Functional Equivalence in Attention: A Comprehensive Study with Applications to Linear Mode Connectivity

Anonymous authors
Paper under double-blind review

Abstract

The parameter space of neural networks serves as a surrogate for the underlying function class; however, the mapping is inherently non-injective, as revealed by functional equivalence, wherein distinct parameter configurations yield identical input-output behaviors. While this phenomenon has been analyzed in classical architectures such as fully connected and convolutional networks, the increasing complexity of modern designs, particularly attention-based models, presents new and significant challenges. Prior analyses of multihead attention have been largely restricted to the vanilla formulation, thereby neglecting crucial components such as positional encodings that fundamentally alter architectural symmetries and render earlier results inapplicable. In this work, we undertake a formal study of functional equivalence in Transformers with positional encodings. Focusing on the two most widely used variants–sinusoidal and rotary–we demonstrate that sinusoidal encodings preserve the equivalence structure of vanilla attention, whereas rotary encodings significantly reduce the associated symmetry group, thereby enhancing expressivity. This theoretical insight offers a principled explanation for the growing prominence of RoPE in practice. Furthermore, we extend our analysis to investigate how positional encodings influence the phenomenon of linear mode connectivity (LMC). By introducing an alignment algorithm, we empirically validate the presence and variability of LMC across a wide range of Transformer configurations, datasets, and modalities, demonstrating that the type of positional encoding plays a decisive role in shaping the connectivity of solutions.

1 Introduction

The training of deep neural networks reveals a seeming paradox: despite the high dimensionality and non-convexity of the loss landscape with numerous local minima, simple optimization methods such as stochastic gradient descent (SGD) consistently discover solutions that generalize well.

(Linear) Mode Connectivity. One influential perspective on this phenomenon is offered by the concept of *mode connectivity* (MC) (Goodfellow et al., 2014; Keskar et al., 2016; Sagun et al., 2017; Venturi et al., 2019; Neyshabur et al., 2020; Tatro et al., 2020; Yunis et al., 2022; Zhou et al., 2023), which reveals that solutions discovered through independent optimization trajectories are rarely isolated; rather, they lie within extensive connected manifolds of parameters yielding comparably low loss. A particularly tractable instance of this principle is *linear mode connectivity* (LMC) (Frankle et al., 2020; Entezari et al., 2021), in which two trained models can be joined by a straight-line interpolation in parameter space that remains confined to a low-loss region. Formally, consider a model $f(\cdot : \theta)$ parameterized by θ, with loss function $\mathcal{L}(\theta) \geq 0$. Optimization amounts to minimizing $\mathcal{L}(\theta)$ over Θ. Two solutions $\theta_A, \theta_B \in \Theta$ are said to exhibit LMC when the associated *loss barrier* (Fran-

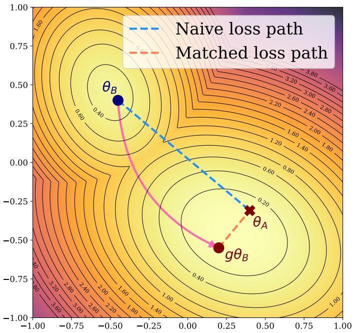


Figure 1: Illustration of Linear Mode Connectivity (up to symmetry)

kle et al., 2020; Entezari et al., 2021) vanishes (or is negligible):

$$B(\theta_A, \theta_B) := \sup_{t\in[0,1]} \left[\mathcal{L}(t\theta_A + (1-t)\theta_B) - (t\mathcal{L}(\theta_A) + (1-t)\mathcal{L}(\theta_B))\right] \approx 0. \quad (1)$$

Empirical investigations have revealed that independently trained networks on small datasets are often connected by low-loss paths (Freeman & Bruna, 2016; Garipov et al., 2018; Draxler et al., 2018), and even that nearly arbitrary pairs of solutions can be joined through curves of low error (Garipov et al., 2018). MC sheds light on the effectiveness of weight-space ensembling, known to improve generalization (Izmailov et al., 2018; Rame et al., 2022; Wortsman et al., 2022) and has been applied to adversarial robustness (Zhao et al., 2020), generalization theory (Pittorino et al., 2022; Juneja et al., 2022; Lubana et al., 2023), loss landscape geometry (Gotmare et al., 2018; Vlaar & Frankle, 2022; Lucas et al., 2021), and more recently, continual learning (Wen et al., 2023; Kozal et al., 2024; Chen et al., 2023a) and ensemble methods (Kanoh & Sugiyama, 2024; Kim et al., 2025).

Attention Mechanism and Positional Encoding. The attention mechanism is inherently permutation invariant, necessitating positional encoding (PE) to capture token order (Vaswani et al., 2017). Early models employed Absolute PEs (APEs), either sinusoidal or learnable embeddings (Gehring et al., 2017), which became standard in seminal architectures such as BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), and ViT (Dosovitskiy et al., 2020). While effective, APEs treat absolute positions as the sole signal, limiting robustness under local reordering. Relative PEs (RPEs) address this by encoding pairwise distances into attention weights (Shaw et al., 2018), a design later adopted in many models (Dai et al., 2019; He et al., 2020; Raffel et al., 2020). Among recent advances, Rotary PE (RoPE) (Su et al., 2024) encodes relative position via angular rotations of query–key vectors, preserving dot-product structure and enabling both translation equivariance and long-sequence extrapolation. RoPE is now widely adopted in state-of-the-art models (Touvron et al., 2023; Chowdhery et al., 2023; Nijkamp et al., 2022; Liu et al., 2024; Guo et al., 2025; Agarwal et al., 2025; Bai et al., 2025; Yang et al., 2025), attesting to its robustness in large-scale settings.

Functional Equivalence. A major difficulty in characterizing LMC lies in the *permutation invariance* of neural networks: reordering hidden units does not alter the underlying function (Brea et al., 2019; Novak et al., 2018), yet such symmetries can cause functionally identical models to appear distant in parameter space (Allen-Zhu et al., 2019; Du et al., 2019; Frankle & Carbin, 2018; Belkin et al., 2019; Neyshabur et al., 2018). This phenomenon is subsumed under the broader framework of *functional equivalence* (Hecht-Nielsen, 1990; Fefferman & Markel, 1993; Kurkova & Kainen, 1994; Albertini & Sontag, 1993b;a), which seeks to describe when distinct parameterizations realize the same input–output mapping. To address this issue, recent studies have examined LMC *up to permutation*, where low-loss paths are revealed once hidden units are properly aligned (Singh & Jaggi, 2020; Ainsworth et al., 2022; Peña et al., 2023; Ito et al., 2024; 2025; Zhao et al., 2025). Theoretical results show that dropout-stable networks naturally exhibit mode connectivity (Kuditipudi et al., 2019; Shevchenko & Mondelli, 2020), while LMC under permutation alignment may already emerge at initialization in the NTK regime (Entezari et al., 2021; Jacot et al., 2018), with rigorous guarantees recently established (Ferbach et al., 2024). These developments lend support to the convexity conjecture (Entezari et al., 2021), which views the SGD solution set as approximately convex once symmetries are accounted for. This view is strengthened by Sharma et al. (2024), who propose simultaneous linear connectivity, where a single model aligns linearly with multiple others. Additional studies explore the geometry of the solution space (Ainsworth et al., 2022; Xiao et al., 2023) and identify star-shaped regions conducive to LMC (Sonthalia et al., 2024).

Alignment Algorithms. These algorithms align parameters to establish LMC. (Entezari et al., 2021) proposed a simulated annealing-based algorithm. Singh & Jaggi (2020) employed Optimal Transport, while Akash et al. (2022) utilized the Wasserstein Barycenter. Ainsworth et al. (2022) introduced three methods: activation matching (using intermediate activations), weight matching (being data-independent), and the Straight-Through Estimator (minimizing interpolation loss via gradients); all are based on solving the Linear Assignment Problem (Kuhn, 1955; Jonker & Volgenant, 1988; Crouse, 2016). Peña et al. (2023) developed Sinkhorn re-basin, a differentiable method that improves alignment but struggles with residual connections due to layer-independent optimization.

Contribution. Recent work on the symmetry of vanilla attention (Tran et al., 2025; Knyazev et al., 2024) shows that head permutations and linear group actions capture all symmetries. Meanwhile, Theus et al. (2025) proposed a Transformer matching method, but it overlooks symmetry in the query-key and key-value components. In this paper, we study LMC in attention-based models, focusing on how PEs influence functional equivalence. The paper is organized as follows:

1. In Section 2, we recall the notion of Multihead Attention and its parameter space, together with the result characterizing functional equivalence in the vanilla case.
2. In Section 3, we analyze how positional encodings alter the internal structure of attention. We focus primarily on the most widely used encodings, Absolute PE and Relative PE. In particular, we study sinusoidal PE as a representative of APE and rotary PE as a representative of RPE, and show why results from the vanilla case do not extend directly to these settings.
3. In Section 4, we present the main result of the paper, which characterizes the full symmetry of attention with widely used positional encodings. This characterization underlies the matching algorithm for Multihead Attention described in Section 5.
4. In Section 6, we present empirical evidence of LMC across a wide range of models and tasks, under diverse settings and across datasets of varying scales and modalities. We also evaluate the effectiveness of our proposed matching algorithms and conduct detailed ablation studies in Section 6.2 to validate their individual components.

Theoretical foundations and experimental details are provided in the Appendix. Specifically, Appendix A offers a *consolidated overview to help readers grasp the overall structure of the work.*

2 Preliminary on Parameter Space of Multihead Attention

We present the formal definition of the Multihead Attention, describe its associated parameter space, and review key results in the literature concerning its functional equivalence properties.

Multihead Attention and its Parameter Space. Let d be a positive integer denoting the token dimension, L a positive integer denoting the sequence length, and h a positive integer denoting the number of heads. Denote the space of all sequences of tokens as $\mathcal{S} := \sqcup_{L=1}^{\infty}\mathbb{R}^{L\times d}$. Given a fixed head dimension d_h, consider $W_i^Q, W_i^K, W_i^V, W_i^O \in \mathbb{R}^{d\times d_h}$ for each $i \in [h]$. For an input sequence $\mathbf{x} = (x_1, \ldots, x_L)^\top \in \mathbb{R}^{L\times d} \subset \mathcal{S}$, the Multihead Attention with h heads is defined by

$$\text{MultiHead}\Big(\mathbf{x}:\ \{W_i^Q, W_i^K, W_i^V, W_i^O\}_{i=1}^h\Big) = \sum_{i=1}^{h} \text{softmax}\left(\left(\mathbf{x}W_i^Q\right)\left(\mathbf{x}W_i^K\right)^\top\right)\cdot\left(\mathbf{x}W_i^V\right)\left(W_i^O\right)^\top. \quad (2)$$

Here, the operator softmax is applied row-wise to the similarity matrix $(\mathbf{x}W_i^Q)(\mathbf{x}W_i^K)^\top \in \mathbb{R}^{L\times L}$, yielding the *attention matrix* associated with $\mathbf{x}$. Each row of this matrix represents a probability distribution that specifies the relative contributions of all input tokens to a given output token. In typical settings, the head dimension is chosen as $d_h = d/h$. The parameters and the parameter space of the MultiHead map is thus denoted as θ and Θ, respectively, and given by

$$\theta := \left(W_i^Q, W_i^K, W_i^V, W_i^O\right)_{i=1}^h \in \Theta(d, d_h, h) := \left(\mathbb{R}^{d\times d_h}\right)^{4h}. \quad (3)$$

Symmetry Group. Define the following group

$$G_{\text{Att}}(d_h, h) := S_h \times \left(\text{GL}(d_h)\times\text{GL}(d_h)\right)^h. \quad (4)$$

This is precisely the direct product between the permutation group S_h and h copies of $\text{GL}(d_h)\times\text{GL}(d_h)$. Each $g \in G_{\text{Att}}(d_h, h)$ has the form $g := (\sigma, (U_i, V_i)_{i=1}^h)$, where $\sigma \in S_h$ and $U_i, V_i \in \text{GL}(d_h)$. The group $G_{\text{Att}}(d_h, h)$ acts naturally on the parameter space $\Theta(d, d_h, h)$, as follows:

$$g\theta := \left(W_{\sigma(i)}^Q\cdot U_i^\top, W_{\sigma(i)}^K\cdot U_i^{-1}, W_{\sigma(i)}^V\cdot V_i^\top, W_{\sigma(i)}^O\cdot V_i^{-1}\right)_{i=1}^h. \quad (5)$$

This action preserves the functionality of MultiHead maps: for $\theta \in \Theta(d, d_h, h)$ and $g \in G_{\text{Att}}(d_h, h)$, one has $\text{MultiHead}(\cdot : \theta) = \text{MultiHead}(\cdot : g\theta)$. The general linear action cancels in the matrix multiplications, while the permutation action induced by σ commutes with addition. Together, these actions determine the symmetry of the multihead attention, as stated in the following result.

Theorem 2.1 (See Tran et al. (2025)). *Consider two* MultiHead *maps with h and $\bar{h}$ heads, parameterized by $\theta = (W_i^Q, W_i^K, W_i^V, W_i^O)_{i=1}^h$ and $\bar{\theta} = (\bar{W}_i^Q, \bar{W}_i^K, \bar{W}_i^V, \bar{W}_i^O)_{i=1}^{\bar{h}}$. Assume that*

1. *All matrices $W_i^Q, W_i^K, W_i^V, W_i^O$ and $\bar{W}_i^Q, \bar{W}_i^K, \bar{W}_i^V, \bar{W}_i^O$, for all feasible i, are of rank d_h.*

2. *From θ, the matrices $\{W_i^Q(W_i^K)^\top\}_{i=1}^h$ are pairwise distinct. The same condition holds for $\bar{\theta}$.*

If the two maps are identical, then $h = \bar{h}$, and there exists $g \in G_{\text{Att}}(d_h, h)$ such that $\bar{\theta} = g\theta$.

Remark 2.2. While the theorem imposes certain assumptions on the parameters of the MultiHead maps, these conditions hold almost surely. Thus, up to a negligible subset of the parameter space (e.g., measure zero or the complement of a dense set), functional equivalence is fully characterized by the symmetry group. This type of assumption aligns with those in the literature on functional equivalence of neural architectures (Hecht-Nielsen, 1990; Fefferman & Markel, 1993; Bui Thi Mai & Lampert, 2020), and we adopt the same perspective in the statement of our result.

3 How Positional Encoding Alters Architectural Symmetry

We investigate how positional encodings (PEs) modify the internal structure of the attention mechanism. Our analysis primarily focuses on *sinusoidal encoding* and *rotary encoding*, which are two widely used PEs. These serve as representatives of the two principal paradigms of positional encoding: absolute and relative, respectively. We examine how the formulation of attention is altered under these schemes, and how the associated architectural symmetries are consequently affected.

For now, we follow the standard implementation practice of assuming that both d and d_h are even.

3.1 Absolute Positional Encoding

Sinusoidal Encoding. In Absolute PEs, let $\mathbf{p} = \{p_i\}_{i=1}^{\infty} \subset \mathbb{R}^d$ denote the sequence of positional vectors, which encodes positional information. In the case of *sinusoidal encoding* from the original Transformer (Vaswani et al., 2017), the components of $p_m \in \mathbb{R}^d$ are defined as

$$p_{m,2k} = \sin\left(\frac{m}{10000^{2k/d}}\right), \text{ and } \quad p_{m,2k+1} = \cos\left(\frac{m}{10000^{2k/d}}\right), \tag{6}$$

for $0 \leq k < d/2$. For an input sequence $\mathbf{x} \in \mathcal{S}$ of length L, i.e., $\mathbf{x} = (x_1, \ldots, x_L)^\top \in \mathbb{R}^{L \times d}$, the positional encoding is incorporated by addition, namely $\mathbf{x} + \mathbf{p} = (x_1 + p_1, \ldots, x_L + p_L)^\top$ (this is an abuse of notation), which is then supplied as input to the multihead attention, yielding

$$\text{MultiHead}_{\text{SinusoidalPE}}\Big(\mathbf{x} : \theta\Big) = \text{MultiHead}\Big(\mathbf{x} + \mathbf{p} : \theta\Big). \tag{7}$$

Symmetry Group. In this formulation, PE does not alter the internal structure of the MultiHead map as in Equation (2); it merely applies a shift to the input. Moreover, the encoding map $\mathcal{S} \to \mathcal{S}$, defined by $\mathbf{x} \mapsto \mathbf{x} + \mathbf{p}$, is bijective. Consequently, the introduction of sinusoidal PE has no effect on the analysis of functional equivalence for multihead attention. Thus, the functional equivalence classes in the presence of sinusoidal PE coincide exactly with those in the absence of PE.

3.2 Relative Positional Encoding

Rotary Positional Encoding. We next recall the *Rotary Positional Encoding* (RoPE) (Su et al., 2024). For a token at position n, define the block-diagonal rotation matrix $R_n \in \mathbb{R}^{d_h \times d_h}$ by

$$R_n = \text{diag}\left(\begin{bmatrix}\cos(n\varphi_1) & -\sin(n\varphi_1)\\ \sin(n\varphi_1) & \cos(n\varphi_1)\end{bmatrix}, \ldots, \begin{bmatrix}\cos(n\varphi_{d_h/2}) & -\sin(n\varphi_{d_h/2})\\ \sin(n\varphi_{d_h/2}) & \cos(n\varphi_{d_h/2})\end{bmatrix}\right), \tag{8}$$

where $\varphi_i = 10000^{-2(i-1)/d}$ for $i \in [d_h/2]$. For brevity, we omit the explicit subscript indicating the head dimension d_h. Note that $R_n = (R_1)^n$. The multihead attention with RoPE is defined as s **Effect of RoPE on Internal Structure and Symmetry Group.** The parameterization and parameter space of $\text{MultiHead}_{\text{RoPE}}$ coincide with those of the standard MultiHead map defined in Equation (3). However, in contrast to the vanilla case, the action of $G_{\text{Att}}(d_h, h)$ on $\Theta(d, d_h, h)$ no longer preserves functionality. Specifically, for $\theta \in \Theta(d, d_h, h)$ and $g \in G_{\text{Att}}(d_h, h)$, one generally has

$$\text{MultiHead}_{\text{RoPE}}(\cdot : \theta) \neq \text{MultiHead}_{\text{RoPE}}(\cdot : g\theta). \tag{9}$$

The essential reason is as follows. While the interaction between W_i^V and W_i^O remains purely multiplicative and thus structurally consistent with the vanilla case, the matrices W_i^Q and W_i^K are now separated by the relative rotary matrix R_{m-n}. This insertion prevents the cancellation of group actions induced by $\mathrm{GL}(d_h)$, thereby violating the invariance property.

Symmetry Group. To define the symmetry group, we first introduce, for $i \in [d_h/2]$, the matrices $P_i, J_i \in \mathbb{R}^{d_h \times d_h}$, defined as block-diagonal matrices with $d_h/2$ consecutive 2×2 diagonal blocks:

$$P_i = \mathrm{diag}\left(0, \ldots, 0, \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{i\text{-th block}}, 0, \ldots, 0\right), \quad J_i = \mathrm{diag}\left((0, \ldots, 0, \underbrace{\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}}_{i\text{-th block}}, 0, \ldots, 0\right). \tag{10}$$

Now define the following group

$$\mathrm{H}(d_h) \coloneqq \left\{ \{U = \sum_{i=1}^{d_h/2} (a_i P_i + b_i J_i) \in \mathbb{R}^{d_h \times d_h} \ : \ (a_i, b_i) \in \mathbb{R}^2 \setminus \{(0,0)\}, \ i \in [d_h/2] \right\}. \tag{11}$$

It is straightforward to verify that $\mathrm{H}(d_h)$ forms an abelian subgroup of $\mathrm{GL}(d_h)$, and is moreover isomorphic to $(\mathbb{C}^\times)^{d_h/2}$, where $\mathbb{C}^\times$ denotes the multiplicative group of nonzero complex numbers. In particular, the rotary matrices R_n are contained in $\mathrm{H}(d_h)$ for all n. We then define

$$G_{\text{RoPE}}(d_h, h) \coloneqq S_h \times (\mathrm{H}(d_h) \times \mathrm{GL}(d_h))^h. \tag{12}$$

It follows immediately that $G_{\text{RoPE}}(d_h, h)$ is a subgroup of $G_{\text{Att}}(d_h, h)$. Furthermore, the canonical action of $G_{\text{Att}}(d_h, h)$ on $\Theta(d, d_h, h)$ restricts to a well-defined group action of $G_{\text{RoPE}}(d_h, h)$ on $\Theta(d, d_h, h)$. Crucially, this restricted action preserves the functionality of the $\text{MultiHead}_{\text{RoPE}}$ map. In particular, for every $\theta \in \Theta(d, d_h, h)$ and every $g \in G_{\text{RoPE}}(d_h, h)$, one has

$$\text{MultiHead}_{\text{RoPE}}(\cdot \colon \theta) = \text{MultiHead}_{\text{RoPE}}(\cdot \colon g\theta). \tag{13}$$

The justification is as follows. Compared to the standard MultiHead map, aside from the head permutation σ and the interaction between W_i^V and W_i^O, the only structural difference lies in the interaction between W_i^Q and W_i^K. Since $\mathrm{H}(d_h)$ is abelian and R_n belongs to $\mathrm{H}(d_h)$, one obtains

$$\begin{aligned} (W_i^Q U^\top) R_n (W_i^K U^{-1})^\top &= W_i^Q U^\top R_n (U^{-1})^\top (W_i^K)^\top \\ &= W_i^Q R_n U^\top (U^{-1})^\top (W_i^K)^\top = W_i^Q R_n (W_i^K)^\top. \end{aligned} \tag{14}$$

Thus the multiplication inside the softmax of the $\text{MultiHead}_{\text{RoPE}}$ map remains invariant under G_{RoPE}.

Remark 3.1. Our main result, presented next, shows that G_{RoPE} fully characterizes the symmetry structure of the $\text{MultiHead}_{\text{RoPE}}$ map. Since $\mathrm{H}(d_h)$ is substantially smaller than $\mathrm{GL}(d_h)$, the function class represented by $\text{MultiHead}_{\text{RoPE}}$ is strictly larger than that of MultiHead or $\text{MultiHead}_{\text{SinusoidalPE}}$. *This finding offers a theoretical rationale for the increasing use of RoPE in attention-based models.*

4 Functional Equivalence of Multihead Attention with RoPE

In this section, we examine the functional equivalence of multihead attention under a general formulation, of which the RoPE-based attention mechanism constitutes a special case.

4.1 General Formulation of Multihead Attention

General Multihead Attention. Consider a Multihead map with h heads, parameterized by two families of matrices $\{\{A_i^{m,n}\}_{m,n\geq 1}\}_{i=1}^h$ and $\{B_i\}_{i=1}^h$, where each $A_i^{m,n}, B_i \in \mathbb{R}^{d \times d}$. For an input sequence $\mathbf{x} = (x_1, \ldots, x_L)^\top \in \mathbb{R}^{L \times d}$, the general Multihead map is defined as

$$\text{MultiHead}\Big(\mathbf{x} : \{\{A_i^{m,n}\}_{m,n}, B_i\}_{i=1}^h\Big) = \sum_{i=1}^h \text{softmax}\left[x_m A_i^{m,n} x_n^\top\right]_{m,n=1,\ldots,L} \cdot \mathbf{x} B_i. \tag{15}$$

To facilitate the subsequent analysis, we impose two structural conditions:

1. *(Stationarity)*: for all $m, n \geq 1$ and all shifts $k \geq 0$, we assume $A_i^{m,n} = A_i^{m+k,n+k}$, reflecting the natural shift-invariance induced by relative positional encodings; and,

2. *(Self-similarity symmetry)*: for each $m \geq 1$, $A_i^{m,m}$ parameterizes the similarity score of the m-th token with itself at head i. Since any quadratic form is uniquely represented by a symmetric matrix, we replace $A_i^{m,m}$ with its symmetrization $\text{sym}(A_i^{m,m}) := \left(A_i^{m,m} + (A_i^{m,m})^\top\right)/2$,
$$A_i^{m,m} \mapsto \text{sym}(A_i^{m,m}), \text{ which implies that } \quad x_m A_i^{m,m} x_m^\top = x_m \text{sym}(A_i^{m,m}) x_m^\top. \tag{16}$$
Henceforth, we assume that all matrices $A_i^{m,m}$ are symmetric.

From now, these two conditions will be assumed whenever the general formulation is considered.

Functional Equivalence of General MultiHead. We now consider the situation where two general MultiHead maps, one with h heads and the other with $\bar{h}$ heads, yield identical outputs:

$$\text{MultiHead}\Big(\mathbf{x}\colon \{\{A_i^{m,n}\}_{m,n}, B_i\}_{i=1}^h\Big) = \text{MultiHead}\Big(\mathbf{x}\colon \{\{\bar{A}_i^{m,n}\}_{m,n}, \bar{B}_i\}_{i=1}^{\bar{h}}\Big). \tag{17}$$

Equation (17) is equivalent to the fact that a MultiHead map with $h + \bar{h}$ heads vanishes identically:

$$0 = \text{MultiHead}\Big(\mathbf{x} : \{\{A_i^{m,n}\}_{m,n}\}_{i=1}^h \sqcup \{\{\bar{A}_i^{m,n}\}_{m,n}\}_{i=1}^{\bar{h}},\ \{B_i\}_{i=1}^h \sqcup \{-\bar{B}_i\}_{i=1}^{\bar{h}}\Big). \tag{18}$$

Before presenting our result, we introduce the following notion. Two families $\{X^{m,n}\}_{m,n\geq 1}$ and $\{Y^{m,n}\}_{m,n\geq 1}$ are *distinct* if there exists indices $m, n \geq 1$ such that $X^{m,n} \neq Y^{m,n}$. We are now in position to state the main result of this section, which provides a fundamental insight into the problem of Functional Equivalence in MultiHead Attention.

Theorem 4.1. *Consider the* MultiHead *map with h heads, parameterized by families of matrices $\{\{A_i^{m,n}\}_{m,n}\}_{i=1}^h$ and $\{B_i\}_{i=1}^h$ in $\mathbb{R}^{d\times d}$, as in Equation (15). Assume that the h parameter families $\{A_1^{m,n}\}_{m,n}, \ldots, \{A_h^{m,n}\}_{m,n}$, are pairwise distinct, and further that $A_i^{m,n}$ is nonzero for all $i \in [h]$ and $m, n \geq 1$. If the* MultiHead *map is identical to zero, then all $B_1, \ldots, B_h$ are equal to zero.*

The proof of Theorem 4.1, given in Appendix D.2, can be viewed as a statement on the linear independence of attention heads. It proceeds by rewriting the identically zero MultiHead map, after clearing denominators from the softmax, as an exponential polynomial identically equal to zero, and then applying tools from this theory. While the proof is lengthy, the functional equivalence of RoPE follows directly as a corollary, requiring only additional arguments for the rotary matrices R.

4.2 The case of Multihead Attention with RoPE

$\text{MultiHead}_{\text{RoPE}}$ is subsumed by the general formulation in Equation (15). Indeed, define

$$A_i^{m,m} := \text{sym}\left(W_i^Q (W_i^K)^\top\right), \quad A_i^{m,n} := W_i^Q R^{m-n} (W_i^K)^\top \text{ if } m \neq n, \quad B_i := W_i^V (W_i^O)^\top.$$

Then $\text{MultiHead}_{\text{RoPE}}$ is precisely a special case of the general MultiHead formulation:

$$\text{MultiHead}_{\text{RoPE}}\Big(\mathbf{x}\colon \{W_i^Q, W_i^K, W_i^V, W_i^O\}_{i=1}^h\Big) = \text{MultiHead}\Big(\mathbf{x}\colon \left\{\{A_i^{m,n}\}_{m,n}, B_i\right\}_{i=1}^h\Big). \tag{19}$$

The following result characterizes the functional equivalence of Multihead Attention with RoPE.

Theorem 4.2. *Consider two* $\text{MultiHead}_{\text{RoPE}}$ *maps with h and $\bar{h}$ heads. They are parameterized by $(W_i^Q, W_i^K, W_i^V, W_i^O)_{i=1}^h$ and $(\bar{W}_i^Q, \bar{W}_i^Q, \bar{W}_i^V, \bar{W}_i^O)_{i=1}^{\bar{h}}$ in $\mathbb{R}^{d\times d_h}$, respectively. Assume that*

1. *In the first* $\text{MultiHead}_{\text{RoPE}}$ *map, the h families below consist solely of nonzero matrices,*
$$\left\{\text{sym}\left(W_i^Q (W_i^K)^\top\right), \{W_i^Q R^n (W_i^K)^\top\}_{n\in\mathbb{Z}, n\neq 0}\right\}, \textit{ for } \ i \in [h],$$
and these constitute h pairwise distinct families. The same condition holds for the second map.

2. *All matrices $W_i^Q, W_i^K, W_i^V, W_i^O$ and $\bar{W}_i^Q, \bar{W}_i^K, \bar{W}_i^V, \bar{W}_i^O$, for all feasible i, are of rank d_h.*

If the two $\text{MultiHead}_{\text{RoPE}}$ *maps are identical, then necessarily $h = \bar{h}$. Moreover, there exists a permutation $\sigma \in S_h$, matrices $\{U_i\}_{i=1}^h \subset \mathrm{H}(d_h)$ and $\{V_i\}_{i=1}^h \subset \mathrm{GL}(d_h)$, such that*

$$\bar{W}_i^Q = W_{\sigma(i)}^Q \cdot U_i^\top, \bar{W}_i^K = W_{\sigma(i)}^K \cdot (U_i)^{-1}, \bar{W}_i^V = W_{\sigma(i)}^V \cdot V_i^\top, \bar{W}_i^O = W_{\sigma(i)}^O \cdot (V_i)^{-1}. \tag{20}$$

The proof of Theorem 4.2 is given in Appendix F. It proceeds as follows: $\text{MultiHead}_{\text{RoPE}}$ is reformulated as a special case of the general formulation (as in Equation (19)); Theorem 4.1 is then applied to derive relations between parameters; finally, a structural property of the rotary matrix, stated in Lemma F.2 of Appendix F.2, is used to recover the relationship between the original parameters.

5 Matching Algorithm for Multihead Attention Layers

As detailed in the above sections, the functionality of a Multihead Attention (MHA) is invariant under relevant group actions–G_{Att} or G_{RoPE}. To align two MHAs A and B, with their parameters denoted by $\theta^A = (W^Q_{i,A}, W^K_{i,A}, W^V_{i,A}, W^O_{i,A})_{i=1}^h$ and $\theta^B = (W^Q_{i,B}, W^K_{i,B}, W^V_{i,B}, W^O_{i,B})_{i=1}^h$, we need to find an optimal group element g that accounts for these symmetries. Inspired by the Weight Matching algorithm (Ainsworth et al., 2022), we propose a data-independent alignment method, applicable to both standard MHA and MHA with RoPE. Our method decomposes into two stages.

1. First, we match the ordering of heads in A and B by formulating the problem as a Linear Assignment Problem (LAP), solved in $O(h^3)$ time using the Hungarian algorithm (Kuhn, 1955).
2. Second, for each matched pair of heads, we find an optimal transformation from the relevant symmetry group ($\mathrm{GL}(d_h)$ or $\mathrm{H}(d_h)$) to align their internal parameters.

This staged approach separates the discrete permutation from continuous transformations, streamlining optimization. We process each stage as follows.

Stage 1 (Head Permutation Matching). Given a cost matrix $C = \{C_{i,j}\}_{i,j=1}^h \in \mathbb{R}^{h\times h}$, the goal of an LAP is to find the optimal permutation $\sigma^* \in S_h$ that maps heads from B to A by minimizing the total assignment cost: $\sigma^* = \arg\min_{\sigma\in S_h} \sum_{i=1}^h C_{i,\sigma(i)}$. To construct the cost matrix, we define:

$$M_i^A = W^Q_{i,A}(W^K_{i,A})^\top \quad \text{and} \quad N_i^A = W^V_{i,A}(W^O_{i,A})^\top, \quad \text{for } i \in [h], \tag{21}$$

where these matrices are in $\mathbb{R}^{d\times d}$. The matrices M_i^B and N_i^B are defined similarly. To capture the softmax translation-invariance, we center each row of M_i^A as $\bar{M}_i^A = M_i^A - \frac{1}{d}(M_i^A\mathbf{1})\mathbf{1}^\top$, and similarly for M_i^B, where $\mathbf{1} = [1,\dots,1]^\top \in \mathbb{R}^d$. The cost matrix $C \in \mathbb{R}^{h\times h}$ is then defined as:

$$C_{i,j} = \left\|\bar{M}_i^A - \bar{M}_j^B\right\|_F^2 + \left\|N_i^A - N_j^B\right\|_F^2, \quad \text{for } i,j \in [h]. \tag{22}$$

This ensures that the cost matrix remains invariant under group actions on W_i^Q, W_i^K or W_i^V, W_i^O.

