C}_N$ form an orthogonal system with equal squared norm $(N-1)$. A direct consequence is that N must be even (for $N > 1$), hence conference matrices cannot exist for arbitrary orders. Moreover, after normalization $\mathbf{Q}_N = \mathbf{C}_N/\sqrt{N-1}$, one obtains an orthogonal transform $\mathbf{Q}_N^\top \mathbf{Q}_N = \mathbf{I}$ used in our deep-mixing construction.

A.6.2 Symmetric vs. skew-symmetric types and congruence conditions

Real conference matrices (up to equivalence operations such as negating rows/columns) fall into two structured types governed by modular constraints on N:

- **Oddly even order ($N \equiv 2 \bmod 4$):** the core submatrix is symmetric, yielding a *symmetric* conference matrix.
- **Evenly even order ($N \equiv 0 \bmod 4$):** the core submatrix is skew-symmetric, yielding a *skew-symmetric* conference matrix.

In particular, skew-symmetric conference matrices can only occur at orders N divisible by 4. These congruence constraints already rule out infinitely many dimensions.

A.6.3 Paley-type constructions and admissible orders

A widely used deterministic infinite family is given by Paley-type constructions based on finite fields. Let q be an odd prime power and set $N = q + 1$. Then:

- If $q \equiv 1 \pmod 4$, Paley graphs yield *symmetric* conference matrices of order $N = q + 1$.
- If $q \equiv 3 \pmod 4$, Paley digraphs (Paley tournaments) yield *skew-symmetric* conference matrices of order $N = q + 1$, hence $N \equiv 0 \pmod 4$.

Therefore, Paley-type constructions do *not* exist for arbitrary N: they only realize orders of the form $N = q + 1$ with q an odd prime power (and with the additional residue condition selecting the symmetric vs. skew-symmetric type).

Examples used in this paper. Our typical block sizes $N \in \{4, 8, 32\}$ correspond to $q \in \{3, 7, 31\}$, all of which are prime powers with $q \equiv 3 \pmod 4$. Hence, a skew-symmetric Paley-type conference matrix $\mathbf{C}_N$ exists for each such N, allowing deterministic and non-random generation of $\mathbf{Q}_N = \mathbf{C}_N/\sqrt{N-1}$.

A.6.4 Why the existence problem is nontrivial beyond Paley orders

The defining constraint $\mathbf{C}_N^\top \mathbf{C}_N = (N-1)\mathbf{I}$ is highly rigid: it requires N sign-vectors with exactly one zero per row/column and pairwise orthogonality. Even when the congruence conditions above are satisfied, existence is not guaranteed for all admissible N, and the set of known orders is incomplete; Paley-type constructions cover only a subset. In addition, further arithmetic necessary conditions are known (e.g., constraints on $N - 1$), highlighting that the existence question is governed by deep combinatorial-design and number-theoretic structure.

A.6.5 Practical handling for arbitrary dimensions

Since conference matrices do not exist for every N, our implementation supports arbitrary target dimen-

sions by composing available blocks: we construct $\mathbf{Q} = \mathrm{diag}(\mathbf{Q}_{N_1}, \ldots, \mathbf{Q}_{N_t}, \mathbf{I})$ using admissible orders $\{N_i\}$ (e.g., Paley-type $4, 8, 32, \ldots$), and pad the remaining dimensions with an identity block. This preserves orthogonality and enables strong mixing within conference blocks while maintaining deterministic reproducibility.

A.7 Passkey Retrieval Dataset Construction

We construct a synthetic *Passkey Retrieval* benchmark that matches the LongRoPE-style evaluation setting, with the goal of isolating long-context *needle-in-a-haystack* retrieval under precise token-length control. Each instance is a JSONL record with fields `{question, answer}`, where the `question` contains a short passkey statement embedded within a large amount of irrelevant text, and the `answer` is the passkey string.

Tokenizer-consistent length measurement. All lengths and constraints are defined in *token space* using the GPT-2 tokenizer (`use_fast=True`) with `add_special_tokens=False`. This ensures that (i) context length targets correspond to the model's effective sequence length, and (ii) dataset construction is reproducible and comparable across runs. We denote by $\tau(\cdot)$ the tokenization operator and by $|\tau(\cdot)|$ the resulting token count.