Stage 2 (Internal Parameter Alignment). After reordering the heads of B with σ^*, we separately align the Query-Key and Value-Output components for each head. For Query-Key, we define:

$$\mathcal{L}_{Q,K}(U_i) = \|W^Q_{i,A} - W^Q_{i,B}U_i^\top\|_F^2 + \|W^K_{i,A} - W^K_{i,B}U_i^{-1}\|_F^2. \tag{23}$$

We then minimize $\mathcal{L}_{Q,K}(U_i)$ over U_i in the appropriate symmetry group.

Standard MHA. The symmetry group is $\mathrm{GL}(d_h)$. We optimize Equation (23) for $U_i \in \mathrm{GL}(d_h)$ via gradient descent, using the gradient in Lemma G.1. The optimization is initialized from the solution to a constrained version of the problem, where U_i is restricted to be orthogonal (Lemma G.2).

MHA with RoPE. The symmetry group is restricted to $U_i \in \mathrm{H}(d_h)$. This constraint decouples the problem into $d_h/2$ independent 2-dimensional subproblems, each reducible to a minimization over a scalar variable, solved efficiently using Brent's method (Brent, 2013), as shown in Lemma G.3.

For both MHA variants, we align Value-Output by finding a matrix $V_i \in \mathrm{GL}(d_h)$ that minimizes:

$$\mathcal{L}_{V,O}(V_i) = \|W_i^{V,A} - W_i^{V,B}V_i^{-1}\|_F^2 + \|W_i^{O,A} - V_iW_i^{O,B}\|_F^2. \tag{24}$$

This problem is solved using the same approach as the Query-Key alignment for standard MHA. The complete procedure is summarized in Algorithm 1.

Remark 5.1. Our experimental implementation extends the theory by incorporating biases through augmented weight matrices (e.g., $\widetilde{W_i^Q} = [W_i^Q; (b_i^Q)^\top]$). Furthermore, for the full Transformer block alignment in Experiment 6.1, we supplement our method with standard Weight Matching (Ainsworth et al., 2022) for the feed-forward networks.

Remark 5.2. To align full Transformer models, Theus et al. (2025) identified a residual-path symmetry under orthogonal group action on the embedding space, though it holds strictly for RMSNorm networks. For LayerNorm models, it requires reparameterization, thus leading to a variant of LMC. Moreover, the approach considers only Query–Key and Key–Value circuits, without addressing the essential symmetry groups of these components. This underscores the novelty of our framework.

Table 1: Experimental setups for LMC under *first attention layer* re-initialization. The table reports datasets, model depths, and head counts, with figure references showing interpolation curves for APE and RoPE variants. Notation $A \rightarrow B$ indicates pretraining on A, fine-tuning on B.

Dataset	Layers	Heads	APE	RoPE	Dataset	Layers	Heads	APE	RoPE
MNIST	1	[4, 8]	[5a, 5b]	[22a, 22b]	AGNews	2	[4, 8]	[13a, 13b]	[30a, 30b]
	2	[4, 8]	[6a, 6b]	[23a, 23b]		6	[4, 8]	[14a, 14b]	[31a, 31b]
CIFAR-10	2	[4, 8]	[7a, 7b]	[24a, 24b]	IMDB	2	[4, 8]	[15a, 15b]	[32a, 32b]
	4	[4, 8]	[8a, 8b]	[25a, 25b]		6	[4, 8]	[16a, 16b]	[33a, 33b]
	6	[4, 8]	[9a, 9b]	[26a, 26b]	DBPedia	2	[4, 8]	[17a, 17b]	[34a, 34b]
CIFAR-100	6	[4, 8]	[10a, 10b]	[27a, 27b]		6	[4, 8]	[18a, 18b]	[35a, 35b]
ImageNet-21k→CIFAR-10	12	[6]	[11a]	[28a]	Enwik8	12	[4, 8, 16]	[19a, 19b, 19c]	[36a, 36b, 36c]
ImageNet-21k→CIFAR-100	12	[6]	[11b]	[28b]	WikiText103	12	[2, 3, 4]	[20a, 20b, 20c]	[37a, 37b, 37c]
ImageNet-1k	12	[8, 12, 16]	[12a,12b,12c]	[29a,29b,29c]	One Billion Word	12	[8, 12, 16]	[21a, 21b, 21c]	[38a, 38b, 38c]

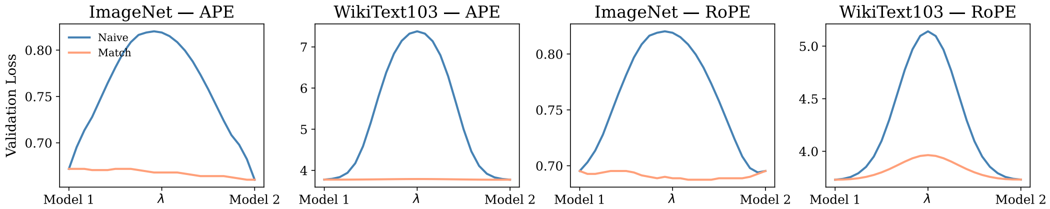


Figure 2: LMC interpolation plots for ViT on ImageNet-1K (subplots 1 and 3) and GPT-2 on WikiText103 (subplots 2 and 4), with APE and RoPE under first attention layer re-initialization.

6 Experimental Results

In this section, we study LMC in attention-based models with two types of PE: APE and RoPE. Four re-initialization strategies are considered: (i) re-initializing only the first attention layer (first attention layer), (ii) stacking re-initialized attention layers sequentially (full attention layers), (iii) re-initializing the first attention-FFN pair (first Transformer layer), and (iv) re-initializing the entire Transformer (full model), including all attention and feedforward blocks. In all cases, only the designated re-initialized parameters are fine-tuned, with others frozen. We emphasize the first layer for its central role in early representations (Appendix H). We assess LMC across three seeds by interpolating between checkpoint pairs and measuring test performance at 25 evenly spaced points.

Datasets and Models. For vision tasks, we adopt ViT (Dosovitskiy et al., 2020) on MNIST (LeCun et al., 1998), CIFAR-10/100 (Krizhevsky et al., 2009), and ImageNet-1K (Deng et al., 2009). For language modeling, we use GPT-2 (Radford et al., 2019) on Enwik8 Mahoney (2011), WikiText103 (Merity et al., 2016), and the One Billion Word benchmark (Chelba et al., 2013). For text classification, we employ BERT (Devlin et al., 2019) on AG News (Zhang et al., 2015), IMDB reviews (Maas et al., 2011), and DBPedia (Lehmann et al., 2015). All hyperparameters, including batch size, optimizer, learning rate schedules, and other configuration details, are provided in Appendix I.

6.1 Empirical Verification of Linear Mode Connectivity

We examine LMC under two extremes: (i) **first attention layer** and (iv) **full model**. Intermediate settings–(ii) full attention layers and (iii) first Transformer layer–are included in Appendix J.2 and J.3. Tables 1 and 2 summarize the experimental setups across tasks, while Figures 2 and 72 show the validation loss curves for the first attention layer and full-model re-initializations. For models using APE, our experiments reveal that LMC is evident with first-layer re-initialization, but vanishes under the full attention layer and full model configurations on *large-scale language modeling datasets* such as Enwik8, WikiText-103, and One Billion Word. With RoPE, LMC consistently appears in vision and smaller NLP tasks (AGNews, IMDB, DBPedia). For the aforementioned language modeling datasets, LMC fails to emerge even when only the first attention layer is re-initialized. This limitation likely arises from the architecture of the models used-specifically, GPT2-style decoder-only Transformers. Unlike encoder-based models, decoder-only architectures employ *causal attention*, where each token can only attend to its leftward context. This results in attention scores between tokens j and i vanishing for all $j > i$, which significantly restricts information flow during interpolation and disrupts symmetry across tokens. Consequently, LMC is less likely to emerge in such models unless stronger inductive biases or broader symmetry-aware methods are introduced.

Table 2: Experimental setups for LMC under full Transformer re-initialization. The table lists datasets, model depths, and attention head counts, along with references to figures comparing APE and RoPE. This configuration represents the most disruptive reset scenario considered in our study.

Dataset	Layers	Heads	APE	RoPE	Dataset	Layers	Heads	APE	RoPE	Dataset	Layers	Heads	APE	RoPE
CIFAR-10	6	[8]	[73a]	[73b]	AGNews	6	[8]	[75a]	[75b]	ImageNet-1k	[12]	[12]	[77a]	[77b]
CIFAR-100	6	[8]	[74a]	[74b]	DBPedia	6	[8]	[76a]	[76b]	Wikitext103	[12]	[3]	[78a]	[78b]

Table 3: Joint comparison of head permutations and ablation variants for 6-layer ViT/BERT models with 4 heads on 4 datasets and 2 PE types under first-layer attention replacement. For head permutations, we report Rank (out of 24 permutations) and $\hat{L}$ for loss and accuracy barriers, averaged over 10 checkpoint pairs. For the ablation study, we report barrier ratios (%) relative to naive interpolation: Variant 1 removes Stage 2, Variant 2 uses Stage 2 with orthogonal initialization only (no gradient descent), and Full method applies the optimization. Lower values indicate better connectivity.

Dataset	PE Type	Stage 1: Head permutation				Stage 2: Component Ablation ratios (%)					
		Rank (out of 24) ↓		$\hat{L} = \frac{L_{\text{method}} - L_{\text{top1}}}{L_{\text{naive}} - L_{\text{top1}}} \times 10^2$ ↓		Loss barrier ratio ↓			Accuracy barrier ratio ↓		
		Loss	Accuracy	Loss	Accuracy	Variant 1	Variant 2	Full	Variant 1	Variant 2	Full
CIFAR-10	APE	2.40 ± 0.54	1.94 ± 0.37	2.60 ± 0.92	2.11 ± 0.48	78.3 ± 19.4	10.2 ± 5.1	**8.7 ± 2.3**	76.5 ± 18.7	10.9 ± 4.8	**8.4 ± 2.1**
	RoPE	2.80 ± 0.65	2.01 ± 0.66	2.90 ± 0.87	2.21 ± 0.53	79.1 ± 20.2	12.5 ± 5.6	**9.2 ± 2.5**	77.8 ± 19.3	11.7 ± 5.2	**9.0 ± 2.4**
CIFAR-100	APE	3.10 ± 0.78	1.11 ± 0.38	3.00 ± 0.72	1.39 ± 0.52	74.6 ± 17.8	10.8 ± 4.3	**7.5 ± 1.9**	73.2 ± 17.1	10.4 ± 4.0	**7.2 ± 1.8**
	RoPE	2.30 ± 0.35	2.11 ± 0.77	3.10 ± 0.83	1.32 ± 0.34	75.9 ± 18.5	12.6 ± 4.7	**8.0 ± 2.1**	74.4 ± 17.9	12.1 ± 4.4	**7.8 ± 2.0**
IMDBreview	APE	4.50 ± 1.63	2.52 ± 1.31	4.70 ± 1.74	2.44 ± 1.43	91.4 ± 21.6	15.7 ± 6.2	**10.3 ± 2.8**	91.2 ± 20.9	15.3 ± 5.9	**10.1 ± 2.7**
	RoPE	4.70 ± 1.22	2.94 ± 1.46	4.80 ± 1.89	2.72 ± 1.32	88.7 ± 22.3	16.4 ± 6.5	**11.1 ± 3.0**	95.5 ± 21.7	15.9 ± 6.3	**10.8 ± 2.9**
DBPedia	APE	2.90 ± 0.91	0.59 ± 0.17	2.40 ± 0.85	0.72 ± 0.23	61.8 ± 16.4	10.9 ± 3.8	**7.1 ± 1.7**	58.5 ± 15.8	10.5 ± 3.6	**6.9 ± 1.6**
	RoPE	2.20 ± 0.44	0.62 ± 0.16	2.70 ± 0.91	0.35 ± 0.12	62.4 ± 16.9	11.3 ± 4.1	**7.4 ± 1.8**	41.1 ± 16.2	10.8 ± 3.9	**7.2 ± 1.7**

6.2 Ablation on the matching algorithm

We perform ablation studies on each component of our matching method (Section 5) using 6-layer ViT/BERT models with 4-head attention layers on CIFAR-10/100, IMDB Reviews, and DBPedia datasets, for both APE and RoPE under first layer replacement scheme.

Stage 1. We assess Stage 1 by ranking the selected head permutation among all 24 possibilities, each with Stage 2 applied after reordering. Table 3 reports the rank and scaled metric $\hat{L} = \frac{L_{\text{method}} - L_{\text{top1}}}{L_{\text{naive}} - L_{\text{top1}}} \times 10^2$, averaged over 10 checkpoint pairs from 4 checkpoints, where L_{method}, L_{top1}, and L_{naive} are the barriers for our method, the best permutation, and naive interpolation. Results show low ranks and near-zero $\hat{L}$, indicating near-optimal matching. Visualizations of LMC across all permutations (Appendix J.5) highlight the need for accurate matching, as poor permutations degrade performance.

Stage 2. To evaluate Stage 2, we ablate its components (Table 3). Variant 1, which omits Stage 2 entirely, yields high and unstable barrier ratios. Variant 2, using only the initial orthogonal alignment, substantially reduces barriers to 10–16%. Our full method, which builds upon Variant 2 by adding gradient descent fine-tuning, achieves the lowest and most stable barriers at 7-12%. This demonstrates that *both initial alignment and subsequent fine-tuning are essential* for optimal performance.

7 Conclusions, Limitations, and Future Directions

In this paper, we present a study of the functional equivalence of Multihead Attention, focusing on how positional encodings alter the symmetry of vanilla attention. We propose a weight-matching algorithm for attention parameters and use it to investigate LMC in Transformer models. Our experiments show that LMC consistently arises in encoder-only architectures but fails in decoder-only models for large-scale language modeling, reflecting a structural limitation of causal attention, in which restricted context flow disrupts token symmetry and prevents smooth interpolation. Future work may extend this analysis to other regimes, such as causal attention, or develop more efficient matching methods for studying LMC in RoPE-based attention and at larger scales, if it exists. Conversely, establishing a provable framework for the non-existence of LMC in certain model classes would be a novel contribution, advancing our understanding of the loss landscape in deep learning.

Ethics Statement. This work does not pose foreseeable risks of negative societal or ethical impact, given its technical and methodological focus.

Reproducibility Statement. We provide the complete source code for all experiments in the supplementary materials. Details of hyperparameters, training protocols, and computational infrastructure are included in Appendix I. All datasets employed in this study are publicly available and can be easily accessed online.

LLM Usage Declaration. Large language models (LLMs) were employed solely for grammar checking and minor language refinements.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Aditya Kumar Akash, Sixu Li, and Nicolás García Trillos. Wasserstein barycenter-based model fusion and linear mode connectivity of neural networks. *arXiv preprint arXiv:2210.06671*, 2022.

Francesca Albertini and Eduardo D Sontag. Identifiability of discrete-time neural networks. In *Proc. European Control Conference*, pp. 460–465. Springer Berlin, 1993a.

Francesca Albertini and Eduardo D Sontag. For neural networks, function determines form. *Neural networks*, 6(7):975–990, 1993b.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Johanni Brea, Berfin Simsek, Bernd Illing, and Wulfram Gerstner. Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape. *arXiv preprint arXiv:1907.02911*, 2019.

Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

Phuong Bui Thi Mai and Christoph Lampert. Functional vs. parametric equivalence of relu networks. In *8th International Conference on Learning Representations*, 2020.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

Qidong Chen, Jun Sun, Vasile Palade, and Zihao Yu. Continual relation extraction via linear mode connectivity and interval cross training. *Knowledge-Based Systems*, 264:110288, 2023a.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

David F Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1309–1318. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/draxler18a.html.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

Charles Fefferman and Scott Markel. Recovering a feed-forward net from its output. In *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pp. 335–342. Morgan Kaufmann, 1993. URL http://papers.nips.cc/paper/748-recovering-a-feed-forward-net-from-its-output.

Damien Ferbach, Baptiste Goujaud, Gauthier Gidel, and Aymeric Dieuleveut. Proving linear mode connectivity of neural networks via optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pp. 3853–3861. PMLR, 2024.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020.

C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pp. 1243–1252. PMLR, 2017.

Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. Using mode connectivity for loss landscape analysis. *arXiv preprint arXiv:1806.06977*, 2018.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Peter Hall. On representatives of subsets. *Journal of The London Mathematical Society-second Series*, pp. 26–30, 1935. URL https://api.semanticscholar.org/CorpusID:23252557.

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

Robert Hecht-Nielsen. On the algebraic structure of feedforward network weight spaces. In *Advanced Neural Computers*, pp. 129–135. Elsevier, 1990.

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.

Akira Ito, Masanori Yamada, and Atsutoshi Kumagai. Analysis of linear mode connectivity via permutation-based weight matching: With insights into other permutation search methods. *arXiv preprint arXiv:2402.04051*, 2024.

Akira Ito, Masanori Yamada, and Atsutoshi Kumagai. Linear mode connectivity between multiple models modulo permutation symmetries. In *Forty-second International Conference on Machine Learning*, 2025.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Roy Jonker and Ton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. In *DGOR/NSOR: Papers of the 16th Annual Meeting of DGOR in Cooperation with NSOR/Vorträge der 16. Jahrestagung der DGOR zusammen mit der NSOR*, pp. 622–622. Springer, 1988.

Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. Linear connectivity reveals generalization strategies. *arXiv preprint arXiv:2205.12411*, 2022.

Ryuichi Kanoh and Mahito Sugiyama. Linear mode connectivity in differentiable tree ensembles. *arXiv preprint arXiv:2405.14596*, 2024.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Byungjai Kim, Chanho Ahn, Wissam J Baddar, Kikyung Kim, Huijin Lee, Saehyun Ahn, Seungju Han, Sungjoo Suh, and Eunho Yang. Test-time ensemble via linear mode connectivity: A path to better adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Boris Knyazev, Abhinav Moudgil, Guillaume Lajoie, Eugene Belilovsky, and Simon Lacoste-Julien. Accelerating training with neuron interaction and nowcasting networks. *arXiv preprint arXiv:2409.04434*, 2024.

Jędrzej Kozal, Jan Wasilewski, Bartosz Krawczyk, and Michał Woźniak. Continual learning with weight interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4187–4195, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Rong Ge, and Sanjeev Arora. Explaining landscape connectivity of low-cost solutions for multilayer nets. *Advances in neural information processing systems*, 32, 2019.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Vera Kurkova and Paul C Kainen. Functionally equivalent feedforward neural networks. *Neural Computation*, 6(3):543–558, 1994.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Ekdeep Singh Lubana, Eric J Bigelow, Robert P Dick, David Krueger, and Hidenori Tanaka. Mechanistic mode connectivity. In *International Conference on Machine Learning*, pp. 22965–23004. PMLR, 2023.

James Lucas, Juhan Bae, Michael R Zhang, Stanislav Fort, Richard Zemel, and Roger Grosse. Analyzing monotonic linear interpolation in neural network loss landscapes. *arXiv preprint arXiv:2104.11044*, 2021.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Matt Mahoney. Large text compression benchmark, 2011.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

Fidel A Guerrero Peña, Heitor Rapela Medeiros, Thomas Dubail, Masih Aminbeidokhti, Eric Granger, and Marco Pedersoli. Re-basin via implicit sinkhorn differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20237–20246, 2023.

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

Fabrizio Pittorino, Antonio Ferraro, Gabriele Perugini, Christoph Feinauer, Carlo Baldassi, and Riccardo Zecchina. Deep networks on toroids: removing symmetries reveals the structure of flat regions in the landscape geometry. In *International Conference on Machine Learning*, pp. 17759–17781. PMLR, 2022.

Robert Piziak and Patrick L Odell. Full rank factorization of matrices. *Mathematics magazine*, 72 (3):193–201, 1999.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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. *Journal of machine learning research*, 21(140):1–67, 2020.

Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022.

Gian-Carlo Rota. On the foundations of combinatorial theory: I. theory of möbius functions. In *Classic Papers in Combinatorics*, pp. 332–360. Springer, 1964.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Ekansh Sharma, Devin Kwok, Tom Denton, Daniel M Roy, David Rolnick, and Gintare Karolina Dziugaite. Simultaneous linear connectivity of neural networks modulo permutation. *arXiv preprint arXiv:2404.06498*, 2024.

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

Alexander Shevchenko and Marco Mondelli. Landscape connectivity and dropout stability of sgd solutions for over-parameterized neural networks. In *International Conference on Machine Learning*, pp. 8773–8784. PMLR, 2020.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33:22045–22055, 2020.

Ankit Sonthalia, Alexander Rubinstein, Ehsan Abbasnejad, and Seong Joon Oh. Do deep neural network solutions form a star domain? *arXiv preprint arXiv:2403.07968*, 2024.

Richard P Stanley. Enumerative combinatorics volume 1 second edition. *Cambridge studies in advanced mathematics*, 2011.

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

Norman Tatro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, and Rongjie Lai. Optimizing mode connectivity via neuron alignment. *Advances in Neural Information Processing Systems*, 33:15300–15311, 2020.

Alexander Theus, Alessandro Cabodi, Sotiris Anagnostidis, Antonio Orvieto, Sidak Pal Singh, and Valentina Boeva. Generalized linear mode connectivity for transformers. *arXiv preprint arXiv:2506.22712*, 2025.

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*, 2023.

Hoang V. Tran, Thieu Vo, An Nguyen The, Tho Tran Huu, Minh-Khoi Nguyen-Nhat, Thanh Tran, Duy-Tung Pham, and Tan Minh Nguyen. Equivariant neural functional networks for transformers. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20:133, 2019.

Tiffany J Vlaar and Jonathan Frankle. What can linear interpolation of neural network loss landscapes tell us? In *International Conference on Machine Learning*, pp. 22325–22341. PMLR, 2022.

Haitao Wen, Haoyang Cheng, Heqian Qiu, Lanxiao Wang, Lili Pan, and Hongliang Li. Optimizing mode connectivity for class incremental learning. In *International Conference on Machine Learning*, pp. 36940–36957. PMLR, 2023.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998. PMLR, 2022.

Tim Z Xiao, Weiyang Liu, and Robert Bamler. A compact representation for bayesian neural networks by removing permutation symmetry. *arXiv preprint arXiv:2401.00611*, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

David Yunis, Kumar Kshitij Patel, Pedro Henrique Pamplona Savarese, Gal Vardi, Jonathan Frankle, Matthew Walter, Karen Livescu, and Michael Maire. On convexity and linear mode connectivity in neural networks. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Understanding mode connectivity via parameter space symmetry. *arXiv preprint arXiv:2505.23681*, 2025.

Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. *arXiv preprint arXiv:2005.00060*, 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.

Zhanpeng Zhou, Yongyi Yang, Xiaojiang Yang, Junchi Yan, and Wei Hu. Going beyond linear mode connectivity: The layerwise linear feature connectivity. *arXiv preprint arXiv:2307.08286*, 2023.

TABLE OF NOTATION

General Mathematical Notation	
$\mathbb{R}^n$	n-dimensional Euclidean space
$\mathbb{R}^{m\times n}$	Space of $m \times n$ real matrices
$\lVert \cdot \rVert_F$	Frobenius norm of a matrix
trace$(\cdot)$	Trace of a square matrix
sym(M)	Symmetrization of a matrix M, defined as $(M + M^\top)/2$
Dimensions and Indices	
d	Dimension of token embeddings
d_h	Dimension of each attention head (typically d/h)
h	Number of attention heads in a model
L	Length of the input token sequence
m, n, k	Indices representing positions in a sequence
i, j, p	Indices representing attention heads
Spaces and Parameters	
$\mathcal{S}$	The space of all token sequences, $\bigsqcup_{L=1}^{\infty} \mathbb{R}^{L\times d}$
$W_i^Q, W_i^K, W_i^V, W_i^O$	Query, key, value, and output matrices of head i, each in $\mathbb{R}^{d\times d_h}$
θ	The complete set of parameters for a multi-head attention layer
$\Theta(d, d_h, h)$	The parameter space for a multi-head attention layer, $(\mathbb{R}^{d\times d_h})^{4h}$
$A_i^{m,n}, B_i$	Parameter matrices for the general multi-head attention formulation
Symmetry Groups	
S_h	The permutation group on a set of h elements
GL(d_h)	The general linear group of invertible $d_h \times d_h$ matrices
$G_{\text{Att}}(d_h, h)$	The symmetry group for standard multi-head attention
H(d_h)	The symmetry group for the RoPE query-key mechanism
$G_{\text{RoPE}}(d_h, h)$	The symmetry group for multi-head attention with RoPE
Positional Encodings	
p_m	The absolute positional encoding vector for position m
R_n	The block-diagonal rotation matrix for position n in RoPE
φ_i	The rotation frequency for the i-th block in RoPE matrices
P_i, J_i	2D block-diagonal matrices used to define H(d_h)
Matching Algorithm	
$C, C_{i,j}$	The cost matrix used for the linear assignment problem and its entries
π^*	The optimal head permutation
$\mathcal{L}_{Q,K}(U)$	The loss function for aligning query-key matrices with matrix U
$\mathcal{L}_{V,O}(V)$	The loss function for aligning value-output matrices with matrix V
$g_j(x)$	The 1D scalar objective function for RoPE alignment in subspace j
$\eta_{Q,j}, \eta_{K,j}$	Constants representing squared Frobenius norms to align RoPE
$\gamma_{Q,j}, \gamma_{K,j}$	Constants representing complex correlation scalars to align RoPE

Supplement to “Functional Equivalence in Attention: A Comprehensive Study with Applications to Linear Mode Connectivity”

Table of Contents

A Organization of the Paper and Appendix

Although this work is lengthy, *its core contributions can be distilled into a compact framework that is accessible even to readers interested solely in theoretical analysis, solely in empirical evaluation, or in both*. This section serves as the preamble to the Appendix, where we provide a comprehensive overview of our main results, encompassing both theoretical developments and experimental findings. The purpose of this summary is to orient the reader before engaging with the detailed technical content that follows, and to clarify how each component contributes to the overarching narrative of the work.

Main Paper. The organization of the main paper is as follows.

1. Section 1 provides an introduction and related work on Linear Mode Connectivity. Related concepts, such as functional equivalence and alignment methods, are also introduced in connection with prior literature.
2. Section 2 reviews vanilla attention, including its parameter space, symmetry group, and a result from literature–Theorem 2.1–which establishes complete functional equivalence for vanilla attention.
3. Section 3 examines how positional encodings may alter the internal structure of attention, thereby rendering the analysis from the vanilla case no longer directly applicable. While absolute PEs of the additive type do not affect the structure, relative PEs (with particular emphasis on Rotary PE) fundamentally change the attention mechanism. The corresponding symmetry group for the RoPE case is presented, which is strictly smaller than in the vanilla or APE setting. This reduction in symmetry implies that the function class realized by RoPE attention is strictly larger, providing a theoretical explanation for its increasing prominence in practice.
4. Section 4 focuses primarily on the RoPE case. First, we extend the RoPE setting to a general attention formulation that accommodates all cases of interest. In this formulation, the similarity score between two tokens at their specific positional indices is expressed as a bilinear form or quadratic norm. The result on functional equivalence of this settings is provided in Theorem 4.1. This framework subsumes vanilla attention, sinusoidal PE, and RoPE. To the best of our knowledge, this constitutes the most general formulation of attention studied under functional equivalence to date. Using this formulation, we then characterize the functional equivalence of the RoPE case, presented in Theorem 4.2.
5. Section 5 introduces an alignment method that serves as a tool for examining linear mode connectivity (LMC) in attention-based models. We propose a two-stage alignment algorithm for multi-head attention layers, applicable to both standard MHA and MHA with RoPE. The first stage matches the ordering of attention heads between two models by solving a linear assignment problem. The second stage aligns the internal parameters of each matched head pair independently for Query-Key and Value-Output components, optimizing over the appropriate symmetry group ($\mathrm{GL}(d_h)$ for standard MHA or $\mathrm{H}(d_h)$ for RoPE) via gradient descent or efficient scalar minimization. Remarks extend the method to include biases, full Transformer blocks, and full Transformer models.
6. Section 6 examines LMC under four re-initialization strategies, with emphasis on the first attention layer and full model resets, while intermediate cases are reported in the Appendix. Experiments are conducted across diverse Vision and NLP tasks. Ablation studies confirm the effectiveness of the two-stage matching algorithm in reducing barriers: Ablation study

for Stage 1 demonstrates that head permutation is crucial for finding LMC, while Ablation study for Stage 2 shows its importance that incorporating gradient descent optimization further improves alignment and reduces barriers.

7. Section 7 summarizes our findings, discusses limitations, and outlines future directions.

Appendix. The appendices provide complete proofs of the theoretical results in the main paper, the proposed matching algorithms, as well as additional experimental details.

Theoretical Proofs. Appendices B, C, D, E, and F contain all theoretical aspects and proofs related to functional equivalence. The main theoretical results of our work are Theorem 4.1 and Theorem 4.2. These two theorems are self-contained and can be understood directly from their statements, with all assumptions and settings specified in the main paper. *For readers not interested in the detailed proofs, this summary should suffice to convey the essence of our theoretical contributions, and the corresponding sections may be safely skipped.*

1. Appendix B formally defines the attention mechanism and its parameter space, followed by a description of how positional encodings are incorporated into attention.
2. Appendix C briefly describes the symmetry structures of vanilla attention, attention with absolute PEs, and attention with relative PEs (with emphasis on RoPE).
3. Appendix D introduces the general attention formulation. Theorem D.1, which is Theorem 4.1 in the main paper, establishes the functional equivalence of this general setting. The proof can be sketched as follows: starting from the softmax operator, we multiply through the denominators to rewrite the expression as an exponential polynomial, and then apply results and techniques from this area to complete the argument. All key intermediate results used as lemmas in the proof are stated in a self-contained manner in Appendix E, which includes
   (a) Appendix E.1 presents a result on the linear independence of exponential polynomials over the field of rational functions.
   (b) Appendix E.2 recalls Hall's Marriage Theorem, a classical result in combinatorics that is employed in some double-counting arguments used in our proof.
   (c) Appendix E.3 provides background on the Möbius function, with a particular focus on the partition lattice, and states a combinatorial identity that is used in our proof.
   (d) Appendix E.4 establishes a lemma on weighted sums over tuples, which is applied in our proof.
4. Appendix F applies the functional equivalence analysis of the general attention case to the specific setting of RoPE. Theorem F.1, corresponding to Theorem 4.2 in the main paper, provides the full details of this analysis. The proof proceeds as follows: RoPE is first reformulated as a special case of the general attention formulation via reparameterization; we then apply Theorem D.1 (4.1), and finally invoke a structural property of the rotary matrix, stated in Lemma F.2 of Appendix F.2, to recover the relationship between the original attention parameters.

Matching Algorithm. Appendix G develops the two-stage alignment procedure: first permuting attention heads via a linear assignment problem, then refining parameters with structured transformations. Key lemmas provide gradients for general linear updates, an SVD-based orthogonal initialization, and a RoPE-specific reduction to 2D subproblems. Algorithm 1 summarizes the complete method.

Experimental Details. Appendix I provides a comprehensive description of the experimental setup, including datasets, training protocols, and hyperparameters, along with additional results to ensure reproducibility. Appendix J further illustrates the interpolation results through detailed figures:

1. Appendix J.1 reports experiments on re-initializing only the first attention layer, highlighting its dominant role in shaping early representations.
2. Appendix J.2 investigates re-initialization of all attention layers, showing the cumulative effect of disrupting contextual interactions across the network.
3. Appendix J.3 studies re-initialization of the first Transformer layer, coupling attention and its adjacent feedforward block to examine early-layer sensitivity.

4. Appendix J.4 evaluates the most extreme setting where the entire Transformer is re-initialized, quantifying the magnitude of barriers introduced by full resets.
5. Appendix J.5 presents ablation studies on head permutation, including the two-stage matching algorithm. Stage 1 demonstrates the necessity of optimal head alignment for preserving linear mode connectivity, while Stage 2 leverages gradient refinement to further reduce interpolation barriers.

The experimental findings indicate that linear mode connectivity (LMC) manifests robustly in encoder-only architectures across a diverse set of vision and text classification benchmarks, including MNIST, CIFAR-10/100, ImageNet-21K → CIFAR transfer, ImageNet-1K, AGNews, IMDB Reviews, and DBpedia. By contrast, for large-scale language modeling datasets such as Enwik8, WikiText-103, and One Billion Word, LMC is exhibited exclusively under first attention layer and first-layer re-initialization. This phenomenon can be attributed to the reliance on GPT-2 models—decoder-only Transformers employing causal attention—which inherently impose more restrictive conditions on interpolation and connectivity.

B Multihead Attention Mechanism

B.1 Multihead Attention

General Formulation of Multihead Attention. Let d be a positive integer presenting the dimension of tokens and L be a positive integer presenting the sequence length. Denote the space of all sequences of tokens as:

$$\mathcal{S} := \bigsqcup_{L=1}^{\infty} \mathbb{R}^{L\times d}. \tag{25}$$

Consider a parameterized similarity map

$$f(\cdot,\cdot : \phi)\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \tag{26}$$

which assigns a score to a pair of tokens, and a parameterized embedding map

$$g(\cdot : \pi)\colon \mathbb{R}^d \to \mathbb{R}^d, \tag{27}$$

which produces token representations. The parameters are denoted $\phi \in \Phi$ and $\pi \in \Pi$, respectively.

Given an input sequence $\mathbf{x} = (x_1, \ldots, x_L)^\top \in \mathbb{R}^{L\times d}$, the multihead attention mechanism with h heads is defined by

$$\begin{aligned}&\mathrm{MultiHead}\Big(\mathbf{x} : \ \{\phi_i\}_{i=1}^h, \{\pi_i\}_{i=1}^h\Big)\\ &= \sum_{i=1}^h \mathrm{softmax}\begin{bmatrix} f(x_1,x_1 : \phi_i) & f(x_1,x_2 : \phi_i) & \cdots & f(x_1,x_L : \phi_i)\\ f(x_2,x_1 : \phi_i) & f(x_2,x_2 : \phi_i) & \cdots & f(x_2,x_L : \phi_i)\\ \vdots & \vdots & \ddots & \vdots\\ f(x_L,x_1 : \phi_i) & f(x_L,x_2 : \phi_i) & \cdots & f(x_L,x_L : \phi_i)\end{bmatrix} \cdot \begin{bmatrix} g(x_1 : \pi_i)\\ g(x_2 : \pi_i)\\ \vdots\\ g(x_L : \pi_i)\end{bmatrix}.\end{aligned} \tag{28}$$

Here the *attention matrix* of $\mathbf{x}$ is obtained by applying the softmax operator row-wise:

$$\mathrm{softmax}\begin{bmatrix} f(x_1,x_1 : \phi_i) & f(x_1,x_2 : \phi_i) & \cdots & f(x_1,x_L : \phi_i)\\ f(x_2,x_1 : \phi_i) & f(x_2,x_2 : \phi_i) & \cdots & f(x_2,x_L : \phi_i)\\ \vdots & \vdots & \ddots & \vdots\\ f(x_L,x_1 : \phi_i) & f(x_L,x_2 : \phi_i) & \cdots & f(x_L,x_L : \phi_i)\end{bmatrix}, \tag{29}$$

so that each row represents a probability distribution over the contributions of input tokens to a given output token.

Weight-Space Formulation. In standard practice, the similarity function is implemented via query–key projections. With a fixed head dimension $d_h \in \mathbb{N}$, one sets

$$\phi = (W^Q, W^K), \quad W^Q, W^K \in \mathbb{R}^{d\times d_h}, \tag{30}$$

and defines

$$f(x, y\colon \phi) = (xW^Q)(yW^K)^\top. \tag{31}$$

The embedding function is parameterized by

$$\pi = (W^V, W^O), \quad W^V, W^O \in \mathbb{R}^{d \times d_h}, \tag{32}$$

and specified as

$$g(x\colon \pi) = (xW^V)(W^O)^\top. \tag{33}$$

Typically, the head dimension is chosen as $d_h = d/h$. In this case, the multihead attention operator takes the form

$$\begin{aligned}\text{MultiHead}\Big(\mathbf{x} :\ \{W_i^Q, W_i^K, W_i^V, W_i^O\}_{i=1}^h\Big) \\ = \sum_{i=1}^h \text{softmax}\left(\left(\mathbf{x}W_i^Q\right)\left(\mathbf{x}W_i^K\right)^\top\right) \cdot \left(\mathbf{x}W_i^V\right)\left(W_i^O\right)^\top.\end{aligned} \tag{34}$$

The parameterization of the multihead operator is thus given by

$$\theta = \left(W_i^Q, W_i^K, W_i^V, W_i^O\right)_{i=1}^h \in \left(\mathbb{R}^{d \times d_h}\right)^{4h}. \tag{35}$$

Consequently, the parameter space of a multihead attention layer with h heads is

$$\Theta(d, d_h, h) = \left(\mathbb{R}^{d \times d_h}\right)^{4h}. \tag{36}$$

B.2 Positional Encoding

The multi-head attention mechanism, as formulated in subsection B.1, is inherently permutation-invariant: the similarity scores $f(x_j, x_k\colon \phi_i)$ and value projections $g(x_k\colon \pi_i)$ depend solely on the token representations, disregarding their sequential order. This property enables parallel computation but renders the model incapable of distinguishing sequences that differ only in token positions. To inject order information, positional encodings (PEs) are essential. We categorize PEs into two primary classes: *absolute positional encodings (APEs)*, which associate a unique vector with each absolute position, and *relative positional encodings (RPEs)*, which encode pairwise relative displacements to promote translation equivariance.

B.2.1 Absolute Positional Encodings

In the absolute paradigm, each position $m \in \{1, \ldots, L\}$ is mapped to a fixed vector $p_m \in \mathbb{R}^d$, independent of the sequence content $\mathbf{x} = (x_1, \ldots, x_L)^\top \in \mathbb{R}^{L \times d}$. The positional vectors are added elementwise to the token embeddings, yielding $\mathbf{x}' = \mathbf{x} + \mathbf{p}$ where $\mathbf{p} = (p_1, \ldots, p_L)^\top \in \mathbb{R}^{L \times d}$. The multi-head attention then processes this augmented input:

$$\begin{aligned}&\text{MultiHead}\Big(\mathbf{x} + \mathbf{p} :\ \{\phi_i\}_{i=1}^h, \{\pi_i\}_{i=1}^h\Big) \\ &= \sum_{i=1}^h \text{softmax}\begin{bmatrix} f(x_1+p_1, x_1+p_1\colon \phi_i) & \cdots & f(x_1+p_1, x_L+p_L\colon \phi_i) \\ f(x_2+p_2, x_1+p_1\colon \phi_i) & \cdots & f(x_2+p_2, x_L+p_L\colon \phi_i) \\ \vdots & \ddots & \vdots \\ f(x_L+p_L, x_1+p_1\colon \phi_i) & \cdots & f(x_L+p_L, x_L+p_L\colon \phi_i) \end{bmatrix} \\ &\qquad\qquad \cdot \begin{bmatrix} g(x_1+p_1\colon \pi_i) \\ g(x_2+p_2\colon \pi_i) \\ \vdots \\ g(x_L+p_L\colon \pi_i) \end{bmatrix}.\end{aligned} \tag{37}$$

Under the standard dot-product parameterization where the parameters for head i are $\phi_i = (W_i^Q, W_i^K)$, with $W_i^Q, W_i^K \in \mathbb{R}^{d \times d_h}$ and $d_h = d/h$, the similarity score expands to

$$\begin{aligned} f(x_m+p_m, x_n+p_n\colon \phi_i) &= (x_m+p_m)W_i^Q\left((x_n+p_n)W_i^K\right)^\top \\ &= x_m W_i^Q (W_i^K)^\top x_n^\top + x_m W_i^Q (W_i^K)^\top p_n^\top \\ &\quad + p_m W_i^Q (W_i^K)^\top x_n^\top + p_m W_i^Q (W_i^K)^\top p_n^\top, \end{aligned} \tag{38}$$

revealing explicit coupling of absolute positions through content-position, position-content, and position-position interactions.

A foundational instantiation is the *sinusoidal encoding* from the original Transformer (Vaswani et al., 2017), where components of $p_m \in \mathbb{R}^d$ (assuming d even) are

$$p_{m,2k} = \sin\left(\frac{m}{10000^{2k/d}}\right), \qquad p_{m,2k+1} = \cos\left(\frac{m}{10000^{2k/d}}\right), \tag{39}$$

for $0 \leq k < d/2$. This deterministic, parameter-free construction embeds positions in a periodic space, allowing relative distances to be recovered via linear combinations of vectors. It supports extrapolation to unseen lengths, though empirical gains are modest (Dai et al., 2019).

Alternatively, *learned absolute embeddings* treat $\{p_m\}_{m=1}^{L}$ as trainable parameters optimized jointly with the model (Devlin et al., 2019). This approach adapts to task-specific patterns, often boosting in-domain performance, but lacks the inductive bias of sinusoids and generalizes poorly beyond the maximum training length L_{train}, as unseen p_m for $m > L_{\text{train}}$ are undefined.

For vision tasks, APEs extend to 2D grids in models like the Vision Transformer (ViT) (Dosovitskiy et al., 2020), where learnable $p_{u,v} \in \mathbb{R}^d$ for patch positions $(u, v) \in \{1, \dots, H\} \times \{1, \dots, W\}$ are added to patch embeddings $x_{u,v}$, preserving absolute spatial structure.

B.2.2 Relative Positional Encodings

Unlike APEs that inject a unique signal for each absolute position, RPEs integrate relational information directly into the self-attention mechanism. The core principle is that the interaction between two tokens should primarily depend on their content and the relative distance between them, ra ther than their absolute positions in the sequence. This property, known as translation equivariance, is crucial for generalization to sequence lengths unseen during training.

Formally, RPEs parameterize the similarity function $f(\cdot,\cdot)$ in the attention mechanism with pairwise terms $\phi_i^{m,n}$ that depend on the positions m and n for each attention head i. The multi-head attention output is then computed as:

$$\begin{aligned}&\text{MultiHead}\Big(\mathbf{x} : \ \{\{\phi_i^{m,n}\}_{m,n}, \pi_i\}_{i=1}^{h}\Big)\\ &= \sum_{i=1}^{h} \text{softmax}\begin{bmatrix} f(x_1, x_1 : \phi_i^{1,1}) & f(x_1, x_2 : \phi_i^{1,2}) & \cdots & f(x_1, x_L : \phi_i^{1,L}) \\ f(x_2, x_1 : \phi_i^{2,1}) & f(x_2, x_2 : \phi_i^{2,2}) & \cdots & f(x_2, x_L : \phi_i^{2,L}) \\ \vdots & \vdots & \ddots & \vdots \\ f(x_L, x_1 : \phi_i^{L,1}) & f(x_L, x_2 : \phi_i^{L,2}) & \cdots & f(x_L, x_L : \phi_i^{L,L}) \end{bmatrix} \\ &\qquad \cdot \begin{bmatrix} g(x_1 : \pi_i) \\ g(x_2 : \pi_i) \\ \vdots \\ g(x_L : \pi_i) \end{bmatrix},\end{aligned} \tag{40}$$

with value projections g unaffected by positions. Translation equivariance is enforced via

$$\phi_i^{m,n} = \phi_i^{m+k,n+k}, \qquad \forall m, n, k \in \mathbb{Z}, \tag{41}$$

so $\phi_i^{m,n}$ depends only on the relative offset $m - n$, making attention scores functions of token content and displacement.
Several influential RPE variants have been proposed. Early work by Shaw et al. (2018) introduced additive relative embeddings, which augment the key vectors with learnable embeddings corresponding to the clipped relative distance between the query and key. A simpler and highly effective approach, popularized by the T5 model, involves adding a learned scalar bias directly to the pre-softmax attention logits, where biases are efficiently parameterized by bucketing nearby relative positions (Raffel et al., 2020). Building on this, ALiBi (Attention with Linear Biases) proposed a parameter-free scheme where the bias is a fixed, head-specific linear penalty proportional to the token distance, a simple yet powerful inductive bias that grants remarkable extrapolation capabilities (Press et al., 2021).

While these additive and bias-based methods are effective, a novel approach, **Rotary Positional Encoding (RoPE)** (Su et al., 2024), has emerged as the predominant method. It is utilized in most of the popular Large Language Models, including the LLaMA (Touvron et al., 2023), PaLM (Chowdhery et al., 2023), CodeGen (Nijkamp et al., 2022), and DeepSeek (Liu et al., 2024) families of models.

Rotary Positional Encoding (RoPE). Instead of adding signals to keys or attention logits, RoPE applies position-dependent orthogonal rotations to the query and key vectors. This elegantly encodes relative position information by leveraging the property that the inner product of two rotated vectors depends only on their original content and the relative rotation angle.

Assuming the embedding dimension d is even, the block-diagonal rotation matrix $R_n^d \in \mathbb{R}^{d\times d}$ for a token at position n is defined as

$$R_n^d = \begin{bmatrix} \cos(n\varphi_1) & -\sin(n\varphi_1) & 0 & \cdots & 0 & \\ \sin(n\varphi_1) & \cos(n\varphi_1) & 0 & \cdots & 0 & \\ 0 & 0 & \cos(n\varphi_2) & \cdots & 0 & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & \cos(n\varphi_{d/2}) & -\sin(n\varphi_{d/2}) \\ 0 & 0 & 0 & \cdots & \sin(n\varphi_{d/2}) & \cos(n\varphi_{d/2}) \end{bmatrix}, \tag{42}$$

with $\varphi_i = 10000^{-2(i-1)/d}$ for $i = 1, \dots, d/2$. Rotations are applied per head to the head dimension d_h via the standard projections $W_i^Q, W_i^K \in \mathbb{R}^{d\times d_h}$:

$$f(x_m, x_n : \phi_i^{m,n}) = \left(x_m W_i^Q R_m^{d_h}\right)\left(x_n W_i^K R_n^{d_h}\right)^\top, \tag{43}$$

where $R_m^{d_h}(R_n^{d_h})^\top = R_{m-n}^{d_h}$, ensuring relative dependence. Values remain unrotated: $g(x_j : \pi_i) = (x_j W_i^V)(W_i^O)^\top$. RoPE's design provides robust relative position encoding with excellent performance in autoregressive modeling, leading to its widespread adoption.

The setting of RoPE is presented as follows. Let D be an even positive integer. For each integer n, the block-diagonal rotation matrix $R_n^D \in \mathbb{R}^{D\times D}$ at position n is defined as

$$R_n^D = \begin{bmatrix} \cos(n\varphi_1) & -\sin(n\varphi_1) & 0 & \cdots & 0 & \\ \sin(n\varphi_1) & \cos(n\varphi_1) & 0 & \cdots & 0 & \\ 0 & 0 & \cos(n\varphi_2) & \cdots & 0 & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & \cos(n\varphi_{D/2}) & -\sin(n\varphi_{D/2}) \\ 0 & 0 & 0 & \cdots & \sin(n\varphi_{D/2}) & \cos(n\varphi_{D/2}) \end{bmatrix}, \tag{44}$$

where the rotation frequencies are given by $\varphi_i = 10000^{-2(i-1)/D}$ for $i = 1, \dots, D/2$. Within the multihead attention mechanism, each attention head operates on a sub-dimension of size $D = d_h$. After the standard linear projections $W_i^Q, W_i^K \in \mathbb{R}^{d\times d_h}$, the rotary matrices are applied. For brevity, when the head dimension is clear, we omit the superscript and write R_n.

The similarity score between two tokens x_m and x_n for head i is expressed as

$$f(x_m, x_n : \phi_i^{m,n}) = \left(x_m W_i^Q R_m\right)\left(x_n W_i^K R_n\right)^\top, \tag{45}$$

A key structural property of the rotation matrices is

$$R_m(R_n)^\top = R_{m-n}, \tag{46}$$

which ensures that the similarity depends solely on the relative offset $(m - n)$, thereby encoding relative position information directly into the attention mechanism. The value transformation remains unchanged:

$$g(x_j : \pi_i) = (x_j W_i^V)(W_i^O)^\top. \tag{47}$$

The RoPE construction embeds tokens in a rotationally parameterized space, thereby enriching the query-key interactions with relative positional dependence while preserving the autoregressive modeling constraints. This property underlies its demonstrated empirical effectiveness and explains its prevalence in contemporary Transformer-based architectures.

For 2D data, such as in vision transformers, this concept can be extended. For instance, *Axial RoPE* (Heo et al., 2024) splits the feature dimension d into two halves, applying independent 1D rotary encodings along the height and width axes, respectively, to capture axial relative displacements.

Remark B.1 (Comparison between Absolute and Relative Encoding)**.** APEs provide a straightforward global anchor via additive vectors p_m, with sinusoids offering extrapolation structure and learned variants task adaptation, though both risk overfitting to training lengths. RPEs, by contrast, emphasize relational offsets through translation-invariant $\phi_i^{m,n}$, yielding superior equivariance and generalization—especially in RoPE and ALiBi, which balance expressivity and efficiency. Recent advances further enhance RPE extrapolation: position interpolation (PI) rescales frequencies for longer contexts (Chen et al., 2023b), YaRN dynamically adjusts rotations (Peng et al., 2023), and data-adaptive methods like DAPE learn offset-specific encodings (Zheng et al., 2024).

C Functional Equivalence of Attention Mechanism with Positional Encoding

In this section, we investigate the functional equivalence of the attention mechanism. Building on the discussion from the previous section, our focus is on how positional encodings influence the functional equivalence of the standard attention formulation. Since a comprehensive analysis of all available positional encoding schemes would be prohibitively lengthy–given the wide variety that have been proposed–we restrict our attention to the two most classical forms that continue to be widely used in contemporary Transformer architectures: sinusoidal positional encoding and rotary positional encoding.

C.1 Attention with no Positional Encoding

Let d, d_h be positive integers with $d \geq d_h$. Recall that, a multihead attention operator with h heads is given by

$$\begin{aligned}\text{MultiHead}&\Big(\mathbf{x} : \{W_i^Q, W_i^K, W_i^V, W_i^O\}_{i=1}^h\Big)\\ &= \sum_{i=1}^h \text{softmax}\left(\left(\mathbf{x}W_i^Q\right)\left(\mathbf{x}W_i^K\right)^\top\right) \cdot \left(\mathbf{x}W_i^V\right)(W_i^O)^\top. \end{aligned} \tag{48}$$

The MultiHead map is parameterized by

$$\theta := \left(W_i^Q, W_i^K, W_i^V, W_i^O\right)_{i=1}^h \in \left(\mathbb{R}^{d\times d_h}\right)^{4h}. \tag{49}$$

and its parameter space is therefore

$$\Theta(d, d_h, h) := \left(\mathbb{R}^{d\times d_h}\right)^{4h}. \tag{50}$$

Group Action on the Parameter Space. Define the following group

$$G_{\text{Att}}(d_h, h) := S_h \times (\text{GL}(d_h) \times \text{GL}(d_h))^h. \tag{51}$$

This is precisely the direct product between the permutation group S_h and h copies of $\text{GL}(d_h) \times \text{GL}(d_h)$. Each group element $g \in G_{\text{Att}}(d_h, h)$ has the form

$$g := (\sigma, (U_i, V_i)_{i=1}^h), \tag{52}$$

where $\sigma \in S_h$ and $U_i, V_i \in \text{GL}(d_h)$. The natural action of $G_{\text{Att}}(d_h, h)$ on the parameter space $\Theta(d, d_h, h)$ is defined by

$$g\theta := \left(W_{\sigma(i)}^Q \cdot U_i^\top, W_{\sigma(i)}^K \cdot U_i^{-1}, W_{\sigma(i)}^V \cdot V_i^\top, W_{\sigma(i)}^O \cdot V_i^{-1}\right)_{i=1}^h \tag{53}$$

This action preserves the functionality of the MultiHead map: for all $\theta \in \Theta(d, d_h, h)$ and all $g \in G_{\text{Att}}(d_h, h)$,

$$\text{MultiHead}(\cdot : \theta) = \text{MultiHead}(\cdot : g\theta). \tag{54}$$

The contribution of the general linear group action vanishes through cancellation in the matrix multiplications, while the action induced by the permutation σ commutes with the addition operator. Taken together, these actions characterize the full symmetry of the multihead attention mechanism, as established in the following result from (Tran et al., 2025).

Theorem C.1 (See (Tran et al., 2025)). *Let*

$$\theta = \left(W_i^Q, W_i^K, W_i^V, W_i^O\right)_{i=1}^{h} \in \Theta(d, d_h, h), \text{ and} \tag{55}$$

$$\bar{\theta} = \left(\bar{W}_i^Q, \bar{W}_i^K, \bar{W}_i^V, \bar{W}_i^O\right)_{i=1}^{\bar{h}} \in \Theta(d, d_h, \bar{h}), \tag{56}$$

be two parameterizations of MultiHead *maps. Suppose that:*

1. *Every $d \times d_h$ matrix appearing in θ and $\bar{\theta}$ has full column rank d_h;*
2. *The matrices $\{W_i^Q (W_i^K)^\top\}_{i=1}^{h}$ are pairwise distinct;*
3. *The matrices $\{\bar{W}_i^Q (\bar{W}_i^K)^\top\}_{i=1}^{\bar{h}}$ are pairwise distinct.*

If the two MultiHead map are identical, i.e.,

$$\text{MultiHead}(\cdot\,:\,\theta) = \text{MultiHead}(\cdot\,:\,\bar{\theta}), \tag{57}$$

then, $h = \bar{h}$, and there exists $g \in G_{\text{Att}}(d_h, h)$ such that $\bar{\theta} = g\theta$.

Remark C.2. While the theorem imposes certain assumptions on the parameters of the MultiHead maps, it is important to emphasize that these conditions hold almost surely. For instance, a randomly chosen real matrix has full column rank with probability one, and a finite collection of real numbers is almost surely pairwise distinct. At a high level, the result may thus be interpreted as follows: after excluding a negligibly small subset of the parameter space (e.g., a set of measure zero or the complement of a dense set), the functional equivalence of MultiHead maps is completely characterized by the action of the symmetry group.

C.2 Sinusoidal Positional Encoding

Consider the case of sinusoidal positional encoding (PE). Let $\mathbf{p} = \{p_i\}_{i=1}^{\infty}$ denote the sequence of positional vectors, which encodes positional information. For an input sequence $\mathbf{x} \in \mathcal{S}$ of length L, i.e. $\mathbf{x} = (x_1, \ldots, x_L)^\top \in \mathbb{R}^{L \times d}$, the positional encoding is incorporated by addition, namely $\mathbf{x} + \mathbf{p} = (x_1 + p_1, \ldots, x_L + p_L)^\top$, which is then supplied as input to the MultiHead attention map.

In this formulation, the positional encoding does not alter the internal structure of the MultiHead mechanism itself; it merely applies a shift to the input sequence. Furthermore, the encoding map

$$\begin{aligned} \text{PE} \quad : \quad & \mathcal{S} \longrightarrow \mathcal{S} \\ & \mathbf{x} \longmapsto \mathbf{x} + \mathbf{p}, \end{aligned} \tag{58}$$

is bijective. Consequently, the introduction of sinusoidal PE has no effect on the analysis of functional equivalence for MultiHead attention. In particular, the functional equivalence classes in the presence of sinusoidal PE coincide exactly with those in the case without positional encoding.

C.3 Rotary Positional Encoding

Assume that d_h is even. For each integer n, recall R_n which is the $d_h \times d_h$ block-diagonal rotation matrix for a token at position n,

$$R_n = \begin{bmatrix} \cos(n\varphi_1) & -\sin(n\varphi_1) & 0 & \cdots & 0 & \\ \sin(n\varphi_1) & \cos(n\varphi_1) & 0 & \cdots & 0 & \\ 0 & 0 & \cos(n\varphi_2) & \cdots & 0 & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & \cos(n\varphi_{d_h/2}) & -\sin(n\varphi_{d_h/2}) \\ 0 & 0 & 0 & \cdots & \sin(n\varphi_{d_h/2}) & \cos(n\varphi_{d_h/2}) \end{bmatrix}, \tag{59}$$

with $\varphi_i = 10000^{-2(i-1)/d_h}$ for $i = 1, \ldots, d_h/2$. As discussed earlier, the subscript indicating the head embedding dimension d_h is omitted for brevity. Note that, $R_n = R_1^n$. The multihead attention mechanism with Rotary Positional Encoding (RoPE) is defined as

$$\begin{aligned}
&\text{MultiHead}_{\text{RoPE}}\Big(\mathbf{x} : \{W_i^Q, W_i^K, W_i^V, W_i^O\}_{i=1}^h\Big) \\
&\qquad = \sum_{i=1}^h \text{softmax}\left(\left(\mathbf{x}W_i^Q R_m\right)\left(\mathbf{x}W_i^K\right)^\top\right) \cdot \left(\mathbf{x}W_i^V R_n\right)(W_i^O)^\top \\
&\qquad = \sum_{i=1}^h \text{softmax}\Big[x_m W_i^Q R_m R_n^\top (W_i^K)^\top x_n^\top\Big]_{m,n=1,\ldots,L} \cdot \mathbf{x}W_i^V (W_i^O)^\top \\
&\qquad = \sum_{i=1}^h \text{softmax}\Big[x_m W_i^Q R_{m-n} (W_i^K)^\top x_n^\top\Big]_{m,n=1,\ldots,L} \cdot \mathbf{x}W_i^V (W_i^O)^\top. \qquad (60)
\end{aligned}$$

The parameterization and parameter space of MultiHeadRoPE coincide with those of the standard multihead attention map, namely

$$\theta = \Big(W_i^Q, W_i^K, W_i^V, W_i^O\Big)_{i=1}^h \in \Theta(d, d_h, h) = \left(\mathbb{R}^{d\times d_h}\right)^{4h}. \qquad (61)$$

Group Action on the Parameter Space. In contrast to the standard MultiHead maps, for MultiHeadRoPE, the action of $G_{\text{Att}}(d_h, h)$ on $\Theta(d, d_h, h)$ no longer preserves functionality. In particular, for $\theta \in \Theta(d, d_h, h)$ and $g \in G_{\text{Att}}(d_h, h)$, one generally has

$$\text{MultiHead}_{\text{RoPE}}(\cdot : \theta) \neq \text{MultiHead}_{\text{RoPE}}(\cdot : g\theta). \qquad (62)$$

The essential reason is as follows. While the interaction between W_i^V and W_i^O remains purely multiplicative and thus structurally consistent with the vanilla case, the query and key matrices W_i^Q and W_i^K are now separated by the relative rotary matrix R_{m-n}. This insertion prevents the cancellation of group actions induced by $\text{GL}(d_h)$, thereby breaking the invariance property enjoyed by the standard multihead attention mechanism.

To define the symmetry group of $\text{MultiHead}_{\text{RoPE}}$, first, denote these following matrices

$$P := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ and } \qquad J := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}. \qquad (63)$$

For each $i \in [d_h/2]$, define the matrices $P_i, J_i \in \mathbb{R}^{d_h \times d_h}$ as block-diagonal matrices with $d_h/2$ consecutive 2×2 diagonal blocks. Specifically, the i-th diagonal block of P_i (resp., J_i) is given by P (resp., J), while all other diagonal blocks are zero matrices,

$$P_i = \text{diag}(0, \ldots, 0, \underset{i\text{-th block}}{P}, 0, \ldots, 0), \text{ and } \quad J_i = \text{diag}(0, \ldots, 0, \underset{i\text{-th block}}{J}, 0, \ldots, 0). \qquad (64)$$

Define the following group

$$\text{H}(d_h) := \left\{ U = \sum_{i=1}^{d_h/2} (a_i P_i + b_i J_i) \in \mathbb{R}^{d_h\times d_h} : (a_i, b_i) \in \mathbb{R}^2 \setminus \{(0,0)\} \text{ for } i \in [d_h/2] \right\}. \qquad (65)$$

It is straightforward to verify that $\text{H}(d_h)$ forms an abelian subgroup of $\text{GL}(d_h)$. Moreover, $\text{H}(d_h)$ is isomorphic to $(\mathbb{C}^\times)^{d_h/2}$, where $\mathbb{C}^\times$ denotes the multiplicative group of nonzero complex numbers. In particular, the rotary matrices R_n are contained in $\text{H}(d_h)$ for all n. Define the following group,

$$G_{\text{RoPE}} := S_h \times \left(\text{H}(d_h) \times \text{GL}(d_h)\right)^h. \qquad (66)$$

The group G_{RoPE} is a subgroup of G_{Att}. The group action of G_{Att} on Θ restricts naturally to a group action of G_{RoPE} on Θ. The central observation of this section is that this action preserves the functionality of the $\text{MultiHead}_{\text{RoPE}}$ map. In particular, for all $\theta \in \Theta(d, d_h, h)$ and all $g \in G_{\text{RoPE}}(d_h, h)$, one has

$$\text{MultiHead}_{\text{RoPE}}(\cdot : \theta) = \text{MultiHead}_{\text{RoPE}}(\cdot : g\theta). \qquad (67)$$

The justification is as follows. In comparison with the case of standard MultiHead maps, apart from the head permutation σ and the linear interaction between W_i^V and W_i^O, the only structural difference lies in the interaction between W_i^Q and W_i^K. However, since the group $H(d_h)$ is abelian and the rotary matrix R_n belongs to $H(d_h)$, we obtain

$$\begin{aligned}(W_i^Q U^\top) R_n (W_i^K U^{-1})^\top &= W_i^Q U^\top R_n (U^{-1})^\top (W_i^K)^\top \\ &= W_i^Q R_n U^\top (U^{-1})^\top (W_i^K)^\top = W_i^Q R_n (W_i^K)^\top. \end{aligned} \tag{68}$$

Thus the multiplication inside the softmax operator of the MultiHead$_{\text{RoPE}}$ map remains invariant under the action of G_{RoPE}.

Remark C.3. In the next section, we present the main result of this work, which establishes that the group G_{RoPE} completely characterizes the symmetry structure of the MultiHead$_{\text{RoPE}}$ map.

D A General Formulation for Multihead Attention and its Functional Equivalence

D.1 A General Formulation for Multihead Attention

We consider a general setting where the functions f and g are parameterized as follows:

$$f(\cdot,\cdot\colon A \in \mathbb{R}^{d\times d}) : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}, \qquad (x,y) \longmapsto xAy^\top, \tag{69}$$

$$g(\cdot\colon B \in \mathbb{R}^{d\times d}) : \mathbb{R}^d \longrightarrow \mathbb{R}, \qquad x \longmapsto xB. \tag{70}$$

The MultiHead map with h heads is parameterized by two families of matrices:

$$\{A_i^{m,n}\}_{i=1}^h, \quad \{B_i\}_{i=1}^h, \tag{71}$$

where each $A_i^{m,n}, B_i \in \mathbb{R}^{d\times d}$, as follows

$$\begin{aligned}&\text{MultiHead}\Big(\mathbf{x} :\ \{\{\phi_i^{m,n}\}_{m,n}, \pi_i\}_{i=1}^h\Big) \\ &= \sum_{i=1}^h \text{softmax} \begin{bmatrix} f(x_1,x_1\colon A_i^{1,1}) & f(x_1,x_2\colon A_i^{1,2}) & \cdots & f(x_1,x_L\colon A_i^{1,L}) \\ f(x_2,x_1\colon A_i^{2,1}) & f(x_2,x_2\colon A_i^{2,2}) & \cdots & f(x_2,x_L\colon A_i^{2,L}) \\ \vdots & \vdots & \ddots & \vdots \\ f(x_L,x_1\colon A_i^{L,1}) & f(x_L,x_2\colon A_i^{L,2}) & \cdots & f(x_L,x_L\colon A_i^{L,L}) \end{bmatrix} \cdot \begin{bmatrix} g(x_1\colon B_i) \\ g(x_2\colon B_i) \\ \vdots \\ g(x_L\colon B_i) \end{bmatrix} \end{aligned} \tag{72}$$

$$= \sum_{i=1}^h \text{softmax}\left[x_m A_i^{m,n} x_n^\top\right]_{m,n=1,\ldots,L} \cdot \mathbf{x} B_i. \tag{73}$$

We begin with two observations that facilitate the subsequent analysis of *Functional Equivalence* of MultiHead Attention.