Prompt template and controllable key position. Each sample is generated from four fixed text components: (i) a short instruction header H, (ii) a filler sentence F repeated many times, (iii) a key block $K(p)$ that explicitly states the passkey p (repeated to reduce ambiguity), and (iv) a query suffix Q that asks for the passkey. Concretely, for integers $x, y \geq 1$, we form

$$\texttt{question}(x, y, p) \;=\; H \parallel F^x \parallel K(p) \parallel F^y \parallel Q, \tag{27}$$

where $\parallel$ denotes string concatenation. The pair (x, y) controls the passkey's location: x determines the distance from the beginning (prefix filler), and y determines the distance from the end (suffix filler before the query). This design decouples *context length* from *key position*, enabling systematic coverage of retrieval difficulty patterns.

Single-token passkey pool. To make evaluation unambiguous at the token level, passkeys are restricted to digit strings that tokenize into a *single token* under GPT-2 tokenization. By default, we enforce that the *leading-space context* (`" "` `+ digits`) is a single token, matching the typical in-context occurrence after natural language. Optionally, we can additionally require the no-space form (`digits`) to be single-token. We scan a user-specified integer range $[v_{\min}, v_{\max}]$, collect all digit strings satisfying the constraint, and cache the resulting pool (with metadata such as range, tokenizer name, and constraints) to ensure reproducibility and to avoid repeated scanning. During generation, each instance samples p uniformly from this pool with a fixed RNG seed.

Length bucketing and target selection. To evaluate over a wide context range while maintaining exact token-based control, we partition samples into token-length buckets of the form $(L, U]$, where U is an upper bound (default width 256 tokens). For very long contexts, we optionally use an *adaptive stride* to reduce the number of buckets while preserving coverage at short lengths. For each bucket, we select a small set of target lengths $\{T^{(j)}\}$ within $(L, U]$ (the number of targets decreases as U grows) and search for an integer sum $s = x + y$ such that the realized length

$$T(x, y, p) = |\tau(\texttt{question}(x, y, p))| \tag{28}$$

falls inside the bucket. Because tokenization is not perfectly linear in characters, we compute T directly by tokenizing a probe instance and search over s in a small neighborhood around a linear estimate.

Coverage over (x, y) without exhaustive enumeration. Enumerating all (x, y) pairs is unnecessary and expensive. Instead, for each bucket and each feasible s, we sample x values using an integer linspace over $[1, s-1]$ with a bucket-dependent number of steps. For each sampled x, we set $y = s - x$. This yields monotonic coverage from early-key (small x, large y) to late-key (large x, small y) placements at approximately fixed total length, providing broad coverage of retrieval positions while keeping dataset size bounded.

Output organization and size budgeting. Each bucket has its own directory named by the bucket upper bound U. Within a bucket, each (x, y) setting is written to a dedicated JSONL file whose filename encodes (x, y), their approximate token contributions $(x \cdot |\tau(F)|,\ y \cdot |\tau(F)|)$, and the realized total length T. To bound storage and generation cost, we enforce both (i) a global byte budget (default $\sim$1 GiB) and (ii) per-bucket byte caps

that *decrease* with U, allocating more samples to shorter contexts and fewer to extremely long contexts. We conservatively estimate bytes per line using a probe passkey with maximal digit length. Generation stops early once either the global or per-bucket budget is exhausted.

Reproducibility and provenance. We always emit a `dataset_meta.json` file recording the tokenizer, bucket configuration, passkey constraints and range, filler/base token counts, RNG seed, and byte budgets. Optionally, a `summary.json` aggregates per-bucket counts (files, lines, bytes) and stop reasons. A `-dry_run` mode performs the complete planning and budget accounting without writing JSONL files, enabling deterministic capacity planning before committing storage.

Overall, this construction follows three principles: (i) *token-faithful* control of sequence length and key position; (ii) *evaluation clarity* via single-token answers; and (iii) *scalable coverage* across long contexts through adaptive bucketing and structured sampling over (x, y) under strict size budgets.