1. (*Relative positional encoding assumption.*) For all $m,n \geq 1$ and for all shifts $k \geq 0$, we assume

$$A^{m,n} = A^{m+k,n+k}. \tag{74}$$

   This corresponds to the natural stationarity condition imposed by relative positional encodings.

2. (*Diagonal self-similarity terms are symmetric.*) For each $m \geq 1$, the matrix $A_i^{m,m}$ parameterizes the function f that computes the similarity score of the m-th token with itself at the i-th head, namely

$$x_m A_i^{m,m} x_m^\top. \tag{75}$$

   Since every quadratic form corresponds uniquely to a symmetric matrix, we may, without loss of generality, symmetrize $A_i^{m,m}$:

$$A_i^{m,m} \ \longmapsto\ \frac{A_i^{m,m} + (A_i^{m,m})^\top}{2}. \tag{76}$$

Note that this transformation leaves the function invariant:

$$x_m A_i^{m,m} x_m^\top \;=\; x_m \left(\frac{A_i^{m,m} + (A_i^{m,m})^\top}{2} \right) x_m^\top. \tag{77}$$

Henceforth, we assume that all $A_i^{m,m}$ are symmetric.

Under this framework, we now consider the situation where two MultiHead maps, one with h heads and the other with $\bar{h}$ heads, yield identical outputs:

$$\begin{aligned} &\text{MultiHead}\Big(\mathbf{x} :\; \{\{A_i^{m,n}\}_{m,n}, B_i\}_{i=1}^{h}\Big) \\ &\qquad\qquad = \text{MultiHead}\Big(\mathbf{x} :\; \{\{\bar{A}_i^{m,n}\}_{m,n}, \bar{B}_i\}_{i=1}^{\bar{h}}\Big). \end{aligned} \tag{78}$$

Since $g(\cdot : B) = -g(\cdot : -B)$, Equation (78) is equivalent to the assertion that a MultiHead map with $h + \bar{h}$ heads vanishes identically:

$$0 = \text{MultiHead}\Big(\mathbf{x} \;:\; \{\{A_i^{m,n}\}_{m,n}\}_{i=1}^{h} \sqcup \{\{\bar{A}_i^{m,n}\}_{m,n}\}_{i=1}^{\bar{h}},\; \{B_i\}_{i=1}^{h} \sqcup \{-\bar{B}_i\}_{i=1}^{\bar{h}}\Big). \tag{79}$$

Thus, the first step in analyzing functional equivalence is to characterize precisely when a MultiHead map is identically zero. Before presenting the proof, we introduce the following notion. We say that two parameter families $\{A^{m,n}\}_{m,n}$ and $\{\bar{A}^{m,n}\}_{m,n}$ are *distinct* if there exist indices $m, n \geq 1$ such that

$$A^{m,n} \neq \bar{A}^{m,n}. \tag{80}$$

We are now in position to state the main result of this section, which provides a fundamental insight into the problem of Functional Equivalence in MultiHead Attention.

D.2 Functional Equivalence of General Multihead Attention

Theorem D.1 (Theorem 4.1 in the main paper)**.** *Consider the* MultiHead *map with h heads, parameterized by families of matrices $\{\{A_i^{m,n}\}_{m,n}\}_{i=1}^{h} \subset \mathbb{R}^{d\times d}$ and $\{B_i\}_{i=1}^{h} \subset \mathbb{R}^{d\times d}$, i.e.,*

$$\text{MultiHead}\Big(\mathbf{x} :\; \{\{A_i^{m,n}\}_{m,n}, B_i\}_{i=1}^{h}\Big). \tag{81}$$

Assume that the attention parameter families

$$\{A_1^{m,n}\}_{m,n}, \{A_2^{m,n}\}_{m,n}, \ldots, \{A_h^{m,n}\}_{m,n}, \tag{82}$$

are pairwise distinct, and further that $A_i^{m,n}$ is nonzero for all $i \in [h]$ and $m, n \geq 1$. If, for all $\mathbf{x} \in \mathcal{S} = \sqcup_{L=1}^{\infty} \mathbb{R}^{L\times d}$, one has

$$\text{MultiHead}\Big(\mathbf{x} :\; \{\{A_i^{m,n}\}_{m,n}, B_i\}_{i=1}^{h}\Big) = 0. \tag{83}$$

then, $B_1, \ldots, B_h$ are equal to 0.

Proof. To enhance clarity, we begin by outlining the main steps of the proof at a high level:

1. **Preliminary setup.** We first record some initial observations and introduce the necessary notation in preparation for the proof. In particular, we note that it suffices to show that at least one of the coefficients B_i must vanish. Once this is established, symmetry in the construction allows us to conclude that in fact all B_i must be equal to zero, thereby proving the theorem.

2. **Reformulation as an exponential polynomial.** Starting from Equation (83), we show that

$$0 = \sum_{(t_1,\ldots,t_h)\in[L]^h} \exp\left(\sum_{i=1}^{h} x_k A_i^{k,t_i} x_{t_i}^\top \right) \left(\sum_{i=1}^{h} x_{t_i} B_i \right). \tag{84}$$

   This identity arises naturally from a double-counting argument. The resulting expression has the structure of an exponential polynomial that is identically zero. To analyze such expressions, we invoke the linear independence results for exponential functions over rational fields, which allow us to isolate relations among the coefficients.

3. **Structural constraints on the B_i.** By applying the above linear independence principle, we identify a fundamental structural constraint on the coefficients B_i. Specifically, the symmetry conditions imposed by the $A_i^{k,t}$ on admissible permutations force the B_i to satisfy a family of linear relations indexed by $i \in [h]$. These constraints form the core of the argument: they reduce the problem of analyzing a complicated exponential sum to verifying the consistency of a system of linear equations in the B_i.

4. **Partition-based refinement.** We next examine the equalities that occur within the sets of h elements $\{A_i^{k,t}\}_{i=1}^h$. This step is preparatory: it shows that the relations identified in the previous step are not only necessary but also sufficient to deduce that at least one B_i must vanish. The analysis exploits the partition structure $\{U_p\}$, together with the existence of carefully chosen subsets V^{t_j}, to sharpen the constraint and isolate specific indices.

5. **Conclusion.** Finally, we combine the above ingredients to conclude the proof. The linear relations obtained in **Step 3**, when applied to the partition refinement of **Step 4**, imply that one of the B_i's must equal zero. By the initial reduction in Step 1, this suffices to deduce that in fact all $B_i = 0$. This completes the proof of the theorem.

We proceed to present the complete details of the proof.

Step 1.

We express the formulation of

$$\text{MultiHead}\Big(\mathbf{x} \ : \ \{\{A_i^{m,n}\}_{m,n}, B_i\}_{i=1}^h\Big) \tag{85}$$

in a token-wise manner. From Equation (83), for every $1 \le k \le L$, one has

$$\sum_{i=1}^h \left(\sum_{j=1}^L \frac{\exp(x_k A_i^{k,j} x_j^\top)}{\sum_{q=1}^L \exp(x_k A_i^{k,q} x_q^\top)} \cdot x_j B_i \right) = 0. \tag{86}$$

Since the families $\{A_1^{m,n}\}_{m,n}, \{A_2^{m,n}\}_{m,n}, \dots, \{A_h^{m,n}\}_{m,n}$ are pairwise distinct, and for each i, $A_i^{m,n}$ depends only on the difference $(m - n)$, one can choose a sufficiently large L and an index k such that the h sets

$$\{A_1^{k,n}\}_{n\ge1},\ \{A_2^{k,n}\}_{n\ge1},\ \dots,\ \{A_h^{k,n}\}_{n\ge1}$$

are pairwise distinct. For the remainder of the proof, we fix such a k and consider all $L \ge k$.

By induction, it suffices to establish that at least one of $B_1, \dots, B_h$ vanishes. Indeed, if this holds, then the problem reduces to a MultiHead Attention mechanism with fewer heads, and repeating the argument shows that all $B_1, \dots, B_h$ must be zero. Consequently, our goal is to prove that there exists at least one index $1 \le i \le h$ such that $B_i = 0$.

Step 2.

First, we rewrite Equation (86) in a more convenient form. By multiplying out all denominators in Equation (86), we obtain

$$\sum_{i=1}^h \left(\sum_{j=1}^L \exp\Big(x_k A_i^{k,j} x_j^\top\Big) \cdot \prod_{p\in[h]\setminus\{i\}} \left(\sum_{q=1}^L \exp\big(x_k A_p^{k,q} x_q^\top\big) \right) \cdot x_j B_i \right) = 0. \tag{87}$$

We now observe that the left-hand side of Equation (87) can be re-expressed as

$$\begin{aligned} &\sum_{i=1}^h \left(\sum_{j=1}^L \exp\Big(x_k A_i^{k,j} x_j^\top\Big) \cdot \prod_{p\in[h]\setminus\{i\}} \left(\sum_{q=1}^L \exp\big(x_k A_p^{k,q} x_q^\top\big) \right) \cdot x_j B_i \right) \\ &\qquad = \sum_{(t_1,\dots,t_h)\in[L]^h} \exp\left(\sum_{i=1}^h x_k A_i^{k,t_i} x_{t_i}^\top \right) \left(\sum_{i=1}^h x_{t_i} B_i \right). \end{aligned} \tag{88}$$

To verify Equation (88), define for $i \in [h]$ and $j \in [L]$,

$$a_{i,j} := \exp\Big(x_k A_i^{k,j} x_j^\top\Big), \qquad b_{i,j} := x_j B_i. \tag{89}$$

In this notation, the claimed identity becomes

$$\sum_{i=1}^{h}\left(\sum_{j=1}^{L} a_{i,j} \prod_{p\in[h]\setminus\{i\}} \sum_{q=1}^{L} a_{p,q} \cdot b_{i,j}\right) = \sum_{(t_1,\dots,t_h)\in[L]^h}\left(\prod_{i=1}^{h} a_{i,t_i}\right)\left(\sum_{i=1}^{h} b_{i,t_i}\right). \tag{90}$$

For $(i,\mathbf{t}) \in [h]\times[L]^h$, define the weight

$$w(i,\mathbf{t}) := \left(\prod_{p=1}^{h} a_{p,t_p}\right) b_{i,t_i}. \tag{91}$$

We will compute the following quantity in two ways,

$$\sum_{(i,\mathbf{t})\in[h]\times[L]^h} w(i,\mathbf{t}). \tag{92}$$

Group by the distinguished index i.

Fix $i \in [h]$. Then

$$\begin{aligned}\sum_{\mathbf{t}\in[L]^h} w(i,\mathbf{t}) &= \sum_{t_i=1}^{L} \sum_{(t_p)_{p\neq i}\in[L]^{h-1}} \left(\prod_{p=1}^{h} a_{p,t_p}\right) b_{i,t_i} \\ &= \sum_{t_i=1}^{L} a_{i,t_i}\, b_{i,t_i} \underbrace{\sum_{(t_p)_{p\neq i}\in[L]^{h-1}} \prod_{p\neq i} a_{p,t_p}}_{(\star)}.\end{aligned} \tag{93}$$

The inner sum $(\star)$ equals

$$\prod_{p\neq i}\sum_{q=1}^{L} a_{p,q}, \tag{94}$$

since expanding the product enumerates every choice of $(t_p)_{p\neq i}$ exactly once. Hence

$$\sum_{\mathbf{t}\in[L]^h} w(i,\mathbf{t}) = \sum_{j=1}^{L} a_{i,j}\left(\prod_{p\neq i}\sum_{q=1}^{L} a_{p,q}\right) b_{i,j}. \tag{95}$$

Summing over $i = 1,\dots,h$ yields the left-hand side of Equation (90).

Group by the tuple $\mathbf{t}$.

Fix $\mathbf{t} = (t_1,\dots,t_h) \in [L]^h$. Then

$$\sum_{i=1}^{h} w(i,\mathbf{t}) = \sum_{i=1}^{h}\left(\prod_{p=1}^{h} a_{p,t_p}\right) b_{i,t_i} = \left(\prod_{p=1}^{h} a_{p,t_p}\right)\left(\sum_{i=1}^{h} b_{i,t_i}\right). \tag{96}$$

Summing over all $\mathbf{t}$ yields the right-hand side of Equation (90).

In conclusion, both groupings compute the same total $\sum_{(i,\mathbf{t})\in\Omega} w(i,\mathbf{t})$, so Equation (90) holds. Substituting back $a_{i,j} = \exp(x_k A_i^{k,j} x_j^\top)$, $b_{i,j} = x_j B_i$ recovers the original identity. From Equation (87) and Equation (88), we conclude that

$$0 = \sum_{(t_1,\dots,t_h)\in[L]^h}\left[\exp\left(\sum_{i=1}^{h} x_k A_i^{k,t_i} x_{t_i}^\top\right)\left(\sum_{i=1}^{h} x_{t_i} B_i\right)\right]. \tag{97}$$

Note that in Equation (97), both sides represent vectors in $\mathbb{R}^d$. If we examine a single coordinate of this vector, the identity remains valid by restricting each B_i to the corresponding column indexed by that coordinate. Hence, without loss of generality, we may interpret Equation (97) under the convention that each B_i is regarded as a column vector in $\mathbb{R}^d$ corresponding to the chosen coordinate.

Step 3.

For $(t_1, \ldots, t_h) \in \mathbb{N}^h$, define

$$g_{(t_1,\ldots,t_h)}(\mathbf{x}) := \sum_{i=1}^{h} x_k A_i^{k,t_i} x_{t_i}^\top \qquad \in \mathbb{R}[\mathbf{x}], \tag{98}$$

$$h_{(t_1,\ldots,t_h)}(\mathbf{x}) := \sum_{i=1}^{h} x_{t_i} B_i \qquad \in \mathbb{R}[\mathbf{x}], \tag{99}$$

$$f_{(t_1,\ldots,t_h)}(\mathbf{x}) := \exp\big(g_{k,(t_1,\ldots,t_h)}(\mathbf{x})\big)\, h_{(t_1,\ldots,t_h)}(\mathbf{x}). \tag{100}$$

Then Equation (97) can be rewritten as

$$\begin{aligned} 0 &= \sum_{(t_1,\ldots,t_h)\in[L]^h} f_{(t_1,\ldots,t_h)}(\mathbf{x}) \\ &= \sum_{(t_1,\ldots,t_h)\in[L]^h} \exp\big(g_{(t_1,\ldots,t_h)}(\mathbf{x})\big)\; h_{(t_1,\ldots,t_h)}(\mathbf{x}). \end{aligned} \tag{101}$$

Observe that each polynomial $g_{(t_1,\ldots,t_h)} \in \mathbb{R}[\mathbf{x}]$ has constant term equal to zero. By Lemma E.1, Equation (101) implies that, for each $g \in \mathbb{R}[\mathbf{x}]$, grouping together all indices $(t_1, \ldots, t_h)$ such that $g_{(t_1,\ldots,t_h)} = g$ yields

$$0 = \sum_{(t_1,\ldots,t_h)\in[L]^h \,:\, g_{(t_1,\ldots,t_h)}=g} \exp\big(g_{(t_1,\ldots,t_h)}(\mathbf{x})\big)\, h_{(t_1,\ldots,t_h)}(\mathbf{x}), \tag{102}$$

and since $\exp(g(\mathbf{x}))$ is common to all such terms, we conclude

$$0 = \sum_{(t_1,\ldots,t_h)\in[L]^h \,:\, g_{(t_1,\ldots,t_h)}=g} h_{(t_1,\ldots,t_h)}(\mathbf{x}). \tag{103}$$

One has the following observation. Consider an arbitrary tuple $(t_1, \ldots, t_h) \in [L]^h$ such that $t_1, \ldots, t_h$ are pairwise distinct. Assume that there exists another tuple $(t'_1, \ldots, t'_h) \in [L]^h$ satisfying

$$g_{(t_1,\ldots,t_h)} = g_{(t'_1,\ldots,t'_h)}. \tag{104}$$

Since all $A_i^{m,n}$ are nonzero and $A_i^{m,m}$ is symmetric, it follows that every polynomial of the form $x_m A_i^{m,n} x_n$ is nonvanishing. Consequently, in $g_{k,(t_1,\ldots,t_h)}$, for each $i \in [h]$, there must exist polynomial terms that involve at least one entry of x_{t_i}. (This requirement that the t_i's be pairwise distinct is crucial, as it prevents possible cancellation of terms.) Hence, for each $i \in [h]$, there exists $j \in [h]$ such that $t_i = t'_j$. Moreover, since the t_i's are pairwise distinct, it follows that $(t'_1, \ldots, t'_h)$ must be a *permutation* of $(t_1, \ldots, t_h)$. From Equation (101) and Lemma E.1, one therefore obtains

$$0 = \sum_{\sigma\in S_h} h_{(t_{\sigma(1)},\ldots,t_{\sigma(h)})}(\mathbf{x}). \tag{105}$$

It should be emphasized, however, that the condition $(t'_1, \ldots, t'_h)$ being a permutation of $(t_1, \ldots, t_h)$ is not sufficient, in itself, to guarantee that $g_{(t_1,\ldots,t_h)} = g_{(t'_1,\ldots,t'_h)}$. To examine this more closely, let $(t'_1, \ldots, t'_h) = (t_{\sigma(1)}, \ldots, t_{\sigma(h)})$ for some $\sigma \in S_h$. From the assumption $g_{(t_1,\ldots,t_h)} = g_{(t'_1,\ldots,t'_h)}$, we have

$$\sum_{i=1}^{h} x_k A_i^{k,t_i} x_{t_i}^\top = \sum_{i=1}^{h} x_k A_i^{k,t_{\sigma(i)}} x_{t_{\sigma(i)}}^\top. \tag{106}$$

By reindexing the summation, this is equivalent to

$$\sum_{i=1}^{h} x_k A_i^{k,t_i} x_{t_i}^\top = \sum_{i=1}^{h} x_k A_{\sigma^{-1}(i)}^{k,t_i} x_{t_i}^\top, \tag{107}$$

which in turn is equivalent to requiring that $A_i^{k,t_i} = A_{\sigma^{-1}(i)}^{k,t_i}$ for all $i \in [h]$. This shows explicitly the additional algebraic condition that must hold in order for two permutations to yield the same polynomial g. Note that this constitutes a sufficient condition on $\sigma \in S_h$ to ensure that $g_{(t_1,\ldots,t_h)} = g_{(t'_1,\ldots,t'_h)}$ whenever $(t'_1,\ldots,t'_h) = (t_{\sigma(1)},\ldots,t_{\sigma(h)})$.

Accordingly, one deduces

$$\begin{aligned}
0 &= \sum_{\sigma \in S_h \;:\; A_j^{k,t_j} = A_{\sigma^{-1}(j)}^{k,t_j} \; \forall j \in [h]} h_{(t_{\sigma(1)},\ldots,t_{\sigma(h)})}(\mathbf{x}) \\
&= \sum_{\sigma \in S_h \;:\; A_j^{k,t_j} = A_{\sigma^{-1}(j)}^{k,t_j} \; \forall j \in [h]} \left(\sum_{i=1}^{h} x_{t_{\sigma(i)}} B_i \right) \\
&= \sum_{\sigma \in S_h \;:\; A_j^{k,t_j} = A_{\sigma^{-1}(j)}^{k,t_j} \; \forall j \in [h]} \left(\sum_{i=1}^{h} x_{t_i} B_{\sigma^{-1}(i)} \right) \\
&= \sum_{\sigma \in S_h \;:\; A_j^{k,t_j} = A_{\sigma(j)}^{k,t_j} \; \forall j \in [h]} \left(\sum_{i=1}^{h} x_{t_i} B_{\sigma(i)} \right) \\
&= \sum_{i=1}^{h} \left(x_{t_i} \cdot \sum_{\sigma \in S_h \;:\; A_j^{k,t_j} = A_{\sigma(j)}^{k,t_j} \; \forall j \in [h]} B_{\sigma(i)} \right).
\end{aligned} \tag{108}$$

Thus, since the entries $t_1,\ldots,t_h$ are pairwise distinct, the monomials x_{t_i} are linearly independent. It therefore follows that, for each $i \in [h]$, one must have

$$0 = \sum_{\sigma \in S_h \;:\; A_j^{k,t_j} = A_{\sigma(j)}^{k,t_j} \; \forall j \in [h]} B_{\sigma(i)}. \tag{109}$$

Equation (109) encapsulates the key structural constraint on the coefficients B_i. It shows that, once the $A_i^{k,t}$'s impose symmetry conditions on admissible permutations, the B_i's must satisfy a family of linear relations indexed by $i \in [h]$. This relation will serve as the main tool in subsequent steps, where we will exploit the partition structure of the U_p's to force specific B_i's to vanish.

Step 4.

For each $t \in \mathbb{N}$, define $\{U_p^t\}_{p=1}^{\alpha_t}$ to be the unique partition of $[h]$ such that, for $i, j \in [h]$, one has $A_i^{k,t} = A_j^{k,t}$ if and only if i and j belong to the same set U_p^t. Since the number of possible partitions of $\{1,\ldots,h\}$ is finite, there exists a partition $\{U_p\}_{p=1}^{\alpha}$ such that the equality

$$\{U_p^t\}_{p=1}^{\alpha_t} = \{U_p\}_{p=1}^{\alpha} \tag{110}$$

holds for infinitely many values of $t \in \mathbb{N}$. Let S denote the set of all such positive integers t.

By reindexing the head indices if necessary, we may assume that

$$U_1 = \{1,\ldots,m\}. \tag{111}$$

Next, observe that since the h sequences

$$\{A_1^{k,n}\}_{n\geq 1}, \quad \{A_2^{k,n}\}_{n\geq 1}, \quad \dots, \quad \{A_h^{k,n}\}_{n\geq 1} \tag{112}$$

are pairwise distinct, there exists a positive integer K such that the truncated sequences

$$\{A_1^{k,n}\}_{n=1}^{K}, \quad \{A_2^{k,n}\}_{n=1}^{K}, \quad \dots, \quad \{A_h^{k,n}\}_{n=1}^{K} \tag{113}$$

are already pairwise distinct. We then discard all integers $t \leq K$ from the set S, and by a slight abuse of notation, continue to denote the resulting subset by the same symbol S.

Finally, for each partition $\{U_p^t\}_{p=1}^{\alpha_t}$, we denote by $U^t(1)$ the unique set that contains the index 1.

(i) The intersection of K sets $U^1(1), U^2(1), \dots, U^K(1)$ is precisely $\{1\}$, i.e.,

$$U^1(1) \cap U^2(1) \cap \cdots \cap U^K(1) = \{1\}. \tag{114}$$

Indeed, since $1 \in U^t(1)$ for all $t = 1, \dots, K$, it follows immediately that

$$1 \in U^1(1) \cap U^2(1) \cap \cdots \cap U^K(1). \tag{115}$$

Suppose, for the sake of contradiction, that there exists some $i \in [h]$ with $i > 1$ such that

$$i \in U^1(1) \cap U^2(1) \cap \cdots \cap U^K(1). \tag{116}$$

By the construction of $U^t(1)$, this assumption implies that $A_1^{k,t} = A_i^{k,t}$ for all $t = 1, \dots, K$. Equivalently, the infinite sequences $\{A_1^{k,n}\}_{n\geq 1}$ and $\{A_i^{k,n}\}_{n\geq 1}$ coincide. This, however, contradicts the fact that their finite truncations

$$\{A_1^{k,n}\}_{n=1}^{K}, \quad \{A_2^{k,n}\}_{n=1}^{K}, \quad \dots, \quad \{A_h^{k,n}\}_{n=1}^{K}$$

are pairwise distinct by the choice of K.

Therefore, no such $i > 1$ can exist. The only common element across all $U^1(1), \dots, U^K(1)$ is the index 1, which establishes the claim.

(ii) For each $t = 1, \dots, K$, define the set

$$V^t := U^t(1) \cap \{1, 2, \dots, m\} \subset \{1, 2, \dots, m\}. \tag{117}$$

Then, one has

$$V^1 \cap V^2 \cap \cdots \cap V^K = \{1\}. \tag{118}$$

Indeed, one computes

$$\begin{aligned} V^1 \cap V^2 \cap \cdots \cap V^K &= \bigcap_{t=1}^{K} \big(U^t(1) \cap \{1, \dots, m\}\big) \\ &= \bigcap_{t=1}^{K} U^t(1) \ \cap \ \{1, \dots, m\} \\ &= \{1\} \cap \{1, \dots, m\} \\ &= \{1\}. \end{aligned} \tag{119}$$

(iii) Among the K sets $V^1, \dots, V^K$, there exists a positive integer $\gamma < m$ such that one can select γ sets, say $V^{t_1}, \dots, V^{t_\gamma}$ with $1 \leq t_1 < t_2 < \cdots < t_\gamma \leq K$, satisfying the following property: the intersection of these γ sets is $\{1\}$, whereas the intersection of any $\gamma - 1$ among them is no longer $\{1\}$.

To prove this, let γ be the smallest positive integer such that there exist γ sets among $V^1, \dots, V^K$ whose intersection equals $\{1\}$. The existence of such a γ is guaranteed since the intersection of all

K sets is $\{1\}$. Denote these γ sets by $V^{t_1}, \ldots, V^{t_\gamma}$. By the minimality of γ, if one removes any single set from $\{V^{t_1}, \ldots, V^{t_\gamma}\}$, the intersection of the remaining $\gamma - 1$ sets cannot be $\{1\}$.

It remains to show that $\gamma < m$. By minimality, it suffices to establish the existence of fewer than m sets among $\{V^1, \ldots, V^K\}$ whose intersection is $\{1\}$. Since

$$V^1 \cap V^2 \cap \cdots \cap V^K = \{1\}, \tag{120}$$

for each $i \in \{2, \ldots, m\}$ there must exist at least one set among $V^1, \ldots, V^K$ that does not contain i. As there are $m - 1$ such indices i, we can collect at most $m - 1$ sets that collectively exclude all of these elements. Consequently, the intersection of these at most $m - 1$ sets is $\{1\}$, which proves $\gamma \le m - 1 < m$.

This completes the proof. The argument is essentially a pigeonhole-type principle: since every element $i \in \{2, \ldots, m\}$ must be excluded by at least one set, and there are $m - 1$ such elements in total, at most $m - 1$ sets suffice to ensure that all of them are removed, leaving only 1 in the intersection.

(iv) In those γ sets $V^{t_1}, \ldots, V^{t_\gamma}$ in (iii), for each $i \in [\gamma]$, one can choose $v_i \in V^{t_i}$ such that $v_1, \ldots, v_\gamma$ are pairwise distinct.

This is a standard application of the Hall Marriage Theorem (see Appendix E.2). For convenience, rename V^{t_i} as W^i for $i \in [\gamma]$. For each $k \in \{1, \ldots, \gamma\}$, by assumption, we may choose

$$b_k \in \Big(\bigcap_{i \neq k} W^i\Big) \setminus \{1\}. \tag{121}$$

By construction, $b_k \neq 1$, and $b_k \in W^i$ for all $i \neq k$. Moreover, $b_k \notin W^k$, since otherwise b_k would belong to $\bigcap_{i=1}^{\gamma} W^i = \{1\}$, a contradiction. Let $B = \{b_1, \ldots, b_\gamma\}$. Consider the bipartite graph with left vertices $\{W^1, \ldots, W^\gamma\}$ and right vertices $\{1\} \cup B \subseteq \{1, \ldots, m\}$, with an edge $W^i \leftrightarrow x$ whenever $x \in W^i$. A system of distinct representatives (SDR) of size γ in this graph yields the desired elements $v_i \in W^i$. By Hall's theorem, it suffices to show that for every nonempty $J \subseteq \{1, \ldots, \gamma\}$, the neighborhood $N(J)$ satisfies $|N(J)| \ge |J|$.

- If $|J| = 1$, say $J = \{i\}$, then $1 \in W^i$. Furthermore, for every $k \neq i$ we have $b_k \in W^i$. Thus

$$|N(J)| \ge 1 + (\gamma - 1) = \gamma \ge |J|. \tag{122}$$

- If $|J| \ge 2$, fix $k \in \{1, \ldots, \gamma\}$.
  - If $k \notin J$, then $b_k \in W^i$ for every $i \in J$, hence $b_k \in N(J)$.
  - If $k \in J$, pick any $j \in J \setminus \{k\}$. Since $b_k \in W^j$, it follows that $b_k \in N(J)$.

  Thus every b_k belongs to $N(J)$, and clearly $1 \in N(J)$. Hence

$$|N(J)| \ge |B| + 1 = \gamma + 1 \ge |J|. \tag{123}$$

Since Hall's condition is satisfied, there exists a matching that assigns to each W^i a distinct element of $\{1\} \cup B$ contained in W^i. These assigned elements provide the required representatives $v_i \in W^i$, which are pairwise distinct.

Step 5.

To deliver the result of this part, we now employ the token indices $t_1, \ldots, t_\gamma$ identified in *(iii)* and *(iv)* of **Step 4**, together with the token indices in the set S also obtained in **Step 4**. We recall the properties of these token indices that will be used:

1. For all $t \in S$, the partition $\{U_p^t\}_{p=1}^{\alpha_t}$, defined in **Step 4**, coincides with $\{U_p\}_{p=1}^{\alpha}$. In particular, by reindexing the head indices, we may assume $U_1 = \{1, \ldots, m\}$. This guarantees that the structure of the partition is stable across infinitely many $t \in S$, providing us with a consistent reference framework.

2. For all t_i with $i \in [\gamma]$, where $\gamma < m$, recall that $V^{t_i} = U^{t_i}(1) \cap \{1, \ldots, m\}$. One can select γ head indices $v_i \in V^{t_i}$ such that they are pairwise distinct. This property will be crucial later when we need to ensure that certain representatives can be chosen without overlap.

We also recall the main result from **Step 3**, namely Equation (109): for any $(s_1, \ldots, s_h) \in [L]^h$ with pairwise distinct entries, and for each $i \in [h]$, one has

$$0 = \sum_{\sigma \in S_h \,:\, A_j^{k,s_j} = A_{\sigma(j)}^{k,s_j} \ \forall j \in [h]} B_{\sigma(i)}. \tag{124}$$

This identity is the foundation of the argument: it asserts that, under the given matching condition on the coefficients A_j^{k,s_j}, a nontrivial linear combination of the B_i's must vanish.

Now, in Equation (124), let us consider $(s_1, \ldots, s_h) \in [L]^h$ constructed as follows. First, observe that the index set $\{1, \ldots, h\}$ can be decomposed into three disjoint parts:

$$\{1, \ldots, h\} = \{v_1, \ldots, v_\gamma\} \ \sqcup \ \big(\{1, \ldots, m\} \setminus \{v_1, \ldots, v_\gamma\}\big) \ \sqcup \ \big(U_2 \sqcup U_3 \sqcup \cdots \sqcup U_\alpha\big). \tag{125}$$

The first component corresponds to the specially chosen distinct representatives v_i, the second to the remaining elements of U_1, and the third to all indices belonging to the other partition classes $U_2, \ldots, U_\alpha$.

Now fix a subset $T \subset [\gamma]$. Define $(s_1, \ldots, s_h) \in [L]^h$ by setting, for each $j \in [h]$,

1. If $j = v_i$ for some $i \in T$, then set $s_j = s_{v_i} = t_i$. In other words, the positions corresponding to T are aligned with the distinguished token indices t_i.
2. If $j \in \{1, \ldots, m\} \setminus \{v_i : i \in T\}$, take s_j to be an arbitrary element of S. This ensures consistency with the partition structure while leaving us flexibility in the assignment.
3. If $j \in U_p$ for some $2 \leq p \leq \alpha$, then take s_j to be an arbitrary element of S. Again, this choice respects the partitioning of indices into classes U_p.

For the chosen $(s_1, \ldots, s_h) \in [L]^h$, we analyze which $\sigma \in S_h$ satisfy the condition $A_j^{k,s_j} = A_{\sigma(j)}^{k,s_j}$ for all $j \in [h]$. We make the following observations, case by case:

1. For $j \in U_2 \sqcup U_3 \sqcup \cdots \sqcup U_\alpha$, say $j \in U_p$ with $2 \leq p \leq \alpha$, the condition $A_j^{k,s_j} = A_{\sigma(j)}^{k,s_j}$ implies $\sigma(j) \in U_p$. Hence
$$\sigma(U_2 \sqcup U_3 \sqcup \cdots \sqcup U_\alpha) = U_2 \sqcup U_3 \sqcup \cdots \sqcup U_\alpha, \tag{126}$$
and consequently $\sigma(U_1) = U_1$. In particular, if $j \in U_1$, then $\sigma(j) \in U_1$.
2. For $j \in \{1, \ldots, m\} \setminus \{v_i : i \in T\}$, if $A_j^{k,s_j} = A_{\sigma(j)}^{k,s_j}$, then necessarily $\sigma(j) \in U_1 = \{1, \ldots, m\}$. Thus the entire set U_1 is stable under σ, but the specific images of these indices may vary within U_1.
3. For $j = v_i$ with $i \in T$, if $A_j^{k,s_j} = A_{\sigma(j)}^{k,s_j}$, then $\sigma(j) \in U^{s_{v_i}}(1) = U^{t_i}(1)$. From the previous point, we also know $\sigma(j) \in U_1$. Taken together, these conditions imply that $\sigma(j) \in V^{t_i} = U^{t_i}(1) \cap U_1$. In other words, the image of v_i under σ is constrained to lie inside the restricted set V^{t_i}.

Therefore, specifying a $\sigma \in S_h$ that satisfies $A_j^{k,s_j} = A_{\sigma(j)}^{k,s_j}$ for all $j \in [h]$ is equivalent to:

1. For each $j = v_i$ with $i \in T$, choosing $\sigma(j) = \sigma(v_i) \in V^{t_i}$,
2. For each $j \in \{1, \ldots, m\} \setminus \{v_i : i \in T\}$, choosing $\sigma(j) \in U_1 \setminus \{\sigma(v_i) : i \in T\}$ arbitrarily,
3. For each $j \in U_p$ with $2 \leq p \leq \alpha$, choosing $\sigma(j) \in U_p$.

In conclusion, the structure of admissible permutations σ in Equation (124) is fully determined by the subset $T \subset [\gamma]$ and the representatives $v_i \in V^{t_i}$ chosen in **Step 4**. This description clarifies how the constraints arising from the partition classes U_p and the distinguished representatives v_i together restrict the allowed form of σ. Consequently, the sum in Equation (124) can be partitioned into contributions indexed by subsets $T \subset [\gamma]$, which will be the key mechanism for deriving vanishing conditions on the B_i's in the subsequent step.

With these observations in hand, we now perform explicit computations. Fix one choice of $(s_1, \dots, s_h) \in [L]^h$ satisfying the above construction, and in Equation (124) take $i = v_i$ for some $i \in T$. The equation then specializes to

$$\begin{aligned} 0 &= \sum_{\sigma \in S_h \,:\, A_j^{k,t_j} = A_{\sigma(j)}^{k,t_j} \ \forall j \in [h]} B_{\sigma(v_i)} \\ &= \sum_{v \in V^{t_i}} B_v \cdot \Big(\text{the number of } h\text{-tuples in the Cartesian product} \\ &\qquad\qquad \prod_{j \in T} V^{t_j} \times U_1^{m-|T|} \times \prod_{p=2}^{\alpha} U_p^{|U_p|}, \\ &\qquad\qquad \text{such that all } h \text{ entries are pairwise distinct, and} \\ &\qquad\qquad \text{the coordinate corresponding to } V^{t_i} \text{ is fixed to be } v \Big). \end{aligned} \tag{127}$$

The interpretation is as follows: each valid permutation σ contributes one admissible tuple, and the contribution is grouped according to which element $v \in V^{t_i}$ is assigned to the coordinate corresponding to V^{t_i}. The factor multiplying B_v therefore counts exactly the number of such admissible tuples.

Now, observe that once the coordinates corresponding to the V^{t_j}'s are chosen, all the remaining coordinates can be filled freely within their respective partition blocks. In particular:

- The indices in $\{1, \dots, m\} \setminus \{v_i : i \in T\}$ may be permuted arbitrarily within U_1, yielding a factor of $(m - |T|)!$.
- For each $p \in \{2, \dots, \alpha\}$, the indices in U_p may also be permuted arbitrarily, contributing a factor of $|U_p|!$.

Hence the above expression simplifies to

$$\begin{aligned} 0 = \sum_{v \in V^{t_i}} B_v \cdot (m - |T|)! \cdot \prod_{p=2}^{\alpha} |U_p|! \\ \cdot \Big(\text{the number of } h\text{-tuples in the Cartesian product } \prod_{j \in T} V^{t_j}, \\ \text{such that all entries are pairwise distinct, and} \\ \text{the coordinate corresponding to } V^{t_i} \text{ equals } v \Big). \end{aligned} \tag{128}$$

Since the factorial factors are nonzero constants independent of the choice of v, we may divide them out to obtain the equivalent condition

$$\begin{aligned} 0 = \sum_{v \in V^{t_i}} B_v \cdot \Big(\text{the number of } h\text{-tuples in the Cartesian product } \prod_{j \in T} V^{t_j}, \\ \text{such that all entries are pairwise distinct, and} \\ \text{the coordinate corresponding to } V^{t_i} \text{ equals } v \Big). \end{aligned} \tag{129}$$

This identity holds for every choice of subset $T \subset [\gamma]$ and for every $v \in V^{t_i}$ with $i \in [\gamma]$. The key point is that the coefficients B_v appear only through such linear relations, weighted by combinatorial counts of admissible tuples. By applying Corollary E.14, we deduce that

$$0 = \sum_{i \in V^{t_1} \cap V^{t_2} \cap \cdots \cap V^{t_\gamma}} B_i. \tag{130}$$

Finally, recall from the construction in *(iii)* of **Step 4** that the intersection $V^{t_1} \cap V^{t_2} \cap \cdots \cap V^{t_\gamma}$ is exactly $\{1\}$. Therefore, the above equation reduces to

$$B_1 = 0, \tag{131}$$

We have established that $B_1 = 0$. By the preceding argument at the beginning of the proof, this immediately implies that all B_i vanish identically. Hence, we conclude that $B_i = 0$ for every i, which completes the proof. □

Remark D.2. Theorem D.1 can be viewed as a statement about the linear independence of attention heads. Although the theorem is formulated under specific assumptions on the parameters of the MultiHead maps, these conditions are satisfied with probability one. In essence, the result asserts that—except for a negligibly small subset of the parameter space (e.g., a measure-zero set or the complement of a dense subset)—the functional equivalence of general MultiHead maps can be completely characterized. The probabilistic nature of these assumptions aligns with those commonly made in prior studies on the functional equivalence of deep neural networks.

We have the following corollary of Theorem D.1.

Corollary D.3. *Consider two* MultiHead *maps with h and $\bar{h}$ heads, parameterized by families of matrices*

$$\begin{aligned} &\{\{A_i^{m,n}\}_{m,n}\}_{i=1}^{h} \subset \mathbb{R}^{d\times d}, \qquad \{B_i\}_{i=1}^{h} \subset \mathbb{R}^{d\times d}, \\ &\{\{\bar{A}_i^{m,n}\}_{m,n}\}_{i=1}^{\bar{h}} \subset \mathbb{R}^{d\times d}, \qquad \{\bar{B}_i\}_{i=1}^{\bar{h}} \subset \mathbb{R}^{d\times d}, \end{aligned} \tag{132}$$

respectively. Assume that $A_i^{m,n}$ and $\bar{A}_i^{m,n}$ are nonzero for all feasible triples (i, m, n). If the two MultiHead *maps are identical, i.e.,*

$$\begin{aligned} &\text{MultiHead}\Big(\mathbf{x} : \ \{\{A_i^{m,n}\}_{m,n}, B_i\}_{i=1}^{h}\Big) \\ &\qquad\qquad = \text{MultiHead}\Big(\mathbf{x} : \ \{\{\bar{A}_i^{m,n}\}_{m,n}, \bar{B}_i\}_{i=1}^{\bar{h}}\Big), \end{aligned} \tag{133}$$

then for every parameter family

$$\{A^{m,n}\}_{m,n} \subset \mathbb{R}^{d\times d}, \tag{134}$$

we have the identity

$$\sum_{i\in[h] \,:\, \{A_i^{m,n}\}_{m,n} = \{A^{m,n}\}_{m,n}} B_i = \sum_{i\in[\bar{h}] \,:\, \{\bar{A}_i^{m,n}\}_{m,n} = \{A^{m,n}\}_{m,n}} \bar{B}_i. \tag{135}$$

Proof. This follows directly from Theorem D.1. □

E Key Lemmas for the Functional Equivalence of General MultiHead Attention

In this section, we introduce the preliminary concepts and fundamental results that will serve as the foundation for the proofs of our main theorems.

E.1 A Result on the Linear Independence of Exponential Polynomials over the Field of Rational Functions

Let n be a positive integer. Recall that $\mathbb{R}[\mathbf{x}] = \mathbb{R}[x_1, \ldots, x_n]$ denotes the polynomial ring in n variables over $\mathbb{R}$. Its field of fractions is denoted by $\mathbb{R}(\mathbf{x})$, that is,

$$\mathbb{R}(\mathbf{x}) = \left\{ \frac{p}{q} \, : \, p, q \in \mathbb{R}[\mathbf{x}], \ q \neq 0 \right\}, \tag{136}$$

the field of all rational functions in the variables $x_1, \ldots, x_n$ with real coefficients.

We now state and prove a standard result concerning the linear independence of exponential polynomials over $\mathbb{R}(\mathbf{x})$.

Lemma E.1. *Let $p_1, \ldots, p_m$ be polynomials in $\mathbb{R}[\mathbf{x}]$ such that $p_i - p_j$ is nonconstant whenever $i \neq j$. Suppose $q_1, \ldots, q_m$ are rational functions in $\mathbb{R}(\mathbf{x})$ satisfying*

$$q_1 \cdot e^{p_1} + \cdots + q_m \cdot e^{p_m} = 0. \tag{137}$$

Then necessarily $q_1 = \cdots = q_m = 0$.

Proof. We proceed by induction on m.

Base case.

For $m = 1$, the statement is immediate. Indeed, if $q_1 \cdot e^{p_1} = 0$, then since e^{p_1} never vanishes, it follows that $q_1 = 0$.

Inductive step.

Assume the result holds for every collection of fewer than m exponentials. Let $q_1, \ldots, q_m \in \mathbb{R}(\mathbf{x})$ satisfy

$$q_1 \cdot e^{p_1} + \cdots + q_m \cdot e^{p_m} = 0. \tag{138}$$

We wish to show that all q_i vanish. Suppose, for contradiction, that not all q_i are zero. Without loss of generality, assume $q_m \neq 0$.

Dividing through Equation (138) by $q_m e^{p_m}$ yields

$$\frac{q_1}{q_m} \cdot e^{p_1 - p_m} + \cdots + \frac{q_{m-1}}{q_m} \cdot e^{p_{m-1} - p_m} + 1 = 0. \tag{139}$$

This expresses 1 as a linear combination of the exponentials $e^{p_j - p_m}$ with coefficients in $\mathbb{R}(\mathbf{x})$.

Now differentiate both sides of Equation (139) with respect to each variable x_i for $i = 1, \ldots, n$. Since the derivative of 1 is zero, we obtain

$$\sum_{j=1}^{m-1} \left(\frac{\partial}{\partial x_i} \left(\frac{q_j}{q_m} \right) + \frac{q_j}{q_m} \cdot \frac{\partial}{\partial x_i} (p_j - p_m) \right) e^{p_j - p_m} = 0. \tag{140}$$

Each coefficient in parentheses lies in $\mathbb{R}(\mathbf{x})$.

Since $p_1 - p_m, \ldots, p_{m-1} - p_m$ are pairwise distinct and nonconstant, the corresponding exponentials $e^{p_j - p_m}$ are linearly independent over $\mathbb{R}(\mathbf{x})$ by the induction hypothesis. Therefore, each coefficient in Equation (140) must vanish, i.e.,

$$\frac{\partial}{\partial x_i} \left(\frac{q_j}{q_m} \right) + \frac{q_j}{q_m} \cdot \frac{\partial}{\partial x_i} (p_j - p_m) = 0, \tag{141}$$

for every $i = 1, \ldots, n$ and $j = 1, \ldots, m-1$. Equivalently,

$$\frac{\partial}{\partial x_i} \left(\frac{q_j}{q_m} \cdot e^{p_j - p_m} \right) = 0. \tag{142}$$

This shows that for each $j = 1, \ldots, m-1$, the function

$$\frac{q_j}{q_m} \cdot e^{p_j - p_m} \tag{143}$$

is independent of all variables $x_1, \ldots, x_n$, and hence must be a constant $c_j \in \mathbb{R}$.

If some $c_j \neq 0$, then $q_j \neq 0$ and we would have

$$e^{p_j - p_m} = \frac{c_j q_m}{q_j}, \tag{144}$$

which would imply that $e^{p_j - p_m}$ is a rational function, and therefore constant. This contradicts the assumption that $p_j - p_m$ is nonconstant.

Thus, each $c_j = 0$, forcing $q_j = 0$ for all $j = 1, \ldots, m-1$. Substituting into Equation (139) then yields $1 = 0$, an impossibility.

Hence our assumption was false, and all $q_i = 0$. By induction, the lemma follows. □

Remark E.2. Lemma E.1 formalizes the intuitive fact that exponential functions with distinct polynomial exponents cannot cancel each other when combined with rational-function coefficients. It can be viewed as a multivariate generalization of the classical result that functions of the form e^{ax} with distinct real numbers a are linearly independent over the field of rational functions in one variable. Here, the same principle extends to exponential polynomials in several variables, with the essential role played by the assumption that the differences $p_i - p_j$ are nonconstant. This generalization is crucial for arguments in Theorem D.1, involving exponential polynomials over $\mathbb{R}(\mathbf{x})$.

E.2 Hall's Marriage Theorem and Systems of Distinct Representatives

In this section, we recall a classical result from combinatorics, known as *Hall's Marriage Theorem* (Hall, 1935), which provides necessary and sufficient conditions for the existence of a system of distinct representatives (SDR). This theorem will play a crucial role in our arguments, as our construction ultimately reduces to the problem of selecting distinct representatives from a family of subsets. Let $\mathcal{A} = \{A_1, A_2, \ldots, A_s\}$ be a finite family of subsets of a ground set X. A *system of distinct representatives* (SDR) for $\mathcal{A}$ is a set $\{a_1, a_2, \ldots, a_s\}$ such that $a_i \in A_i$ for each i and all $a_1, \ldots, a_s$ are pairwise distinct. Equivalently, an SDR is an injective choice function assigning to each A_i an element $a_i \in A_i$.

The existence of an SDR is a classical question in combinatorics, and Hall's theorem provides a complete characterization.

Theorem E.3 (Hall's Marriage Theorem)**.** *Let $\mathcal{A} = \{A_1, A_2, \ldots, A_s\}$ be a finite family of subsets of a set X. Then $\mathcal{A}$ admits a system of distinct representatives if and only if the following condition (Hall's condition) holds:*

$$\Big|\bigcup_{i \in J} A_i\Big| \geq |J| \quad \textit{for every subset } J \subseteq \{1, 2, \ldots, s\}. \tag{145}$$

In words, Hall's condition states that for every subcollection of the sets A_i, the total number of available elements in their union must be at least as large as the number of sets in the subcollection. This condition is clearly necessary: if $|J|$ sets are assigned representatives, then at least $|J|$ distinct elements are required. The theorem asserts that this necessary condition is also sufficient. Hall's theorem has many applications in combinatorics, graph theory, and matching theory. In the language of bipartite graphs, it gives a necessary and sufficient condition for the existence of a perfect matching from the left vertex set into the right vertex set.

Remark E.4. Hall's Marriage Theorem plays a central role in the argument of Theorem D.1. Moreover, its application is closely connected to the statements of Theorem E.11 and Corollary E.14.

E.3 The Möbius Function on the Partition Lattice

This section introduces the necessary background on incidence algebras and Möbius inversion over finite posets. We then establish an identity for the Möbius function that will serve as a fundamental tool throughout the remainder of the paper. We also present several connections between this identity and other well-studied combinatorial concepts, with the aim of providing readers with greater intuition about its significance. For comprehensive treatments of these topics, we refer the reader to (Rota, 1964; Stanley, 2011).

E.3.1 Incidence Algebras and Möbius Inversion on Finite Posets

Let $(P, \leq)$ be a finite poset. The *incidence algebra* $I(P)$ over $\mathbb{C}$ consists of all functions

$$f \coloneqq \{(x, y) \in P \times P : x \leq y\} \longrightarrow \mathbb{C}. \tag{146}$$

with convolution

$$(f * g)(x, y) \coloneqq \sum_{x \leq z \leq y} f(x, z)\, g(z, y), \qquad \text{for all } x \leq y. \tag{147}$$

The identity for convolution is the Kronecker delta $\delta(x, y)$ (i.e. $\delta(x, y) = 1$ if $x = y$, and 0 otherwise). The *zeta function* $\zeta \in I(P)$ is $\zeta(x, y) \equiv 1$ for $x \leq y$. An element $f \in I(P)$ is invertible if and only if $f(x, x) \neq 0$ for all $x \in P$; in that case f^{-1} is its inverse under convolution.

Möbius function. The *Möbius function* $\mu = \mu_P \in I(P)$ is defined as the convolution inverse of ζ:

$$\mu * \zeta = \zeta * \mu = \delta. \tag{148}$$

Equivalently, for all $x \leq y$ in P, one has

$$\sum_{x \leq z \leq y} \mu(x, z) = \delta(x, y). \tag{149}$$

As a consequence, if $f, g : P \to \mathbb{C}$ satisfy

$$f(x) = \sum_{y \geq x} g(y), \qquad \text{for all } \ x \in P, \tag{150}$$

then *Möbius inversion* yields

$$g(x) = \sum_{y \geq x} \mu(x, y)\, f(y), \qquad \text{for all } \ x \in P. \tag{151}$$

Products of posets. If P, Q are finite posets, their product $P \times Q$ is ordered componentwise. Define

$$(\zeta_P \otimes \zeta_Q)\big((p_1, q_1), (p_2, q_2)\big) \coloneqq \zeta_P(p_1, p_2)\, \zeta_Q(q_1, q_2). \tag{152}$$

A direct computation in $I(P \times Q)$ shows

$$\zeta_{P \times Q} = \zeta_P \otimes \zeta_Q, \tag{153}$$

$$(\mu_P \otimes \mu_Q) * (\zeta_P \otimes \zeta_Q) = \delta_P \otimes \delta_Q = \delta_{P \times Q}. \tag{154}$$

Hence

$$\mu_{P \times Q}\big((p_1, q_1), (p_2, q_2)\big) = \mu_P(p_1, p_2)\, \mu_Q(q_1, q_2). \tag{155}$$

E.3.2 The Partition Lattice and Interval Factorization

Let U be a finite set with $|U| = n$. The set $\Pi(U)$ of all set partitions of U, ordered by refinement, forms a finite lattice with minimum $\hat{0}$ (all singletons) and maximum $\hat{1}$ (one block). The goal of this section is to derive the following explicit formula, stated in the following theorem:

Theorem E.5. *For $\pi \in \Pi(U)$, one has:*

$$\mu_{\Pi(U)}(\hat{0}, \pi) = \prod_{B \in \pi} (-1)^{|B|-1}(|B| - 1)!. \tag{156}$$

For clarity, we begin with an outline of the proof. The reasoning unfolds in two stages.

1. **Interval factorization.** Restriction to blocks induces a canonical isomorphism:

$$[\hat{0}, \pi] \cong \prod_{B \in \pi} \Pi(B). \tag{157}$$

By multiplicativity of the Möbius function on products, one has:

$$\mu_{\Pi(U)}(\hat{0}, \pi) = \prod_{B \in \pi} \mu_{\Pi(B)}(\hat{0}_B, \hat{1}_B). \tag{158}$$

2. **One–block evaluation.** Using the exponential formula for labelled set partitions, for all $n \geq 1$, one has:

$$\mu_{\Pi([n])}(\hat{0}, \hat{1}) = (-1)^{n-1}(n-1)!. \tag{159}$$

Substituting into the product from Step 1 yields

$$\mu_{\Pi(U)}(\hat{0}, \pi) = \prod_{B \in \pi} (-1)^{|B|-1}(|B|-1)!. \tag{160}$$

Having outlined the strategy, we now provide the full proof with all intermediate steps made explicit.

Proof. We structure the proof into several steps for the sake of clarity and readability.

Step 1 (Interval factorization in the partition lattice).

A partition $\pi \in \Pi(U)$ is a set of disjoint nonempty blocks $B \subseteq U$ covering U. For $\sigma, \pi \in \Pi(U)$ write $\sigma \leq \pi$ if every block of σ is contained in a block of π. For $\sigma \leq \pi$ and a block $B \in \pi$, let $\sigma|_B$ be the restriction of σ to B (intersect each block of σ with B and remove empties). Denote by $\hat{1}_B$ the one-block partition of B. We have the following result.

Lemma E.6 (Interval factorization). *For $\sigma \leq \pi$ in $\Pi(U)$, restriction induces a poset isomorphism*

$$\Phi : [\sigma, \pi] \longrightarrow \prod_{B \in \pi} \Pi\big(\sigma|_B, \hat{1}_B\big), \qquad \Phi(\tau) \colon \big(\tau|_B\big)_{B \in \pi}. \tag{161}$$

Its inverse maps $(\rho_B)_{B\in\pi}$ to the join $\bigvee_{B\in\pi} \rho_B$, which coincides with the partition whose restriction to each B equals ρ_B.

Proof. If $\tau \in [\sigma, \pi]$, then $\sigma \leq \tau \leq \pi$ implies that each block of τ lies inside some block of π, so $\tau|_B$ is a partition of B refining $\sigma|_B$, hence $\sigma|_B \leq \tau|_B \leq \hat{1}_B$. Thus Φ is well-defined and order-preserving. Conversely, if $(\rho_B)_{B\in\pi}$ satisfies $\sigma|_B \leq \rho_B \leq \hat{1}_B$, define ρ by declaring that $x, y \in U$ lie in the same block of ρ iff either $x, y \in B$ and $x \sim_{\rho_B} y$ for some $B \in \pi$, or x, y lie in different blocks of π (which never happens since we work blockwise). Then ρ is a partition with $\sigma \leq \rho \leq \pi$ and $\rho|_B = \rho_B$. One checks $\Phi(\rho) = (\rho_B)$ and $\bigvee_{B\in\pi}(\tau|_B) = \tau$, hence Φ is an isomorphism. □

Setting $\sigma = \hat{0}$ in Lemma E.6 yields

$$[\hat{0}, \pi] \;\cong\; \prod_{B\in\pi} \Pi(B). \tag{162}$$

Applying the multiplicativity Equation (155) to Equation (162), one has

$$\mu_{\Pi(U)}(\hat{0}, \pi) = \prod_{B\in\pi} \mu_{\Pi(B)}(\hat{0}_B, \hat{1}_B). \tag{163}$$

Therefore, to compute $\mu_{\Pi(U)}(\hat{0}, \pi)$ for arbitrary π, it suffices to evaluate the single-block quantity

$$m(n) \coloneqq \mu_{\Pi_n}(\hat{0}, \hat{1}), \tag{164}$$

for $n \in \mathbb{N}$, where Π_n denotes the partition lattice on an n-element set.

Step 2 (The one-block value via the exponential formula for labeled set partitions).

We now determine $m(n)$ exactly. One has a Möbius sum constraint as follows: by Equation (149), for every finite poset and any $x < y$, one has

$$\sum_{x\leq z\leq y} \mu(x, z) = 0. \tag{165}$$

In Π_n, taking $x = \hat{0}$ and $y = \hat{1}$ gives

$$\sum_{\tau\in\Pi_n} \mu_{\Pi_n}(\hat{0}, \tau) = 0, \tag{166}$$

for all $n \geq 2$. For $n = 0, 1$, the sum equals 1 (the unique element of the interval). By Equation (163) applied inside Π_n, one has

$$\mu_{\Pi_n}(\hat{0}, \tau) = \prod_{B \in \tau} m(|B|). \tag{167}$$

Define

$$F_n := \sum_{\tau \in \Pi_n} \prod_{B \in \tau} m(|B|). \tag{168}$$

Then, for $n \geq 2$, one has

$$F_0 = 1, \qquad F_1 = 1, \qquad F_n = 0. \tag{169}$$

A standard labeled-partition identity (the exponential formula) asserts that for any sequence $(a_k)_{k \geq 1}$,

$$\sum_{n \geq 0} \left(\sum_{\tau \in \Pi_n} \prod_{B \in \tau} a_{|B|} \right) \frac{z^n}{n!} = \exp\left(\sum_{k \geq 1} a_k \frac{z^k}{k!} \right). \tag{170}$$

Applying this with $a_k = m(k)$ yields

$$\sum_{n \geq 0} F_n \frac{z^n}{n!} = \exp\left(\sum_{k \geq 1} m(k) \frac{z^k}{k!} \right). \tag{171}$$

Using Equation (169), the left-hand side of Equation (171) equals $1 + z$. Taking the formal logarithm gives

$$\sum_{k \geq 1} m(k) \frac{z^k}{k!} = \log(1 + z) = \sum_{k \geq 1} (-1)^{k-1} \frac{z^k}{k}. \tag{172}$$

Equating coefficients, for $k \geq 1$, one has

$$m(k) = k! \cdot \frac{(-1)^{k-1}}{k} = (-1)^{k-1} (k-1)!. \tag{173}$$

Substituting Equation (173) into the block factorization Equation (163) gives the desired expression in Equation (156):

$$\mu_{\Pi(U)}(\hat{0}, \pi) = \prod_{B \in \pi} (-1)^{|B|-1} (|B| - 1)!. \tag{174}$$

This concludes the proof. □

The identity established in Theorem E.5 plays a pivotal role in the proof of Theorem E.11, which, in turn, functions as a supporting lemma for the proof of Theorem D.1–the main result of this work. To conclude this section, we highlight several connections of this identity with other well-studied combinatorial concepts, including the generalized inclusion–exclusion principle, Stirling and Bell numbers, rook polynomials, and systems of distinct representatives.

Remark E.7 (Generalized inclusion-exclusion principle)**.** For any statistic F on partitions that collapses indices within blocks, Möbius inversion on the partition lattice gives

$$G(\hat{0}) = \sum_{\pi \in \Pi([n])} \mu_{\Pi}(\hat{0}, \pi)\, F(\pi), \tag{175}$$

where

$$\mu_{\Pi}(\hat{0}, \pi) = \prod_{B \in \pi} (-1)^{|B|-1} (|B| - 1)!. \tag{176}$$

A canonical instance is counting injections $f : [n] \to [m]$. Let $F(\pi) = m^{|\pi|}$ be the number of maps constant on each block of π. Then

$$m^{\underline{n}} = \sum_{\pi \in \Pi([n])} \mu_{\Pi}(\hat{0}, \pi)\, m^{|\pi|}, \tag{177}$$

which is the inclusion-exclusion count of injections (no two domain points collide).

Remark E.8 (Stirling Numbers and Bell Numbers)**.** The previous identity specializes to the Stirling transforms:

$$x^{\underline{n}} = \sum_{\pi\in\Pi([n])} \mu_\Pi(\hat{0},\pi)\, x^{|\pi|} = \sum_{k=0}^{n} s(n,k)\, x^k, \qquad s(n,k) := \sum_{\substack{\pi\in\Pi([n])\\ |\pi|=k}} \mu_\Pi(\hat{0},\pi), \tag{178}$$

and its inverse

$$x^n = \sum_{k=0}^{n} S(n,k)\, x^{\underline{k}}, \qquad B_n = \sum_{k=0}^{n} S(n,k), \tag{179}$$

where $S(n,k)$ are Stirling numbers of the second kind and B_n are Bell numbers. As a quick consequence, setting $x = 1$ in the first display yields

$$\sum_{\pi\in\Pi([n])} \mu_\Pi(\hat{0},\pi) = 1^{\underline{n}} = \begin{cases} 1, & n = 1, \\ 0, & n \geq 2, \end{cases} \tag{180}$$

a useful checksum for μ_Π.

Remark E.9 (Rook Polynomials)**.** For a board $B \subseteq [n] \times [n]$, let $r_k(B)$ be the number of ways to place k non-attacking rooks on B and $R_B(t) = \sum_{k=0}^{n} r_k(B)\, t^k$ the rook polynomial. The number of permutations of $[n]$ whose graph avoids B is

$$|\{\sigma \in S_n \,:\, (i,\sigma(i)) \notin B, \forall i\}| = \sum_{k=0}^{n} (-1)^k\, r_k(B)(n-k)! = \big(R_B(-\Delta)n!\big), \tag{181}$$

the standard rook-theoretic inclusion-exclusion.

Remark E.10 (Systems of Distinct Representatives)**.** For systems of distinct representatives, let $A_1, \ldots, A_n \subseteq U$. The number of SDRs $(a_1, \ldots, a_n)$ with $a_i \in A_i$ and a_i all distinct is

$$|\mathrm{SDR}(A_1, \ldots, A_n)| = \sum_{\pi\in\Pi([n])} \mu_\Pi(\hat{0},\pi) \prod_{B\in\pi} \Big| \bigcap_{i\in B} A_i \Big|. \tag{182}$$

Here $\prod_{B\in\pi} | \cap_{i\in B} A_i|$ counts choices when entries are forced equal within each block B; Möbius inversion removes those equalities to enforce distinctness-exactly paralleling rook inclusion-exclusion.

E.4 A Technical Result on Weighted Sums over Distinct Tuples

We now present a result concerning the problem of weighted sums over distinct tuples. The results developed in this section form the backbone of our argument in the proof of Theorem D.1, the main result of this work.

Theorem E.11. *Given positive integers $m, n \geq 1$. For each $i \in [m]$, let A_i be a subset of $[n]$. Let $x_1, \ldots, x_n$ be n real numbers. For any nonempty $S \subseteq [m]$, define*

$$F_S := \Big\{(a_i)_{i\in S} \,:\, a_i \in A_i \text{ for all } i \in S, \text{ and all } a_i\text{'s are pairwise distinct}\Big\}. \tag{183}$$

For $i \in S$ and $a \in A_i$, define the fiber

$$F_{S,i,a} := \{(a_j)_{j\in S} \in F_S \,:\, a_i = a\}. \tag{184}$$

For any nonempty $T \subseteq [m]$, define $A_T := \bigcap_{i\in T} A_i$, and

$$G(T) := \sum_{a\in A_T} x_a. \tag{185}$$

Assume that, for every nonempty $S \subseteq [m]$ and every $i \in S$, one has

$$\sum_{a\in A_i} |F_{S,i,a}|\, x_a = 0. \tag{186}$$

Then, for every nonempty $T \subseteq [m]$, one has

$$G(T) = \sum_{a\in A_T} x_a = 0. \tag{187}$$

Proof. Let S be a nonempty finite set. Denote by $\Pi(S)$ the lattice of set partitions of S ordered by refinement: For $\sigma, \pi \in \Pi(S)$, we write $\sigma \leq \pi$ if every block of σ is contained in a block of π. Any $\pi \in \Pi(S)$ is a family of disjoint nonempty blocks whose union is S. For a block $B \subseteq S$ define

$$A_B := \bigcap_{j \in B} A_j, \qquad \text{and} \qquad |A_B| := \Big| \bigcap_{j \in B} A_j \Big|. \tag{188}$$

Let μ denote the Möbius function of $\Pi(S)$ (with respect to refinement). μ is determined by $\sum_{\sigma:\, \sigma \leq \pi} \mu(\sigma) = \mathbf{1}_{\{\pi = \hat{0}\}}$, where $\hat{0}$ is the discrete partition. Formula (1) follows by multiplicativity of μ over blocks and the known one-block value $(-1)^{r-1}(r-1)!$ for a block of size r. It is well-known that:

$$\mu(\pi) = \prod_{B \in \pi} (-1)^{|B|-1} \, (|B| - 1)!. \tag{189}$$

Fix a nonempty $S \subseteq [m]$, an index $i \in S$, and an element $a \in [n]$. Let $\mathcal{G}_S$ be the set of all functions $g : S \to [n]$ satisfying $g(j) \in A_j$ for all $j \in S$ (note that, there is no distinctness condition). For $g \in \mathcal{G}_S$, define its equality partition $\pi(g) \in \Pi(S)$ by:

$$j \sim_{\pi(g)} k \quad \text{if and only if} \quad g(j) = g(k). \tag{190}$$

Thus $\pi(g)$ records which indices are assigned the same value by g. One has g is injective on S if and only if $\pi(g) = \hat{0}$. The set F_S of injective choices can be described as:

$$F_S = \Big\{ g \in \mathcal{G}_S : \ \pi(g) = \hat{0} \Big\}, \tag{191}$$

and the *fiber* fixing the value at the distinguished index i is:

$$F_{S,i,a} = \Big\{ g \in \mathcal{G}_S : \ g(i) = a, \ \pi(g) = \hat{0} \Big\}. \tag{192}$$

For $\pi \in \Pi(S)$ and $i \in S$, let $B_i(\pi)$ denote the unique block of π containing i. Define:

$$N_{S,i,a}(\pi) := \Big| \Big\{ g \in \mathcal{G}_S : \ g \text{ is constant on each block of } \pi, \ g(i) = a \Big\} \Big|. \tag{193}$$

That is, $N_{S,i,a}(\pi)$ counts maps that are constant along blocks of π (so the only equalities allowed among coordinates are those forced by π) and take the prescribed value a at the index i. For every $\pi \in \Pi(S)$, one has:

$$N_{S,i,a}(\pi) = \mathbf{1}_{\{a \in A_{B_i(\pi)}\}} \prod_{\substack{B \in \pi \\ B \neq B_i(\pi)}} |A_B|. \tag{194}$$

Indeed, if g is constant on each block of π, the value on the block $B_i(\pi)$ must equal $g(i) = a$. This is possible exactly when $a \in \bigcap_{j \in B_i(\pi)} A_j = A_{B_i(\pi)}$, which contributes the indicator $\mathbf{1}_{\{a \in A_{B_i(\pi)}\}}$. Then, for any other block $B \in \pi$ with $B \neq B_i(\pi)$, the common value of g on B can be chosen arbitrarily from the intersection $A_B = \bigcap_{j \in B} A_j$, independently across distinct blocks. Therefore there are $|A_B|$ choices for each such block, and multiplying over all $B \neq B_i(\pi)$ yields the product in Equation (194). Now, for $g \in \mathcal{G}_S$, define the two indicator functions on $\Pi(S)$:

$$E(g) := \mathbf{1}_{\{\pi(g) = \hat{0}\}}, \text{ and } \quad C_\pi(g) := \mathbf{1}_{\{\pi(g) \geq \pi\}} \quad (\pi \in \Pi(S)). \tag{195}$$

Here $\pi(g) \geq \pi$ means that g is constant on every block of π. By general Möbius inversion on posets, one has:

$$E(g) \;=\; \sum_{\pi \in \Pi(S)} \mu(\pi) \, C_\pi(g), \tag{196}$$

since

$$\sum_{\sigma \leq \pi(g)} \mu(\sigma) = \mathbf{1}_{\{\pi(g) = \hat{0}\}}. \tag{197}$$

Now fix $i \in S$ and $a \in [n]$, multiply the last identity by $\mathbf{1}_{\{g(i) = a\}}$, and sum over all $g \in \mathcal{G}_S$, one has:

$$\big| F_{S,i,a} \big| = \sum_{g \in \mathcal{G}_S} \mathbf{1}_{\{g(i) = a\}} E(g) = \sum_{\pi \in \Pi(S)} \mu(\pi) \sum_{g \in \mathcal{G}_S} \mathbf{1}_{\{g(i) = a\}} C_\pi(g). \tag{198}$$

The inner sum is precisely $N_{S,i,a}(\pi)$ by definition. Using Equation (194), one therefore obtains the explicit expansion:

$$|F_{S,i,a}| = \sum_{\pi\in\Pi(S)} \mu(\pi)\,\mathbf{1}_{\{a\in A_{B_i(\pi)}\}} \prod_{\substack{B\in\pi\\ B\neq B_i(\pi)}} |A_B|. \tag{199}$$

Multiply Equation (199) by x_a and sum over all $a \in A_i$ (equivalently, over all $a \in [n]$, since the indicator in Equation (199) already forces $a \in A_i$ when $i \in B_i(\pi)$):

$$\sum_{a\in A_i} |F_{S,i,a}|\, x_a = \sum_{\pi\in\Pi(S)} \mu(\pi) \left(\prod_{\substack{B\in\pi\\ B\neq B_i(\pi)}} |A_B| \right) \left(\sum_{a\in A_{B_i(\pi)}} x_a \right). \tag{200}$$

With the shorthand $G(T) := \sum_{a\in A_T} x_a$ this becomes

$$\sum_{a\in A_i} |F_{S,i,a}|\, x_a = \sum_{\pi\in\Pi(S)} \mu(\pi) \left(\prod_{\substack{B\in\pi\\ B\neq B_i(\pi)}} |A_B| \right) G\big(B_i(\pi)\big). \tag{201}$$

By the hypothesis, the left–hand side of Equation (201) is 0. Hence

$$0 = \sum_{\pi\in\Pi(S)} \mu(\pi) \left(\prod_{\substack{B\in\pi\\ B\neq B_i(\pi)}} |A_B| \right) G\big(B_i(\pi)\big), \tag{202}$$

for every nonempty $S \subseteq [m]$ and every $i \in S$. Observe that, in Equation (202), the term $G\big(B_i(\pi)\big)$ only involves nonempty subsets $B_i(\pi)$ with $i \in B_i(\pi) \subseteq S$.

Back to the problem. We now show that $G(T) = 0$ for every nonempty $T \subseteq [m]$ by induction on $k := |T|$. We use the Equation (189) and Equation (202) a lots.

Base case.

Let $T = \{i\}$ for some $i \in [m]$. Take $S = \{i\}$ in the given hypothesis, one has

$$\sum_{a\in A_i} |F_{S,i,a}| x_a = 0. \tag{203}$$

Since S has one element, an injective choice on S is just a choice of a value in A_i, hence $|F_{\{i\},i,a}| = \mathbf{1}_{\{a\in A_i\}}$. Therefore

$$0 = \sum_{a\in A_i} |F_{\{i\},i,a}| x_a = \sum_{a\in A_i} x_a = G(\{i\}), \tag{204}$$

which establishes the base case.

Inductive step.

Fix $k \geq 2$ and assume the claim holds for all nonempty $U \subseteq [m]$ with $|U| < k$, i.e., $G(U) = 0$ whenever $1 \leq |U| \leq k-1$. Let $T \subseteq [m]$ with $|T| = k$, and fix any distinguished index $i \in T$. Apply Equation (202) with $S = T$, we analyze the sum over $\pi \in \Pi(T)$ by separating the one–block partition from the rest.

(a) The contribution of the one–block partition.

There is a unique partition $\pi^\star = \{T\}$ with a single block. For this partition we have $B_i(\pi^\star) = T$, and the product over $B \neq B_i(\pi^\star)$ is an empty product, hence equals 1 by convention. By Equation (189) with $|T| = k$, one has:

$$\mu(\pi^\star) = (-1)^{k-1}(k-1)!. \tag{205}$$

Thus, the term of Equation (202) corresponding to $\pi^\star$ equals

$$\mu(\pi^\star) \cdot 1 \cdot G\big(B_i(\pi^\star)\big) = (-1)^{k-1}(k-1)!G(T). \tag{206}$$

(b) The contribution of all other partitions.

Let $\pi \in \Pi(T)$ with $\pi \neq \pi^\star$. Then $B_i(\pi)$ is a proper, nonempty subset of T (it still contains i but does not equal T). Consequently $|B_i(\pi)| \le k-1$. By the inductive hypothesis,

$$G\big(B_i(\pi)\big) = 0.$$

Hence every summand in Equation (202) with $\pi \neq \pi^\star$ vanishes, regardless of the multiplicative factor $\prod_{B \neq B_i(\pi)} |A_B|$ and the value of $\mu(\pi)$.

Collecting (a) and (b), identity Equation (202) with $S = T$ reduces to

$$0 = (-1)^{k-1}(k-1)!G(T). \tag{207}$$

Since $(-1)^{k-1}(k-1)! \neq 0$, we conclude $G(T) = 0$.

By induction on k, the relation $G(T) = 0$ holds for every nonempty $T \subseteq [m]$. □

Remark E.12 (Combinatorial intuition)**.** Viewed combinatorially, F_S is precisely the set of systems of distinct representatives (SDRs) for the family $\{A_i : i \in S\}$ For a fixed index $i \in S$ and value $a \in A_i$, the fiber $F_{S,i,a}$ enumerates those SDRs that assign the representative a to position i. Assumption in Equation (210) therefore states that the weighted sum $\sum_{a \in A_i} |F_{S,i,a}|\, x_a$ vanishes for every nonempty $S \subseteq [m]$ and every $i \in S$; equivalently, the vector $x = (x_a)_{a \in [n]}$ is orthogonal to the vector of SDR–completion counts at coordinate i. Applying Möbius inversion on the Boolean lattice $(2^{[m]}, \subseteq)$ transfers these linear relations, with coefficients given by SDR multiplicities, into relations with unit coefficients, thereby collapsing the fiber-weighted sums to the unweighted intersection sums $\sum_{a \in \cap_{j \in T} A_j} x_a$. This mirrors the classical rook-polynomial/inclusion-exclusion paradigm: counts of placements with multiplicities invert to simple intersection counts once the incidence algebra is diagonalized by the Möbius function.

Remark E.13 (Discussion and structural role of Equation (199))**.** Equation (199) is a blockwise inclusion-exclusion (rook-theoretic) decomposition of the fiber cardinality. Each partition π encodes a pattern of equalities among coordinates; the Möbius weight $\mu(\pi)$ alternates and corrects for overcounting, while the multiplicative factor $\prod_{B \neq B_i(\pi)} |A_B|$ captures the independent choices of common values on the remaining blocks, and the indicator $\mathbf{1}_{\{a \in A_{B_i(\pi)}\}}$ enforces compatibility of the distinguished value a with the block containing i.

We have a direct corollary of Theorem E.11.

Corollary E.14. *Given positive integers $m, n \ge 1$. For each $i \in [m]$, let A_i be a subset of $[n]$. Let $x_1, \dots, x_n$ be n real numbers. For any nonempty $S \subseteq [m]$, define*

$$F_S := \Big\{(a_i)_{i \in S} :\ a_i \in A_i \ \textit{for all } i \in S,\ \textit{and all } a_i\textit{'s are pairwise distinct}\Big\}. \tag{208}$$

For $i \in S$ and $a \in A_i$, define the fiber

$$F_{S,i,a} := \{(a_j)_{j \in S} \in F_S :\ a_i = a\}\,. \tag{209}$$

Assume that, for every nonempty $S \subseteq [m]$ and every $i \in S$, one has

$$\sum_{a \in A_i} |F_{S,i,a}|\, x_a = 0. \tag{210}$$

Then, one has

$$G(T) = \sum_{a \in A_1 \cap \ldots \cap A_m} x_a = 0. \tag{211}$$

Proof. By taking $T = [m]$ in Theorem E.11, one obtains the asserted main conclusion. □

F Functional Equivalence of Multihead Attention with Rotary Positional Encoding

F.1 Main Result on Functional Equivalence

Theorem F.1 (Theorem 4.2 in the main paper)**.** *Given two positive integers d and d_h with $d > d_h$. Consider two* $\text{MultiHead}_{\text{RoPE}}$ *maps with h and $\bar{h}$ heads, with rotary positional encoding. They are parameterized by families of matrices*

$$\begin{aligned} &\{W_i^Q\}_{i=1}^h, \quad \{W_i^K\}_{i=1}^h, \quad \{W_i^V\}_{i=1}^h, \quad \{W_i^O\}_{i=1}^h \subset \mathbb{R}^{d\times d_h}, \text{ and} \\ &\{\bar{W}_i^Q\}_{i=1}^{\bar{h}}, \quad \{\bar{W}_i^Q\}_{i=1}^{\bar{h}}, \quad \{\bar{W}_i^V\}_{i=1}^{\bar{h}}, \quad \{\bar{W}_i^O\}_{i=1}^{\bar{h}} \subset \mathbb{R}^{d\times d_h}, \end{aligned} \tag{212}$$

respectively. Assume that

1. *In the first* $\text{MultiHead}_{\text{RoPE}}$ *map, for each head $i \in [h]$, the similarity score between two arbitrary tokens does not vanish, i.e.,*
$$W_i^Q(W_i^K)^\top + W_i^K(W_i^Q)^\top \text{ and } W_i^Q R^n (W_i^K)^\top, \tag{213}$$
for all non-zero integer n, are non-zero.

2. *In the second* $\text{MultiHead}_{\text{RoPE}}$ *map, for each head $i \in [\bar{h}]$, the similarity score between two arbitrary tokens does not vanish, i.e.,*
$$\bar{W}_i^Q(\bar{W}_i^K)^\top + \bar{W}_i^K(\bar{W}_i^Q)^\top \text{ and } \bar{W}_i^Q R^n (\bar{W}_i^K)^\top, \tag{214}$$
for all non-zero integer n, are non-zero.

3. *In the first* $\text{MultiHead}_{\text{RoPE}}$ *map, the similarity score maps are pairwise distinct, i.e.,*
$$\Big\{W_i^Q(W_i^K)^\top + W_i^K(W_i^Q)^\top, \{W_i^Q R^n (W_i^K)^\top\}_{n\in\mathbb{Z}, n\neq 0}\Big\}, \tag{215}$$
for $i = 1, \ldots, h$, are h pairwise distinct families.

4. *In the second* $\text{MultiHead}_{\text{RoPE}}$ *map, the similarity score maps are pairwise distinct, i.e.,*
$$\Big\{\bar{W}_i^Q(\bar{W}_i^K)^\top + \bar{W}_i^K(\bar{W}_i^Q)^\top, \{\bar{W}_i^Q R^n (\bar{W}_i^K)^\top\}_{n\in\mathbb{Z}, n\neq 0}\Big\}, \tag{216}$$
for $i = 1, \ldots, \bar{h}$, are h pairwise distinct families.

5. *In the first* $\text{MultiHead}_{\text{RoPE}}$ *map, all matrices $W_i^Q, W_i^K, W_i^V, W_i^O$ for $i \in [h]$ are of rank d_h.*

6. *In the second* $\text{MultiHead}_{\text{RoPE}}$ *map, all matrices $\bar{W}_i^Q, \bar{W}_i^K, \bar{W}_i^V, W_i^O$ for $i \in [h]$ are of rank d_h.*

If the two $\text{MultiHead}_{\text{RoPE}}$ *maps are identical, i.e.,*

$$\begin{aligned} &\text{MultiHead}_{\text{RoPE}}\Big(\mathbf{x} : \{W_i^Q, W_i^K, W_i^V, W_i^O\}_{i=1}^h\Big) \\ &\qquad\qquad = \text{MultiHead}_{\text{RoPE}}\Big(\mathbf{x} : \{\bar{W}_i^Q, \bar{W}_i^Q, \bar{W}_i^V, \bar{W}_i^O\}_{i=1}^{\bar{h}}\Big), \end{aligned} \tag{217}$$

then $h = \bar{h}$, and there exists a permutation $\sigma \in S_h$ and invertible matrices $\{U_i\}_{i=1}^h \subset \mathrm{H}(d_h)$ and $\{V_i\}_{i=1}^h \subset \mathrm{GL}(d_h)$ such that

$$\begin{aligned} \bar{W}_i^Q &= W_{\sigma(i)}^Q \cdot U_i^\top, \quad \bar{W}_i^K = W_{\sigma(i)}^K \cdot (U_i)^{-1}, \\ \bar{W}_i^V &= W_{\sigma(i)}^V \cdot V_i^\top, \quad \bar{W}_i^O = W_{\sigma(i)}^O \cdot (V_i)^{-1}. \end{aligned} \tag{218}$$

Proof. For $i \in [h]$ and $m, n \geq 1$, denote

$$A_i^{m,n} := W_i^Q R^{m-n} (W_i^K)^\top, \text{ if } m \neq n \tag{219}$$

$$A_i^{m,m} := \frac{W_i^Q (W_i^K)^\top + W_i^K (W_i^Q)^\top}{2}, \text{ and} \tag{220}$$

$$B_i := W_i^V (W_i^O)^\top. \tag{221}$$

For $i \in [\bar{h}]$ and $m, n \geq 1$, denote

$$\bar{A}_i^{m,n} := \bar{W}_i^Q R^{m-n} (\bar{W}_i^K)^\top, \text{ if } m \neq n \tag{222}$$

$$\bar{A}_i^{m,m} := \frac{\bar{W}_i^Q (\bar{W}_i^K)^\top + \bar{W}_i^K (\bar{W}_i^Q)^\top}{2}, \text{ and} \tag{223}$$

$$\bar{B}_i := \bar{W}_i^V (\bar{W}_i^O)^\top. \tag{224}$$

Then, one has

$$\begin{aligned} &\text{MultiHead}\Big(\mathbf{x} : \ \{\{A_i^{m,n}\}_{m,n}, B_i\}_{i=1}^h\Big) \\ &\qquad = \text{MultiHead}_{\text{RoPE}}\Big(\mathbf{x} : \ \{W_i^Q, W_i^K, W_i^V, W_i^O\}_{i=1}^h\Big), \end{aligned} \tag{225}$$

and

$$\begin{aligned} &\text{MultiHead}\Big(\mathbf{x} : \ \{\{\bar{A}_i^{m,n}\}_{m,n}, \bar{B}_i\}_{i=1}^{\bar{h}}\Big) \\ &\qquad = \text{MultiHead}_{\text{RoPE}}\Big(\mathbf{x} : \ \{\bar{W}_i^Q, \bar{W}_i^Q, \bar{W}_i^V, \bar{W}_i^O\}_{i=1}^{\bar{h}}\Big). \end{aligned} \tag{226}$$

Thus,

$$\begin{aligned} &\text{MultiHead}\Big(\mathbf{x} : \ \{\{A_i^{m,n}\}_{m,n}, B_i\}_{i=1}^h\Big) \\ &\qquad = \text{MultiHead}\Big(\mathbf{x} : \ \{\{\bar{A}_i^{m,n}\}_{m,n}, \bar{B}_i\}_{i=1}^{\bar{h}}\Big). \end{aligned} \tag{227}$$

From the condition 3, 4, the property of parameters from these maps fit to the setting of Corollary D.3, which is that $A_i^{m,n}$ and $\bar{A}_i^{m,n}$ are nonzero for all feasible triples (i, m, n). Thus, for every parameter family

$$\{A^{m,n}\}_{m,n} \subset \mathbb{R}^{d \times d}, \tag{228}$$

we have the identity

$$\sum_{i \in [h] \, : \, \{A_i^{m,n}\}_{m,n} = \{A^{m,n}\}_{m,n}} B_i = \sum_{i \in [\bar{h}] \, : \, \{\bar{A}_i^{m,n}\}_{m,n} = \{A^{m,n}\}_{m,n}} \bar{B}_i. \tag{229}$$

From condition 3, one has h families of parameters

$$\{A_1^{m,n}\}_{m,n\geq 1}, \{A_2^{m,n}\}_{m,n\geq 1}, \ldots, \{A_h^{m,n}\}_{m,n\geq 1}, \tag{230}$$

are pairwise distinct. Together with Equation (229), consider

$$\{A^{m,n}\}_{m,n} = \{A_i^{m,n}\}_{m,n}, \tag{231}$$

one has the left-hand side of Equation (229) is equal to B_i. Thus,

$$B_i = \sum_{j \in [\bar{h}] \, : \, \{\bar{A}_j^{m,n}\}_{m,n} = \{A_i^{m,n}\}_{m,n}} \bar{B}_j. \tag{232}$$

Note that, since all the matrices W_i^V and W_i^O have rank d_h, it implies that all B_i are non-zero. From Equation (232), for each $i \in [h]$, since the left-hand side is non-zero, the right-hand side has at least one index $j \in [\bar{h}]$ such that $\bar{B}_j$ is non-zero and $\{\bar{A}_j^{m,n}\}_{m,n} = \{A_i^{m,n}\}_{m,n}$. Since h families of parameters

$$\{A_1^{m,n}\}_{m,n\geq 1}, \{A_2^{m,n}\}_{m,n\geq 1}, \ldots, \{A_h^{m,n}\}_{m,n\geq 1}, \tag{233}$$

are pairwise distinct, one implies that each i has its corresponding j's distinctly from others. Thus, $h \le \bar{h}$. By a symmetric argument, one also has $h \ge \bar{h}$. In conclusion, one has $h = \bar{h}$. Moreover, by the above argument, for each i, there exists exactly one $j \in [h]$ such that $\{\bar{A}_j^{m,n}\}_{m,n} = \{A_i^{m,n}\}_{m,n}$. Moreover, this also implies that $B_j = B_i$.

In conclusion, there exists a permutation $\sigma \in S_h$ such that

$$\bar{A}_i^{m,n} = A_{\sigma(i)}^{m,n}, \quad \text{for all } m, n \ge 1, \quad \text{and} \quad \bar{B}_{\sigma(i)} = B_i. \tag{234}$$

From Lemma F.2, there exists matrices $\{U_i\}_{i=1}^h \subset \mathrm{H}(d_h)$ such that

$$\bar{W}_i^Q = W_{\sigma(i)}^Q \cdot U_i^\top, \qquad \bar{W}_i^K = W_{\sigma(i)}^K \cdot (U_i)^{-1}. \tag{235}$$

From the rank factorization (Piziak & Odell, 1999), there exists matrices $\{V_i\}_{i=1}^h \subset \mathrm{GL}(d_h)$ such that

$$\bar{W}_i^V = W_{\sigma(i)}^V \cdot V_i^\top, \qquad \bar{W}_i^O = W_{\sigma(i)}^O \cdot (V_i)^{-1}. \tag{236}$$

This concludes the proof. □

F.2 A Lemma Concerning the Rotary Matrix

Given $d = 2m$ be an even integer. Consider the RoPE matrix at position 1 as

$$R = \mathrm{diag}\big(R(\theta_1), \ldots, R(\theta_{d/2})\big) \in \mathbb{R}^{d\times d}, \quad \text{where } R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{237}$$

Denote the $n \times n$ identity matrix as I_n. For $i = 1, \ldots, m$, define the 2-dimensional coordinate plane

$$E_i := \mathrm{span}\{e_{2i-1}, e_{2i}\} \subset \mathbb{R}^d, \tag{238}$$

where e_{2i-1}, e_{2i} are the $(2i-1)$-th and $2i$-th coordinate basis vectors. Definethe orthogonal projection matrix

$$P_i := e_{2i-1}e_{2i-1}^\top + e_{2i}e_{2i}^\top \in \mathbb{R}^{d\times d}. \tag{239}$$

In words, P_i is the $d \times d$ matrix has the i-th 2×2 diagonal block is the 2×2 identity matrix. We also define the matrix

$$J_i := e_{2i}e_{2i-1}^\top - e_{2i-1}e_{2i}^\top \in \mathbb{R}^{d\times d}. \tag{240}$$

In words, J_i is the $d \times d$ matrix has the i-th 2×2 diagonal block is the following 2×2 matrix

$$J := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}. \tag{241}$$

The matrix R now can be written as

$$R = \sum_{i=1}^m \left(\cos\theta_i P_i + \sin\theta_i J_i\right). \tag{242}$$

Moreover, for $n \in \mathbb{Z}$, one has

$$R^n = \sum_{i=1}^m \left(\cos(n\theta_i) P_i + \sin(n\theta_i) J_i\right). \tag{243}$$

We have the following result.

Lemma F.2. *Given an integer $D \ge d$. Consider matrices $X, Z \in \mathbb{R}^{D\times d}$ and $Y, T \in \mathbb{R}^{d\times D}$. Assume that, for all non zero interger n,*

$$XR^nY = ZR^nT. \tag{244}$$

If

1. *All the angles $\theta_i \in (0, \pi)$ are pairwise distinct, and*

2. *For all $i = 1, \dots, m$, XP_i and P_iY have rank 2.*

Then, there exists an invertible matrix $U \in \mathbb{R}^{d\times d}$ of the form

$$U = \sum_{i=1}^{m}(a_iP_i + b_iJ_i) \ \text{ with } \ (a_i, b_i) \in \mathbb{R}^2 \setminus \{(0,0)\} \text{ for } \ i = 1, \dots, m, \tag{245}$$

such that

$$Z = XU \qquad \text{and} \qquad T = U^{-1}Y. \tag{246}$$

Proof. We structure the proof into several steps for the sake of clarity and readability.

Step 1.

Define

$$A_{1,i} \coloneqq XP_iY \in \mathbb{R}^{D\times D}, \tag{247}$$
$$B_{1,i} \coloneqq XJ_iY \in \mathbb{R}^{D\times D}, \tag{248}$$
$$A_{2,i} \coloneqq ZP_iT \in \mathbb{R}^{D\times D}, \tag{249}$$
$$B_{2,i} \coloneqq ZJ_iT \in \mathbb{R}^{D\times D}. \tag{250}$$

Using

$$R^n = \sum_{i=1}^{m}\left(\cos(n\theta_i)P_i + \sin(n\theta_i)J_i\right), \tag{251}$$

one has

$$\begin{aligned} XR^nY &= \sum_{i=1}^{m} X\left(\cos(n\theta_i)P_i + \sin(n\theta_i)J_i\right)Y \\ &= \sum_{i=1}^{m}\left(\cos(n\theta_i)XP_iY + \sin(n\theta_i)XJ_iY\right) \\ &= \sum_{i=1}^{m}\left(\cos(n\theta_i)A_{1,i} + \sin(n\theta_i)B_{1,i}\right), \end{aligned} \tag{252}$$

and

$$\begin{aligned} ZR^nT &= \sum_{i=1}^{m} Z\left(\cos(n\theta_i)P_i + \sin(n\theta_i)J_i\right)T \\ &= \sum_{i=1}^{m}\left(\cos(n\theta_i)ZP_iT + \sin(n\theta_i)ZJ_iT\right) \\ &= \sum_{i=1}^{m}\left(\cos(n\theta_i)A_{2,i} + \sin(n\theta_i)B_{2,i}\right). \end{aligned} \tag{253}$$

Since $XR^nY = ZR^nT$ for all $n \neq 0$, and $\theta_1, \theta_2, \dots, \theta_m$ are pairwise distinct, one has $A_{1,i} = A_{2,i}$ and $B_{1,i} = B_{2,i}$ for all $i = 1, \dots, m$, or

$$XP_iY = ZP_iT, \qquad \text{and} \qquad XJ_iY = ZJ_iT. \tag{254}$$

Step 2.

Now fix an number $i \in \{1, \dots, m\}$. Let X_i is the $D\times 2$ matrix constructed by concating the $(2i-1)$-th and $2i$-th columns of X, Y_i be the $2 \times D$ matrix constructed by concating the $(2i-1)$-th and $2i$-th rows of Y. Similarly, we construct Z_i, T_i for Z, T, respectively. By the second assumption, we have both X_i and Y_i have rank 2. Moreover, from

$$XP_iY = ZP_iT, \qquad \text{and} \qquad XJ_iY = ZJ_iT, \tag{255}$$

one has

$$X_iY_i = Z_iT_i, \qquad \text{and} \qquad X_iJY_i = Z_iJT_i. \tag{256}$$

Let $V_X \in \mathbb{R}^{2\times D}$ be the left inverse matrix of X_i and $V_Y \in \mathbb{R}^{D\times 2}$ be the right inverse matrix of Y_i,

$$V_XX_i = Y_iV_Y = I_2. \tag{257}$$

One has

$$\begin{aligned} I_2 = (V_XX_i)(Y_iV_Y) &= V_X(X_iY_i)V_Y \\ &= V_X(Z_iT_i)V_Y = (V_XZ_i)(T_iV_Y). \end{aligned} \tag{258}$$

Let $U_i = V_XZ_i$. Then $U_i^{-1} = T_iV_Y$. Moreover, one has

$$\begin{aligned} X_i = X_i(Y_iV_Y) &= (X_iY_i)V_Y \\ &= (Z_iT_i)V_Y = Z_i(T_iV_Y) = Z_iU_i^{-1}, \end{aligned} \tag{259}$$

so $Z_i = X_iU_i$. Similarly, one has

$$\begin{aligned} Y_i = (V_XX_i)Y_i &= V_X(X_iY_i) \\ &= V_X(Z_iT_i) = (V_XZ_i)T_i = U_iT_i, \end{aligned} \tag{260}$$

so $T_i = U_i^{-1}Y_i$. Now, from $X_iJY_i = Z_iJT_i$, one has

$$\begin{aligned} J = (V_XX_i)J(Y_iV_Y) &= V_X(X_iJY_i)V_Y \\ &= V_X(Z_iJT_i)V_Y = (V_XZ_i)J(T_iV_Y) = U_iJU_i^{-1}. \end{aligned} \tag{261}$$

In other words, one has $U_iJ = JU_i$. Then, there exists $(a_i, b_i) \in \mathbb{R}^2 \setminus \{(0,0)\}$ such that $U_i = a_iI_2 + b_iJ$. In conclusion, one has

$$Z_i = X_iU_i, \qquad \text{and} \qquad T_i = U_i^{-1}Y_i, \tag{262}$$

where $U_i = a_iI_2 + b_iJ$ with $(a_i, b_i) \in \mathbb{R}^2 \setminus \{(0,0)\}$.

Step 3.

Define $U = \text{diag}(U_1, \dots, U_m)$. From the property of U_i's, we have

$$U = \sum_{i=1}^{m}(a_iP_i + b_iJ_i) \;\text{ with }\; (a_i, b_i) \in \mathbb{R}^2 \setminus \{(0,0)\} \;\text{ for }\; i = 1, \dots, m, \tag{263}$$

and $Z = XU$ and $T = U^{-1}Y$. This concludes the proof. □

This result will be invoked in the proof of Theorem F.1.

Remark F.3 (On the assumptions of Lemma F.2)**.** If angles are not distinct or some equal 0 or π, first merge blocks with equal θ and repeat the argument within each frequency class; the conclusion remains that U must commute with R (hence with each J_i) on the active subspaces. If $\operatorname{rank}(XP_i) < 2$ or $\operatorname{rank}(P_iY) < 2$ for some i, the same derivation shows C_i must commute with J_i on the image subspace; C_i may be non-unique, but the global relation $Z = XU$, $T = U^{-1}Y$ with U commuting with R still describes the solution set restricted to the active coordinates.

Remark F.4 (Concrete matrix forms)**.** We provide the explicit form of the matrices used in the above argument for the case $d = 6$ (i.e., $m = 3$), expressed in the standard basis $(e_1, \dots, e_6)$, to facilitate readability.

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad J_1 = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$P_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad J_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$P_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \qquad J_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

$$R = \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 & 0 & 0 & 0 \\ \sin\theta_1 & \cos\theta_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos\theta_2 & -\sin\theta_2 & 0 & 0 \\ 0 & 0 & \sin\theta_2 & \cos\theta_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cos\theta_3 & -\sin\theta_3 \\ 0 & 0 & 0 & 0 & \sin\theta_3 & \cos\theta_3 \end{bmatrix},$$

$$U = \begin{bmatrix} a_1 & -b_1 & 0 & 0 & 0 & 0 \\ b_1 & a_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_2 & -b_2 & 0 & 0 \\ 0 & 0 & b_2 & a_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_3 & -b_3 \\ 0 & 0 & 0 & 0 & b_3 & a_3 \end{bmatrix},$$

$$U^{-1} = \begin{bmatrix} \frac{a_1}{a_1^2+b_1^2} & \frac{b_1}{a_1^2+b_1^2} & 0 & 0 & 0 & 0 \\ -\frac{b_1}{a_1^2+b_1^2} & \frac{a_1}{a_1^2+b_1^2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{a_2}{a_2^2+b_2^2} & \frac{b_2}{a_2^2+b_2^2} & 0 & 0 \\ 0 & 0 & -\frac{b_2}{a_2^2+b_2^2} & \frac{a_2}{a_2^2+b_2^2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{a_3}{a_3^2+b_3^2} & \frac{b_3}{a_3^2+b_3^2} \\ 0 & 0 & 0 & 0 & -\frac{b_3}{a_3^2+b_3^2} & \frac{a_3}{a_3^2+b_3^2} \end{bmatrix}.$$

G Supplementary Details on the Matching Algorithm

G.1 Lemmas and Proofs for the Algorithm

Lemma G.1. *Given matrices $X, X', Y, Y' \in \mathbb{R}^{m \times n}$ with $m \geq n$, find a matrix $A \in GL(n)$ that minimizes the following objective function:*

$$f(A) = \|X - X'A^{\top}\|_F^2 + \|Y - Y'A^{-1}\|_F^2, \tag{264}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $GL(n)$ is the general linear group of invertible $n \times n$ matrices. Then $\frac{\partial f}{\partial A}$, the gradient w.r.t A, is:

$$2(AX'^{\top} - X^{\top})X' + 2(A^{-1})^T Y'^T (Y - Y'A^{-1})(A^{-1})^T. \tag{265}$$

Proof. We adopt the matrix convention where the gradient is represented as a column vector. We aim to find an invertible matrix $A \in \mathrm{GL}(n)$ that minimizes the objective function

$$f(A) = \|X - X'A^{\top}\|_F^2 + \|Y - Y'A^{-1}\|_F^2, \tag{266}$$

where $X, X', Y, Y' \in \mathbb{R}^{m\times n}$ with $m \geq n$, and $\|\cdot\|_F$ denotes the Frobenius norm.

Step 1. Express the Objective Function Using the Trace

Since the Frobenius norm satisfies $\|M\|_F^2 = \operatorname{trace}(M^T M)$, we can write

$$\begin{aligned} f(A) &= \|X - X'A^\top\|_F^2 + \|Y - Y'A^{-1}\|_F^2 \\ &= \operatorname{trace}\left((X - X'A^\top)^\top (X - X'A^\top)\right) \\ &\quad + \operatorname{trace}\left((Y - Y'A^{-1})^\top (Y - Y'A^{-1})\right). \end{aligned} \tag{267}$$

Step 2. Compute the Gradient

To find the minimum, we compute the gradient of $f(A)$ with respect to A and set it to zero. Define $g_1(A) = \|X - X'A^\top\|_F^2$ and $g_2(A) = \|Y - Y'A^{-1}\|_F^2$, so $f(A) = g_1(A) + g_2(A)$.

To compute the gradient of $g_1(A) = \|X - X'A^\top\|_F^2$ with respect to A, we first expand the expression using the trace property $\|M\|_F^2 = \operatorname{trace}(M^\top M)$:

$$\begin{aligned} g_1(A) &= \operatorname{trace}\left((X - X'A^\top)^\top (X - X'A^\top)\right) \\ &= \operatorname{trace}\left((X^\top - AX'^\top)(X - X'A^\top)\right) \\ &= \operatorname{trace}\left(X^\top X - X^\top X'A^\top - AX'^\top X + AX'^\top X'A^\top\right) \\ &= \operatorname{trace}(X^\top X) - 2\operatorname{trace}(X^\top X'A^\top) + \operatorname{trace}(AX'^\top X'A^\top). \end{aligned} \tag{268}$$

Now, we compute the gradient of each term with respect to A:

$$\begin{aligned} \frac{\partial}{\partial A}\operatorname{trace}(X^\top X) &= 0, \\ \frac{\partial}{\partial A}(-2\operatorname{trace}(X^\top X'A^\top)) &= -2X^\top X', \\ \frac{\partial}{\partial A}\operatorname{trace}(AX'^\top X'A^\top) &= 2AX'^\top X'. \end{aligned} \tag{269}$$

Summing these results, we obtain the gradient of $g_1(A)$:

$$\begin{aligned} \frac{\partial g_1(A)}{\partial A} &= 0 - 2X^\top X' + 2AX'^\top X' \\ &= 2(AX'^\top - X^\top)X' \end{aligned} \tag{270}$$

For the second term $g_2(A)$, since it involves A^{-1}, we use the differential. Note that $d(A^{-1}) = -A^{-1}dAA^{-1}$. The differential of $g_2(A)$ is

$$\begin{aligned} dg_2 &= d\left[\operatorname{trace}\left((Y - Y'A^{-1})^T (Y - Y'A^{-1})\right)\right] \\ &= 2\operatorname{trace}\left((Y - Y'A^{-1})^T d(Y - Y'A^{-1})\right) \\ &= 2\operatorname{trace}\left((Y - Y'A^{-1})^T Y'(A^{-1}dAA^{-1})\right) \\ &= 2\operatorname{trace}\left((Y - Y'A^{-1})^T Y'A^{-1}dAA^{-1}\right). \end{aligned} \tag{271}$$

Using the cyclic property of the trace, $\operatorname{trace}(PQRS) = \operatorname{trace}(SPQR)$, we adjust the expression:

$$\begin{aligned} dg_2 &= 2\operatorname{trace}\left((Y - Y'A^{-1})^T Y'A^{-1}dAA^{-1}\right) \\ &= 2\operatorname{trace}\left(A^{-1}(Y - Y'A^{-1})^T Y'A^{-1}dA\right). \end{aligned} \tag{272}$$

Since $dg_2 = \operatorname{trace}\left(\left(\frac{\partial g_2}{\partial A}\right)^T dA\right)$, we identify

$$\begin{aligned} \frac{\partial g_2}{\partial A} &= (2A^{-1}(Y - Y'A^{-1})^T Y'A^{-1})^T \\ &= 2(A^{-1})^T Y'^T (Y - Y'A^{-1})(A^{-1})^T. \end{aligned} \tag{273}$$

Thus, the total gradient of $f(A)$ is

$$\begin{aligned}\frac{\partial f}{\partial A} &= \frac{\partial g_1}{\partial A} + \frac{\partial g_2}{\partial A} \\ &= 2(AX'^{\top} - X^{\top})X' + 2(A^{-1})^T Y'^T (Y - Y'A^{-1})(A^{-1})^T. \qquad (274)\end{aligned}$$

□

Lemma G.2. *Given matrices $X, X', Y, Y' \in \mathbb{R}^{m\times n}$ with $m \geq n$, the orthogonal matrix $A \in \mathbb{R}^{n\times n}$ satisfying $A^{\top}A = I$ that minimizes the objective function:*

$$f(A) = \|X - X'A^{\top}\|_F^2 + \|Y - Y'A^{-1}\|_F^2, \qquad (275)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, is:

$$A = UV^{\top}, \quad (13) \qquad (276)$$

where U, Σ, V are from the singular value decomposition $B = U\Sigma V^{\top}$, with $B = X^{\top}X' + Y^{\top}Y'$.

Proof. Since A is orthogonal, $A^{-1} = A^{\top}$, so the objective can be rewritten as:

$$f(A) = \|X - X'A^{\top}\|_F^2 + \|Y - Y'A^{\top}\|_F^2. \quad (2) \qquad (277)$$

Step 1. Expand the Objective Function

The Frobenius norm squared is $\|M\|_F^2 = \text{trace}(M^{\top}M)$. We expand the first term of $f(A)$:

$$\begin{aligned}\|X - X'A^{\top}\|_F^2 &= \text{trace}\left((X - X'A^{\top})^{\top}(X - X'A^{\top})\right) \\ &= \text{trace}\left((X^{\top} - AX'^{\top})(X - X'A^{\top})\right) \\ &= \text{trace}\left(X^{\top}X - X^{\top}X'A^{\top} - AX'^{\top}X + AX'^{\top}X'A^{\top}\right) \\ &= \text{trace}(X^{\top}X) - 2\text{trace}(X^{\top}X'A^{\top}) + \text{trace}(AX'^{\top}X'A^{\top}). \qquad (278)\end{aligned}$$

For an orthogonal matrix A, since $A^{\top}A = AA^{\top} = I$ and the trace is invariant under cyclic permutations, we have:

$$\text{trace}(A^{\top}X'^{\top}X'A) = \text{trace}(X'^{\top}X'AA^{\top}) = \text{trace}(X'^{\top}X'). \qquad (279)$$

Thus,

$$\begin{aligned}\|X - X'A^{\top}\|_F^2 = \text{trace}(X^{\top}X) &- 2\text{trace}(X^{\top}X'A^{\top}) \\ &+ \text{trace}(X'^{\top}X'). \qquad (280)\end{aligned}$$

Similarly, for the second term regarding Y:

$$\|Y - Y'A^{\top}\|_F^2 = \text{trace}(Y^{\top}Y) - 2\text{trace}(Y^{\top}Y'A^{\top}) + \text{trace}(Y'^{\top}Y'). \qquad (281)$$

Substituting into $f(A)$:

$$\begin{aligned}f(A) = \text{trace}(X^{\top}X) &+ \text{trace}(X'^{\top}X') + \text{trace}(Y^{\top}Y) + \text{trace}(Y'^{\top}Y') \\ &- 2\left(\text{trace}(X^{\top}X'A^{\top}) + \text{trace}(Y^{\top}Y'A^{\top})\right). \qquad (282)\end{aligned}$$

The terms $\text{trace}(X^{\top}X)$, $\text{trace}(X'^{\top}X')$, $\text{trace}(Y^{\top}Y)$, and $\text{trace}(Y'^{\top}Y')$ are constant with respect to A. Thus, minimizing $f(A)$ is equivalent to maximizing:

$$g(A) = \text{trace}(X^{\top}X'A^{\top}) + \text{trace}(Y^{\top}Y'A^{\top}). \qquad (283)$$

Using the linearity of the trace and the property $\text{trace}(MN) = \text{trace}(NM)$:

$$g(A) = \text{trace}(A^{\top}(X^{\top}X' + Y^{\top}Y')). \qquad (284)$$

Define $B = X^{\top}X' + Y^{\top}Y'$. Then, the problem reduces to maximizing $\text{trace}(A^{\top}B)$ over all orthogonal matrices A.

Step 2. Singular Value Decomposition (SVD)

Compute the singular value decomposition of B:

$$B = U\Sigma V^\top, \tag{285}$$

where $U, V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ is a diagonal matrix with non-negative singular values $\sigma_i \geq 0$.

Then,

$$\text{trace}(A^\top B) = \text{trace}(A^\top U\Sigma V^\top) = \text{trace}(V^\top A^\top U\Sigma). \tag{286}$$

Define $C = V^\top A^\top U$. Since A, U, V are orthogonal, C is also an orthogonal matrix. Thus,

$$\text{trace}(C\Sigma) = \sum_{i=1}^{n} c_{ii}\sigma_i, \tag{287}$$

where c_{ii} are the diagonal elements of C. Since C is orthogonal, its columns (and rows) are orthonormal vectors, which implies $|c_{ii}| \leq 1$ for all i. Therefore:

$$\text{trace}(C\Sigma) \leq \sum_{i=1}^{n} \sigma_i, \tag{288}$$

with equality when $C = I$, i.e., $c_{ii} = 1$ for all i (assuming all $\sigma_i \geq 0$).

Step 3. Verification and Optimality

The maximum value of trace$(A^\top B)$ is $\sum_{i=1}^{n} \sigma_i$, achieved when $C = I$:

$$\begin{aligned} V^\top A^\top U = I &\implies A^\top U = V \implies A^\top = VU^\top \\ &\implies A = (VU^\top)^\top = UV^\top. \end{aligned} \tag{289}$$

Since $A = UV^\top$ maximizes $g(A)$, and $f(A)$ is of the form constant $- 2g(A)$, this choice of A minimizes $f(A)$.

Final Answer

The orthogonal matrix A that minimizes $f(A)$ is:

$$A = UV^\top, \quad (13) \tag{290}$$

where U, Σ, V are from the singular value decomposition $B = U\Sigma V^\top$, with $B = X^\top X' + Y^\top Y'$.

□

Lemma G.3 (Optimal Alignment for RoPE Query-Key Matrices)**.** *Let $W_Q^a, W_K^a \in \mathbb{R}^{d\times d_h}$ and $W_Q^b, W_K^b \in \mathbb{R}^{d\times d_h}$ be the query and key weight matrices for a single attention head from two models, denoted a and b. The problem of finding an alignment matrix $U \in \mathrm{H}(d_h)$ that minimizes the loss function*

$$\mathcal{L}_{Q,K}(U) = \left\|W_Q^a - W_Q^b U^\top\right\|_F^2 + \left\|W_K^a - W_K^b U^{-1}\right\|_F^2 \tag{291}$$

over $U \in H(d_h)$ decouples into $d_h/2$ independent subproblems. For each subspace $j = 1, \dots, d_h/2$, the subproblem of finding the optimal 2×2 matrix U_j is equivalent to finding the minimizer $x_j^\star = \arg\min_{x>0} g_j(x)$ of the 1D scalar objective function

$$g_j(x) = x\,\eta_{Q,j} + \frac{\eta_{K,j}}{x} - 4\sqrt{|\gamma_{Q,j}|^2 x + \frac{|\gamma_{K,j}|^2}{x} + 2\,\mathrm{Re}(\gamma_{Q,j}\bar{\gamma}_{K,j})}, \tag{292}$$

where the constants $\eta_{Q,j}, \eta_{K,j}, \gamma_{Q,j}, \gamma_{K,j}$ are derived from the corresponding weight submatrices as defined in the proof below (with η denoting squared Frobenius norms and γ denoting complex correlation scalars). The optimal matrix $U_j^\star$ is then determined by the optimal value $x_j^\star$.

Proof. **Step 1. Decomposition of the Loss Function.** The loss $\mathcal{L}_{Q,K}(U)$ decouples independently across the $d_h/2$ orthogonal 2D subspaces. For each subspace $j = 1, \dots, d_h/2$, the corresponding loss term is

$$\mathcal{L}^{(j)}_{Q,K}(U_j) = \left\|Q^a_j - Q^b_j U_j^\top\right\|_F^2 + \left\|K^a_j - K^b_j U_j^{-1}\right\|_F^2 , \tag{293}$$

where $Q^m_j = W^m_{Q,j}$ and $K^m_j = W^m_{K,j} \in \mathbb{R}^{d\times 2}$ are the submatrices for model $m \in \{a, b\}$, and $U_j = \begin{pmatrix} a_j & -b_j \\ b_j & a_j \end{pmatrix} \in \mathrm{H}(2)$. Each $\mathcal{L}^{(j)}_{Q,K}$ can be minimized independently.

For simplicity, we drop the index j in the following. The goal is to find the matrix $U = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ that minimizes the loss:

$$\mathcal{L}(a, b) = \|Q^a - Q^b U^\top\|_F^2 + \|K^a - K^b U^{-1}\|_F^2. \tag{294}$$

Step 2. Reduction to a 1D Scalar Objective Function. The problem simplifies by identifying the matrix U with a complex number $z = a + ib$. The squared magnitude is $r^2 = a^2 + b^2 = |z|^2 = \det(U)$. Key properties are $U^\top U = UU^\top = r^2 I$ and $U^{-1} = \frac{1}{r^2}U^\top$.

Using the property $\|M\|_F^2 = \mathrm{tr}(M^\top M)$, expand the loss function. Dropping the constant terms $\|Q^a\|_F^2 + \|K^a\|_F^2$, the objective to minimize is:

$$\begin{aligned}\mathcal{L} &= -2\,\mathrm{tr}((Q^a)^\top Q^b U^\top) + \mathrm{tr}(U(Q^b)^\top Q^b U^\top) \\ &\quad - 2\,\mathrm{tr}((K^a)^\top K^b U^{-1}) + \mathrm{tr}((U^{-1})^\top (K^b)^\top K^b U^{-1}) \\ &= -2\,\mathrm{tr}(C_Q U^\top) + r^2\|Q^b\|_F^2 - 2\,\mathrm{tr}(C_K U^{-1}) + \frac{1}{r^2}\|K^b\|_F^2 \\ &= r^2\eta_Q + \frac{\eta_K}{r^2} - 2\,\mathrm{tr}(C_Q U^\top) - \frac{2}{r^2}\,\mathrm{tr}(C_K U^\top),\end{aligned} \tag{295}$$

where the constants are defined as $\eta_Q = \|Q^b\|_F^2$, $\eta_K = \|K^b\|_F^2$, $C_Q = (Q^a)^\top Q^b$, and $C_K = (K^a)^\top K^b$.

To express the trace terms in complex form, note that $U^\top = aI - bJ$ where $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. This yields the identity $\mathrm{tr}(CU^\top) = a\,\mathrm{tr}(C) - b\,\mathrm{tr}(CJ) = 2\,\mathrm{Re}(\gamma z)$, where the complex scalar $\gamma = \frac{1}{2}(\mathrm{tr}(C) + i\,\mathrm{tr}(CJ))$. Applying this, the loss becomes:

$$\mathcal{L}(z) = |z|^2\eta_Q + \frac{\eta_K}{|z|^2} - 4\,\mathrm{Re}(\gamma_Q z) - \frac{4}{|z|^2}\,\mathrm{Re}(\gamma_K z), \tag{296}$$

where γ_Q and γ_K are complex constants derived from C_Q and C_K respectively.

Express z in polar form as $z = re^{i\theta}$, where $r = |z| > 0$. The loss function can be rewritten to isolate terms dependent on the phase angle θ:

$$\mathcal{L}(r, \theta) = r^2\eta_Q + \frac{\eta_K}{r^2} - 4\,\mathrm{Re}\left(\left(r\gamma_Q + \frac{1}{r}\gamma_K\right)e^{i\theta}\right). \tag{297}$$

First, optimize the phase θ for a fixed magnitude r. The expression is minimized by maximizing the real part term. The maximum value of $\mathrm{Re}(Ce^{i\theta})$ is $|C|$, achieved when $e^{i\theta}$ has angle $-\arg(C)$. Thus, the optimal phase $\theta^\star$ for a given r is:

$$\theta^\star(r) = -\arg\left(r\gamma_Q + \frac{1}{r}\gamma_K\right). \tag{298}$$

Substituting $\theta^\star$ back into the loss yields a 1D scalar objective function depending only on r:

$$g(r) = r^2\eta_Q + \frac{\eta_K}{r^2} - 4\left|r\gamma_Q + \frac{1}{r}\gamma_K\right|. \tag{299}$$

For algebraic convenience, substitute $x = r^2 > 0$. The squared norm term expands as:

$$\begin{aligned}\left|r\gamma_Q + \frac{1}{r}\gamma_K\right|^2 &= \left(\sqrt{x}\gamma_Q + \frac{1}{\sqrt{x}}\gamma_K\right)\left(\sqrt{x}\bar{\gamma}_Q + \frac{1}{\sqrt{x}}\bar{\gamma}_K\right) \\ &= x|\gamma_Q|^2 + \frac{1}{x}|\gamma_K|^2 + 2\,\mathrm{Re}(\gamma_Q\bar{\gamma}_K). \end{aligned} \tag{300}$$

Letting $A = |\gamma_Q|^2$, $B = |\gamma_K|^2$, and $C = 2\,\mathrm{Re}(\gamma_Q\bar{\gamma}_K)$, the objective function in terms of x is:

$$g(x) = x\eta_Q + \frac{\eta_K}{x} - 4\sqrt{Ax + \frac{B}{x} + C}. \tag{301}$$

Step 3. Analysis of the Stationarity Condition. To minimize $g(x)$ for $x > 0$, find stationary points by solving $g'(x) = 0$:

$$g'(x) = \eta_Q - \frac{\eta_K}{x^2} - \frac{2\left(A - \frac{B}{x^2}\right)}{\sqrt{Ax + \frac{B}{x} + C}} = 0. \tag{302}$$

Isolate the square root term and square both sides (noting this may introduce extraneous solutions):

$$\left(\eta_Q - \frac{\eta_K}{x^2}\right)^2 = \frac{4\left(A - \frac{B}{x^2}\right)^2}{Ax + \frac{B}{x} + C}. \tag{303}$$

Multiplying by the denominator and clearing fractions by multiplying by x^4 yields:

$$(\eta_Q x^2 - \eta_K)^2(Ax^2 + Cx + B) = 4x(Ax^2 - B)^2. \tag{304}$$

The left side has degree 6 in x, while the right side has degree 5, so the stationarity condition corresponds to finding roots of a 6th-degree polynomial.

Step 4. Numerical Approach. Since solving a 6th-degree polynomial analytically is generally infeasible and numerical root-finding can be unstable, a more robust approach is to directly minimize the scalar function $g(x)$ using a 1D optimization method. The procedure is as follows:

1. Compute the scalar constants η_Q, η_K and the complex constants γ_Q, γ_K.
2. Define the objective function $g(x) = x\eta_Q + \frac{\eta_K}{x} - 4\sqrt{|\gamma_Q|^2 x + \frac{|\gamma_K|^2}{x} + 2\,\mathrm{Re}(\gamma_Q\bar{\gamma}_K)}$.
3. Find the minimizer $x^\star = \arg\min_{x>0} g(x)$ using a numerical optimization routine. Here we use the Brent's method (Brent, 2013).
4. Compute the optimal solution:
   - Magnitude: $r^\star = \sqrt{x^\star}$.
   - Phase: $\theta^\star = -\arg\left(r^\star\gamma_Q + \frac{1}{r^\star}\gamma_K\right)$.
   - Parameters: $a = r^\star\cos(\theta^\star)$, $b = r^\star\sin(\theta^\star)$.

This yields the optimal alignment matrix U_j for each subspace j. □

G.2 Algorithm Description

Algorithm 1 Attention Layer Alignment

```
Input: $\theta^A$, $\theta^B$.
Output: Aligned $\theta^{B,\text{aligned}}$.
% Stage 1: Head Permutation
Compute cost matrix $C$.
Solve LAP for $\pi^*$.
Reorder $\theta^B \leftarrow \pi^*(\theta^B)$.
% Stage 2: Internal Parameter Alignment
for $i = 1$ to $h$ do
  % Align $Q, K$
  if standard MHA then
    Minimize $\mathcal{L}_{Q,K}(U_i)$ over GL$(d_h)$.
  else
    Minimize $\mathcal{L}_{Q,K}(U_i)$ over H$(d_h)$.
  end if
  Update: $W^Q_{i,B} \leftarrow W^Q_{i,B} U_i^\top$, $W^K_{i,B} \leftarrow W^K_{i,B} U_i^{-1}$.
  % Align $V, O$
  Minimize $\mathcal{L}_{V,O}(V_i)$ over GL$(d_h)$.
  Update: $W^V_{i,B} \leftarrow W^V_{i,B} V_i^{-1}$, $W^O_{i,B} \leftarrow V_i W^O_{i,B}$.
end for
return $\theta^{B,\text{aligned}}$
```

H Impact of Attention Reinitialization on Pretrained Transformer Performance

We investigate the effect of targeted attention reinitialization on pretrained Transformer models. Unlike feedforward blocks, attention layers govern contextual interactions and strongly influence early representations. To assess their contribution, we reset the parameters of individual attention modules using standard initialization, while keeping embeddings, LayerNorms, and feedforward blocks fixed. Models are then evaluated directly on their pretrained tasks without fine-tuning. Our study considers ViT-Base on ImageNet-1K for image classification and GPT-2 on WikiText103 for language modeling, with performance measured in accuracy and perplexity, respectively. Figures 3 and 4 summarize the results across layers.

We find that reinitializing attention layers beyond the first generally leads to only modest degradation, whereas resetting the initial layer produces a pronounced drop in performance. This asymmetry indicates that early attention plays a uniquely critical role in anchoring representations, while deeper layers remain more resilient due to residual connections and redundancy in the architecture. Based on these findings, subsequent experiments on linear mode connectivity focus on reinitializing the first attention layer, as it provides the most consistent and informative signal of model sensitivity.

I Experimental Details and Hyperparameters

Our experiments assess Linear Mode Connectivity (LMC) across a broad spectrum of benchmarks in both vision and natural language processing. The vision suite covers MNIST, CIFAR-10, CIFAR-100, ImageNet-1k, and transfer from ImageNet-21k to smaller classification datasets. For language, we include generative modeling with WikiText103, Enwik8, and the One Billion Word benchmark, together with supervised classification tasks such as AGNews, IMDB reviews, and DBpedia. Each experiment builds on pretrained Transformer architectures, where the core weights remain fixed and only selected attention modules are re-initialized for fine-tuning. Vision tasks use Vision Transformer (ViT) backbones, autoregressive language modeling relies on GPT-2, and text classification tasks are handled by BERT.

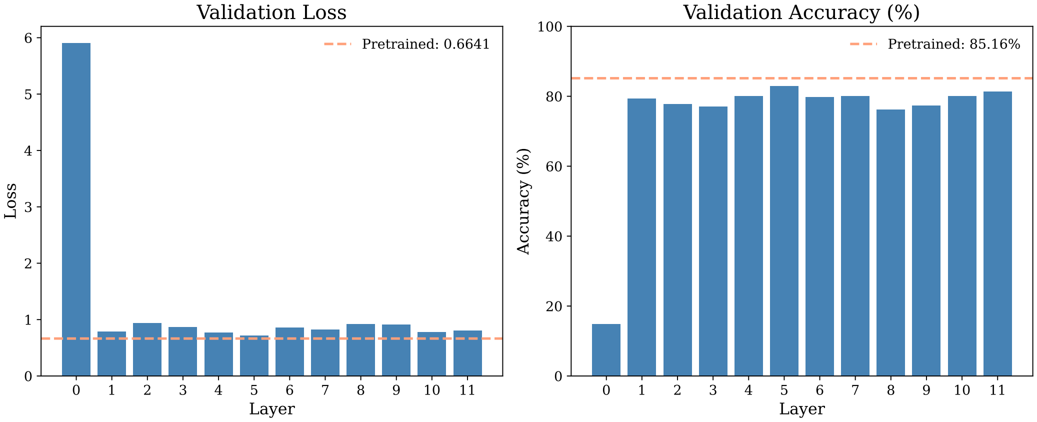


Figure 3: Performance degradation in ViT-Base on ImageNet due to attention reinitialization at different layers.

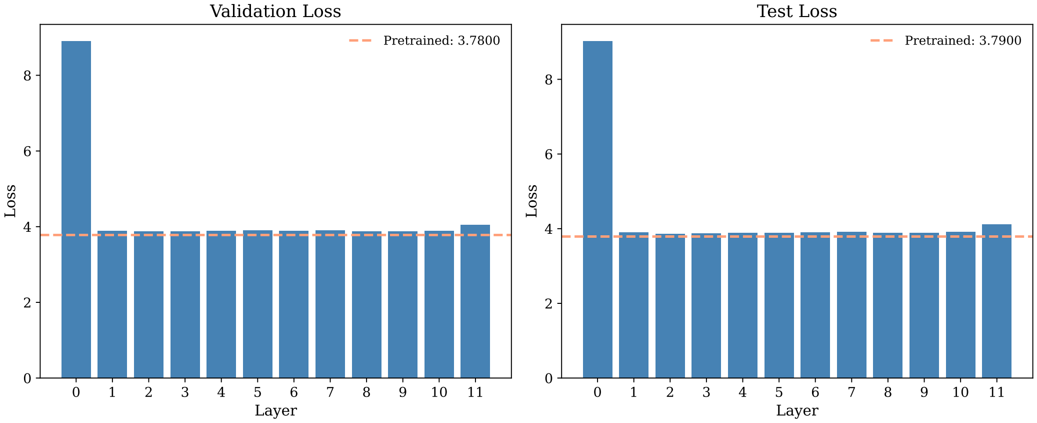


Figure 4: Effect of attention reinitialization on GPT-2 perplexity across layers on WikiText103.

AGNews. For the AGNews dataset, we adopt a compact BERT-style encoder with embedding dimension 96, hidden size 384, and vocabulary size 15,000. Models are trained with depths of 2 or 6 layers and attention configurations of 4 or 8 heads. Pretraining is carried out using the Adam optimizer with a batch size of 512 and learning rate $1 \cdot 10^{-3}$, for up to 6 epochs until convergence.

IMDBreview. For the IMDB dataset, we adopt a compact BERT-style encoder with embedding dimension 96, hidden size 384, and vocabulary size 15,000. Models are trained with depths of 2, or 6 layers and attention configurations of 4 or 8 heads. Pretraining is performed using the Adam optimizer with a batch size of 128 and learning rate $3 \cdot 10^{-4}$, for up to 7 epochs until convergence.

DBPedia. For the DBPedia dataset, we adopt a compact BERT-style encoder with embedding dimension 96, hidden size 384, vocabulary size 30,522, and 219 output classes (max sequence length 256). Models are trained with depths of 2 or 6 layers and attention configurations of 4 or 8 heads. Pretraining is carried out using the Adam optimizer with a batch size of 256 and learning rate 1×10^{-3} under a linear decay schedule, for up to 5 epochs until convergence.

Enwik8. For the Enwik8 dataset, we employ a GPT-2 style Transformer with 12 layers, hidden size of 512, 8 attention heads, and an intermediate size of 2048. The context length is set to 512 tokens, with memory length 512 and evaluation length 128. Pretraining is performed using the Adam optimizer with a batch size of 24 and an initial learning rate of $2.5 \cdot 10^{-4}$, following a cosine decay schedule without warmup, for a total of 60000 steps. During fine-tuning, we replace the pretrained attention modules with variants containing 4, 8, or 16 heads, and train for 60000 steps.

WikiText103. For the WikiText103 benchmark, we adopt a GPT-2 style Transformer with 12 layers, hidden size of 192, 3 attention heads, and an intermediate size of 768. The model uses learned attention biases, with context length, memory length, and evaluation length all set to 256 tokens. Training is conducted with the Adam optimizer using a batch size of 64 and an initial learning rate of $2.5 \cdot 10^{-4}$. A linear warmup of 2000 steps is followed by a cosine decay learning rate schedule. The pretraining phase runs for 60k steps. For fine-tuning, we replace the attention modules with variants containing 2, 3, or 4 heads, and train each configuration for 60000 steps.

One Billion Word. For the One Billion Word benchmark, we employ a GPT-2 style Transformer-based language model with sinusoidal positional embeddings, 12 layers, hidden size of 768, 12 attention heads, and an intermediate size of 3072. The vocabulary size is 793,470. Pretraining is performed with target sequence length 256, memory length 256, and evaluation sequence length 256. The model is trained using Adam with a batch size of 96, an initial learning rate of $2.5 \cdot 10^{-4}$, and a cosine decay learning rate schedule with 2000 warmup steps. Training is run for 500000 steps with random seed fixed at 0 for reproducibility. For fine-tuning, we replace the attention mechanism with variants containing 8, 12, or 16 heads. Each configuration is fine-tuned for 100000 steps.

MNIST. For the MNIST dataset, we adopt a lightweight Vision Transformer with patch size 7, embedding dimension 16, hidden size 64, and depths of 1 or 2 layers paired with 4 or 8 attention heads. Pretraining is carried out using the Adam optimizer with a learning rate of 5×10^{-3}, training to validation convergence (typically 60–80 epochs, depending on configuration).

CIFAR-10. For CIFAR-10, we use a Vision Transformer with patch size 4, embedding dimension 128, hidden size 512, and depths of 2, 4, or 6 layers paired with 4 or 8 attention heads. Images are normalized with CIFAR-10 statistics and augmented using random resized crop, horizontal flip, and rotation. Pretraining is performed with the Adam optimizer at a learning rate of 5×10^{-3} for 100 epochs with batch size 100.

CIFAR-100. For CIFAR-100, we adopt a Vision Transformer with patch size 4, embedding dimension 128, hidden size 512, and depths of 6 layers paired with 4 or 8 attention heads. Images are normalized using standard CIFAR-100 statistics and augmented with random resized crop, horizontal flip, and rotation. Pretraining is conducted with the Adam optimizer at a learning rate of 5×10^{-3} for 100 epochs and batch size 100.

Imagenet21k→CIFAR10. We adopt the ViT-Small-Patch16-224 model, pretrained on ImageNet-21k and subsequently fine-tuned on CIFAR-10. The model consists of 12 layers, a hidden size of 384, an MLP size of 1536, and 6 attention heads, resulting in approximately 22.2M parameters. It employs a patch size and stride of 16. Dropout is disabled (set to 0.0), and the activation function is `gelu`. Stochastic Gradient Descent (SGD) is employed during fine-tuning.

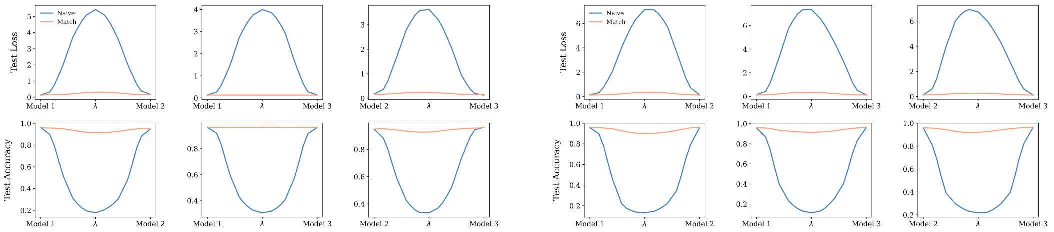
Imagenet21k→CIFAR100. We adopt the ViT-Small-Patch16-224 model, pretrained on ImageNet-21k and subsequently fine-tuned on CIFAR-100. The model consists of 12 layers, a hidden size of 384, an MLP size of 1536, and 6 attention heads, resulting in approximately 22.2M parameters. It employs a patch size and stride of 16. Dropout is disabled (set to 0.0), and the activation function is `gelu`. Stochastic Gradient Descent (SGD) is employed during fine-tuning.

ImageNet-1k. For ImageNet-1k, we utilize a pretrained Vision Transformer with the following configuration: hidden size of 768, 12 Transformer layers, 12 attention heads, and an intermediate size of 3072. Training is performed for 300 epochs with a batch size of 256 using the Adam optimizer and an initial learning rate of $5 \cdot 10^{-4}$. The learning rate follows a cosine decay schedule with 5 epochs of linear warmup. The `gelu` activation function is employed throughout the network, and both attention and hidden dropout rates are set to 0.0. During fine-tuning, we systematically replace the pretrained attention layers with variants containing 8, 12, or 16 heads. Depending on the number of re-initialized layers, the fine-tuning budget is set to 30, 50, 100, or 300 epochs, respectively.

Runtime Environment. All experiments were executed on NVIDIA H100 GPUs with 80GB of memory. A single GPU was sufficient for every task, except for the One Billion Word benchmark, which required two GPUs. Since training was implemented in JAX, approximately 75% of the GPU memory (about 60GB) was pre-allocated by default. For data loading and preprocessing, the number of CPU workers was limited to 10. In terms of wall-clock time, small-scale benchmarks–including MNIST, CIFAR-10, CIFAR-100, transfer learning from ImageNet-21k, and text classification datasets (AGNews, IMDB reviews, DBPedia)–each completed in under 30 minutes. For language modeling, both WikiText103 and Enwik8 required about 2 hours for pretraining and fine-tuning. The One Billion Word benchmark was more computationally demanding, requiring up to 2 days. On the vision side, ImageNet-1k fine-tuning could take as long as 6 days, depending on the configuration.

J Experiments

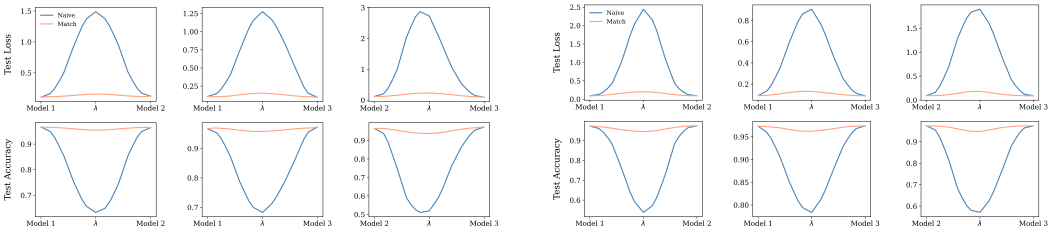
J.1 Linear Mode Connectivity for Attention First Layer



(a) 4 attention heads

(b) 8 attention heads

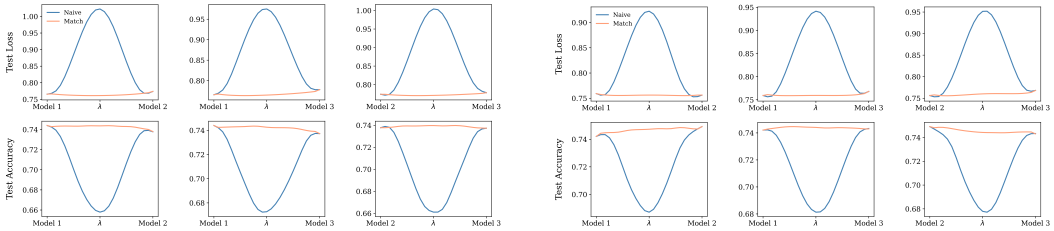
Figure 5: Linear Mode Connectivity for ViT on MNIST with 1 layer



(a) 4 attention heads

(b) 8 attention heads

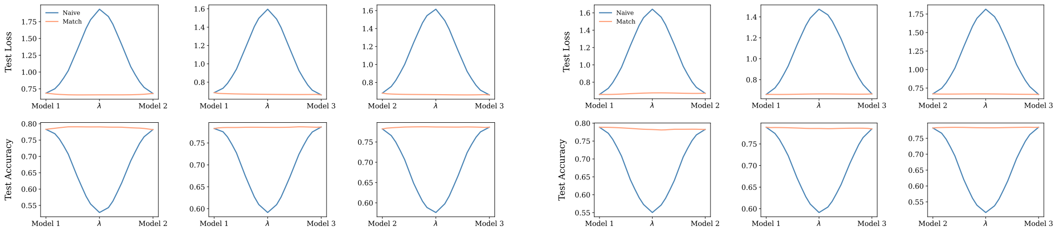
Figure 6: Linear Mode Connectivity for ViT on MNIST with 2 layers



(a) 4 attention heads

(b) 8 attention heads

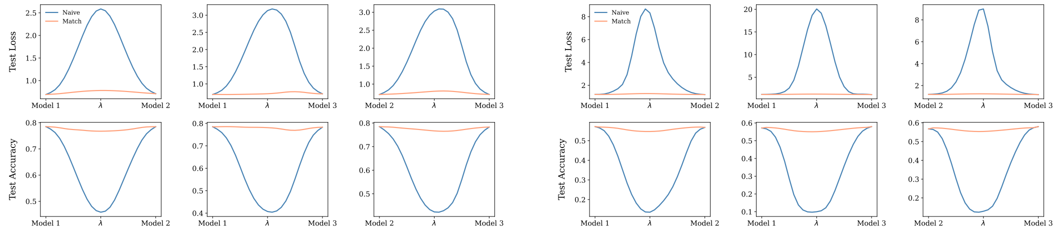
Figure 7: Linear Mode Connectivity for ViT on CIFAR-10 with 2 layers



(a) 4 attention heads

(b) 8 attention heads

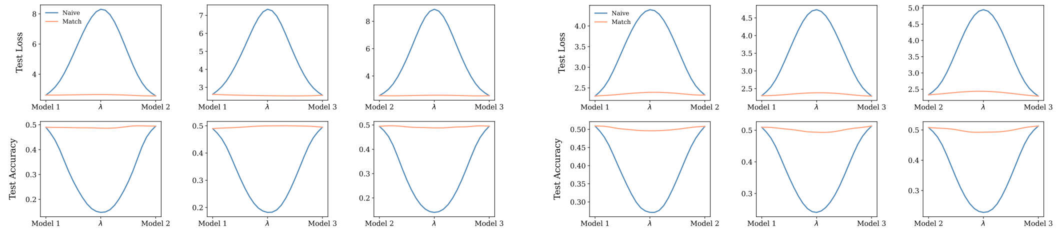
Figure 8: Linear Mode Connectivity for ViT on CIFAR-10 with 4 layers



(a) 4 attention heads

(b) 8 attention heads

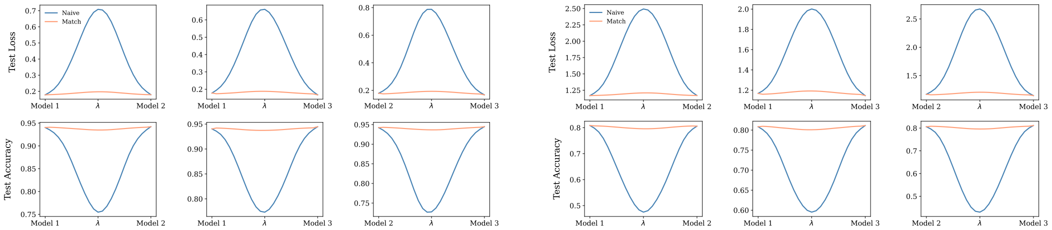
Figure 9: Linear Mode Connectivity for ViT on CIFAR-10 with 6 layers



(a) 4 attention heads

(b) 8 attention heads

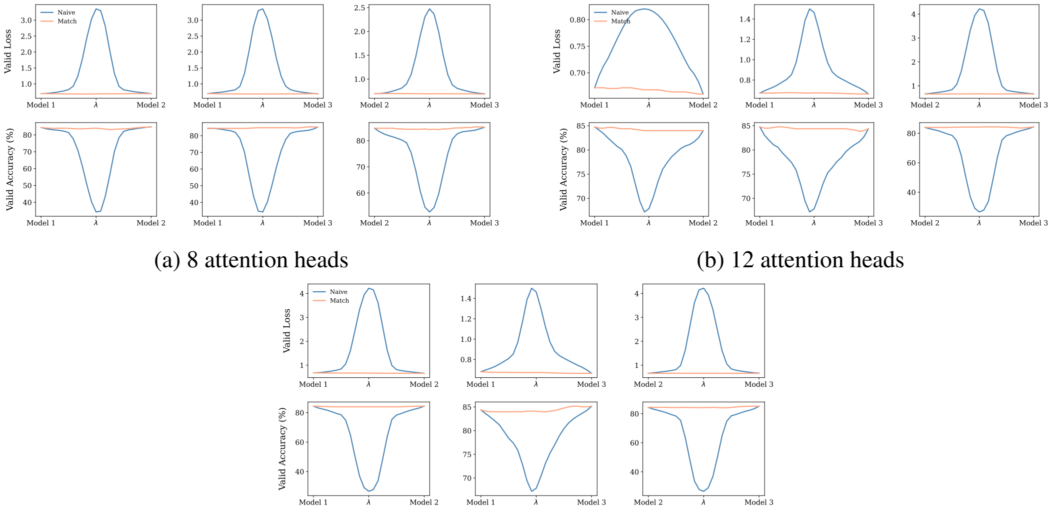
Figure 10: Linear Mode Connectivity for ViT on CIFAR-100 with 6 layers



(a) CIFAR-10

(b) CIFAR-100

Figure 11: Linear Mode Connectivity for ViT on ImageNet21k→CIFAR-10/100 with 12 layers and 6 heads

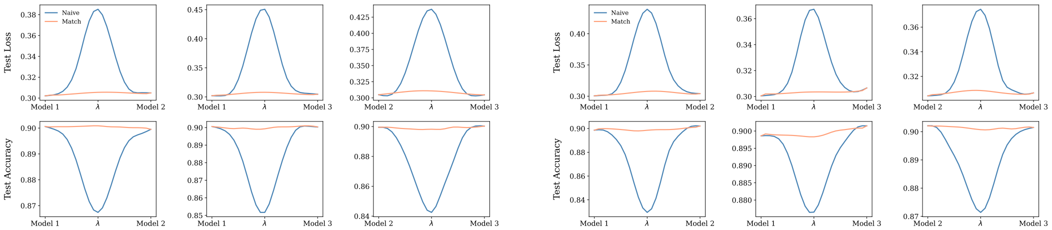


(a) 8 attention heads

(b) 12 attention heads

(c) 16 attention heads

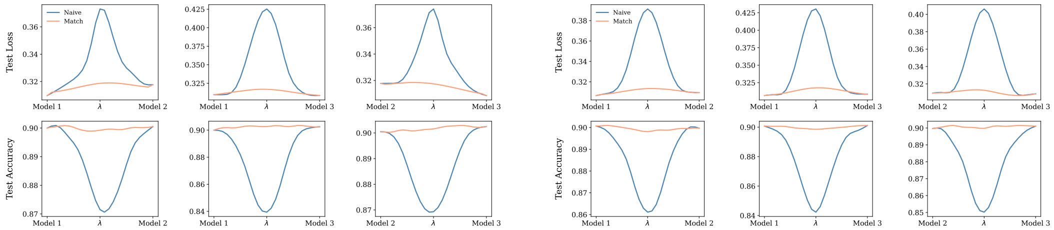
Figure 12: Linear Mode Connectivity for ViT on ImageNet with 12 layers.



(a) 4 attention heads

(b) 8 attention heads

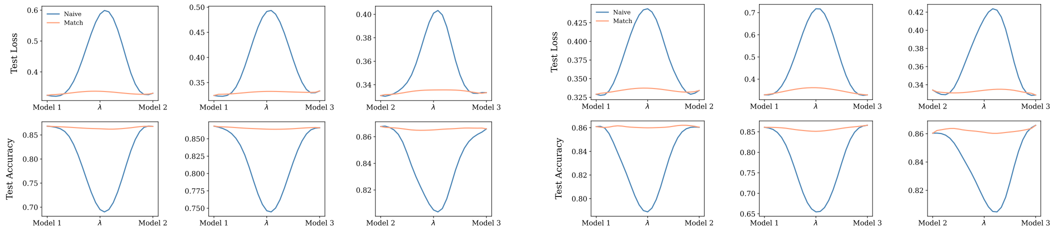
Figure 13: Linear Mode Connectivity for BERT on AGnews with 2 layers



(a) 4 attention heads

(b) 8 attention heads

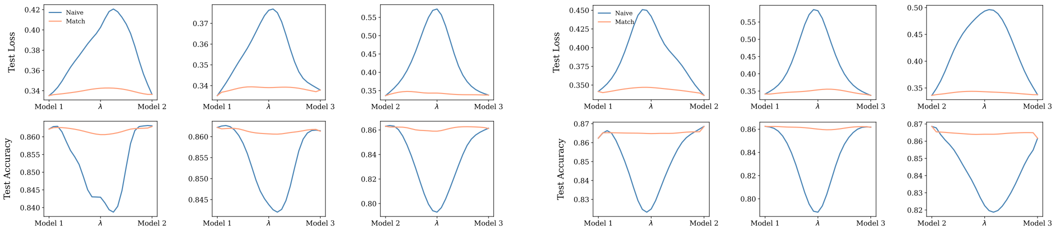
Figure 14: Linear Mode Connectivity for BERT on AGnews with 6 layers



(a) 4 attention heads

(b) 8 attention heads

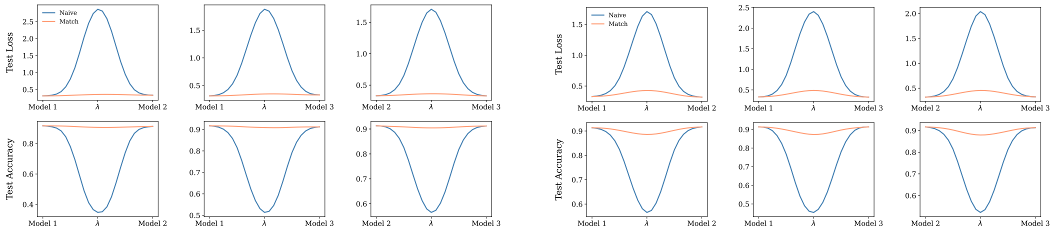
Figure 15: Linear Mode Connectivity for BERT on IMDBreview with 2 layers



(a) 4 attention heads

(b) 8 attention heads

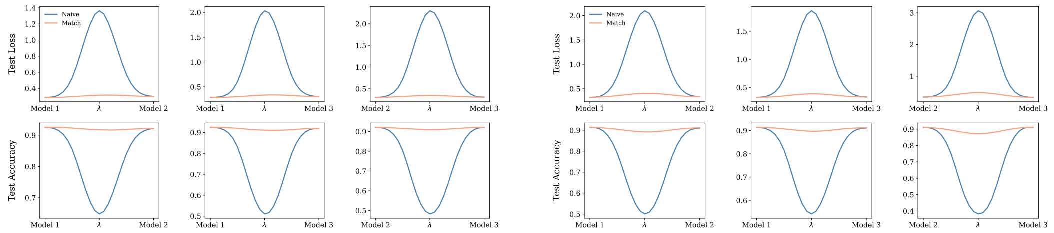
Figure 16: Linear Mode Connectivity for BERT on IMDBreview with 6 layers



(a) 4 attention heads

(b) 8 attention heads

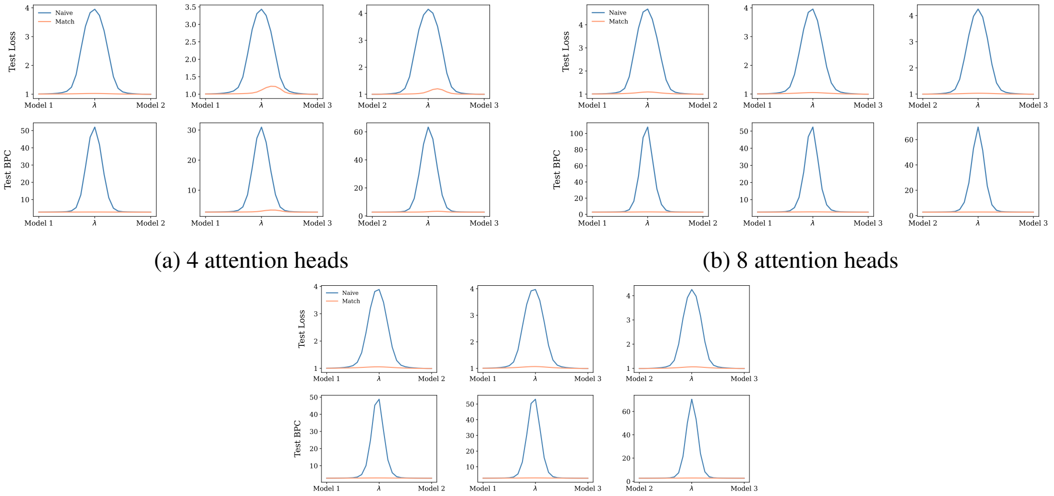
Figure 17: Linear Mode Connectivity for BERT on DBPedia with 2 layers



(a) 4 attention heads

(b) 8 attention heads

Figure 18: Linear Mode Connectivity for BERT on DBPedia with 6 layers

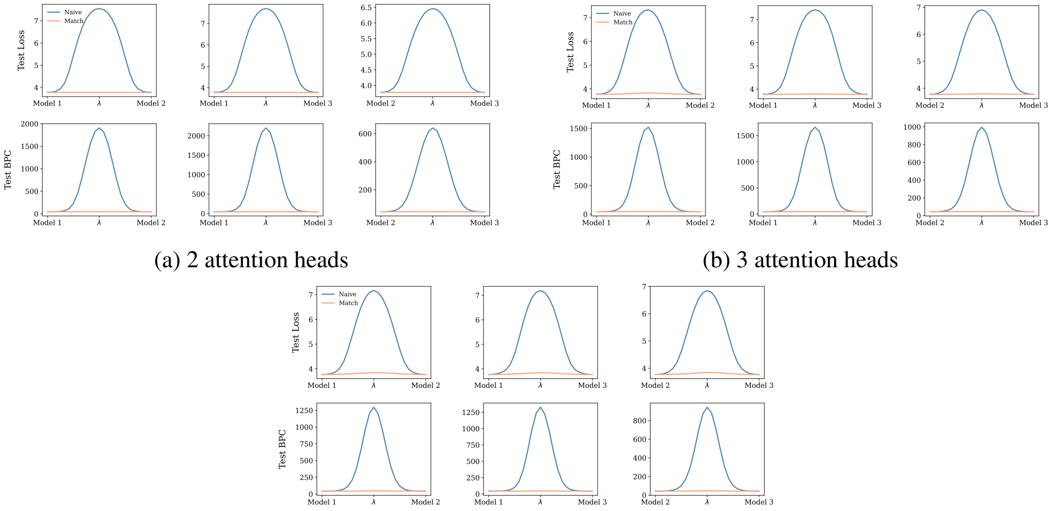


(a) 4 attention heads

(b) 8 attention heads

(c) 16 attention heads

Figure 19: Linear Mode Connectivity for GPT2 on Enwik8 with 12 layers.

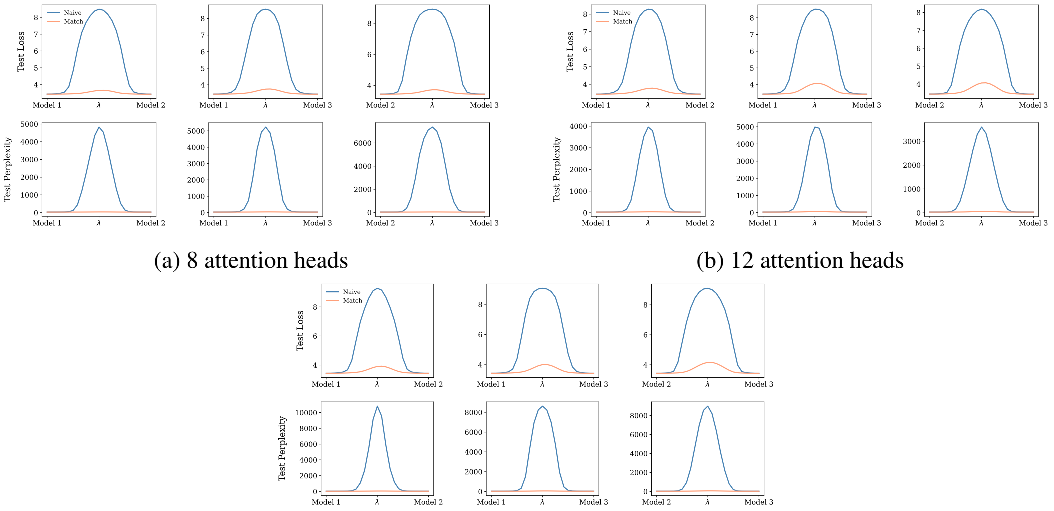


(a) 2 attention heads

(b) 3 attention heads

(c) 4 attention heads

Figure 20: Linear Mode Connectivity for GPT2 on Wikitext103 with 12 layers.

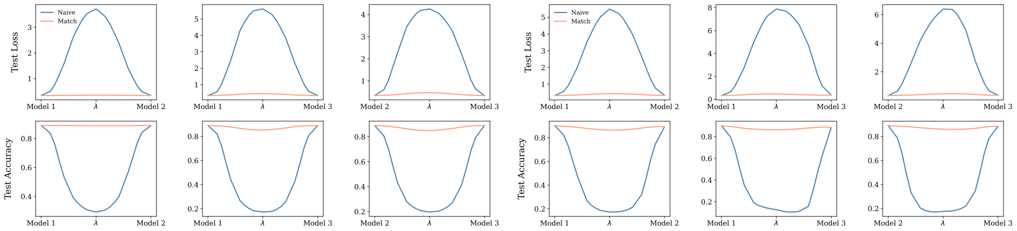


(a) 8 attention heads

(b) 12 attention heads

(c) 16 attention heads

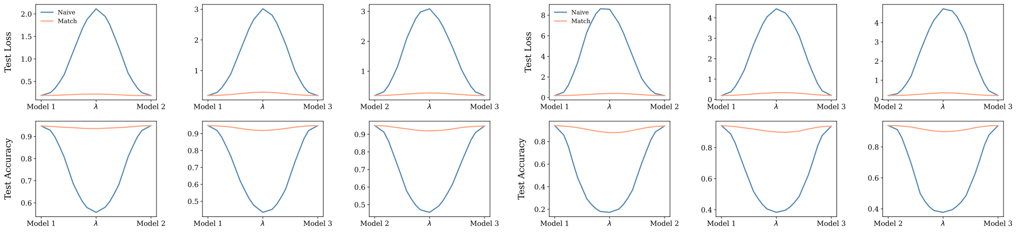
Figure 21: Linear Mode Connectivity for GPT2 on One Billion Words with 12 layers.



(a) 4 attention heads

(b) 8 attention heads

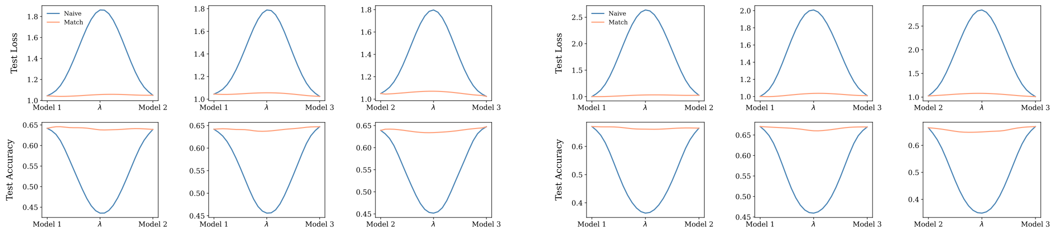
Figure 22: Linear Mode Connectivity for ViT-RoPE on MNIST with 1 layer



(a) 4 attention heads

(b) 8 attention heads

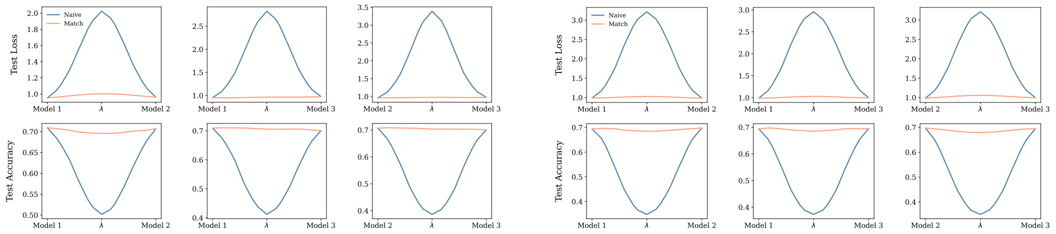
Figure 23: Linear Mode Connectivity for ViT-RoPE on MNIST with 2 layers



(a) 4 attention heads

(b) 8 attention heads

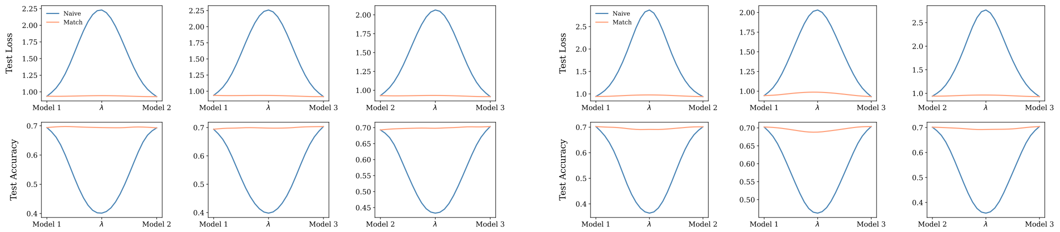
Figure 24: Linear Mode Connectivity for ViT-RoPE on CIFAR-10 with 2 layers



(a) 4 attention heads

(b) 8 attention heads

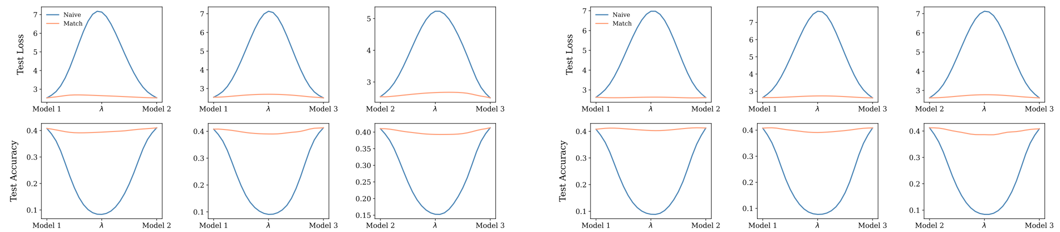
Figure 25: Linear Mode Connectivity for ViT-RoPE on CIFAR-10 with 4 layers



(a) 4 attention heads

(b) 8 attention heads

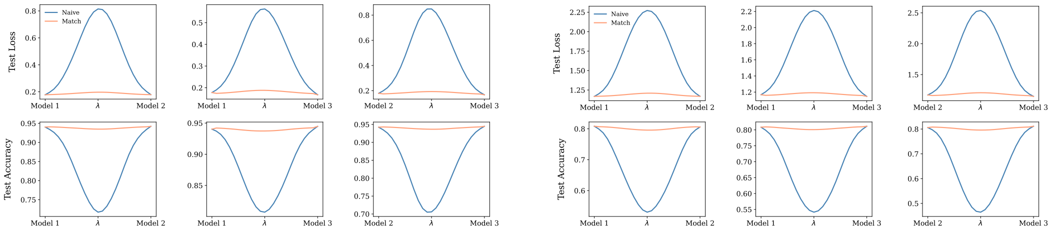
Figure 26: Linear Mode Connectivity for ViT-RoPE on CIFAR-10 with 6 layers



(a) 4 attention heads

(b) 8 attention heads

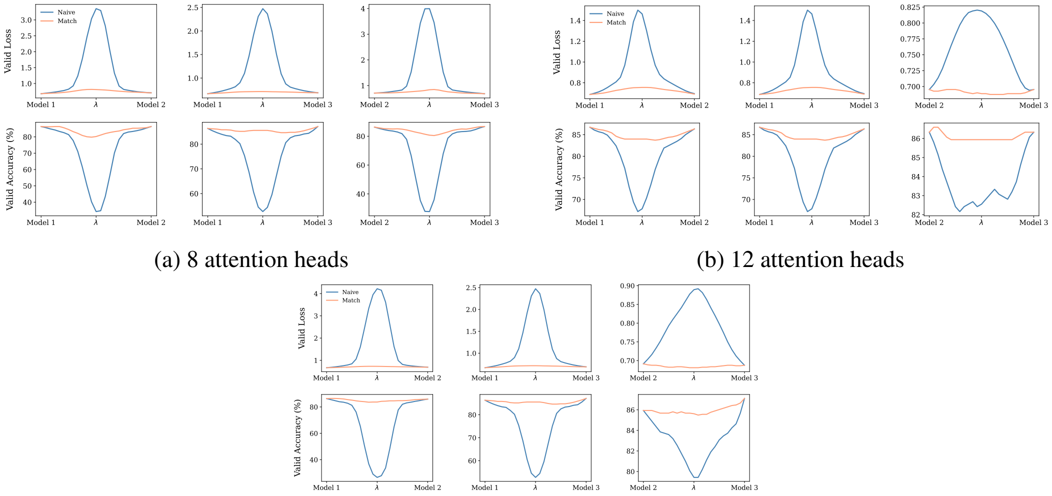
Figure 27: Linear Mode Connectivity for ViT-RoPE on CIFAR-100 with 6 layers



(a) CIFAR-10

(b) CIFAR-100

Figure 28: Linear Mode Connectivity for ViT-RoPE on ImageNet21k→CIFAR-10/100 with 12 layers and 6 heads

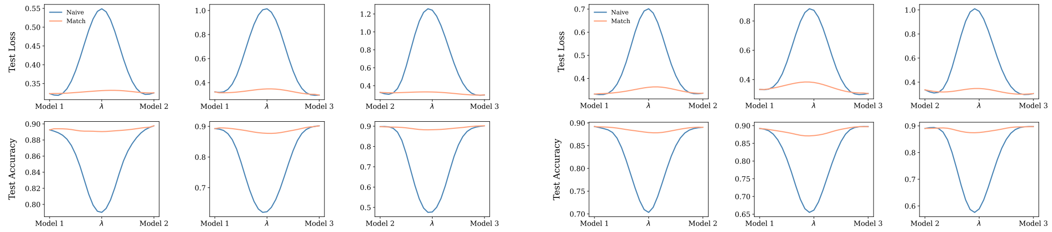


(a) 8 attention heads

(b) 12 attention heads

(c) 16 attention heads

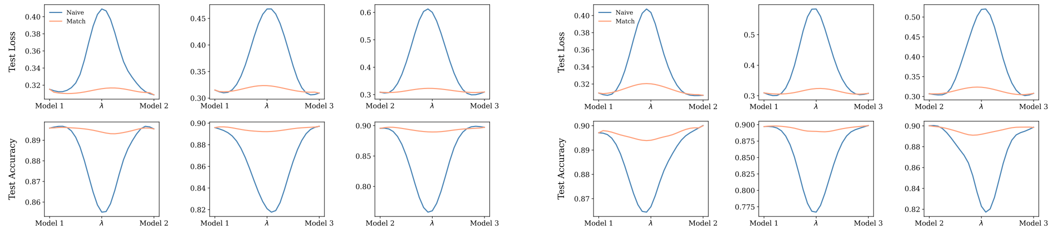
Figure 29: Linear Mode Connectivity for ViT-RoPE on ImageNet with 12 layers.



(a) 4 attention heads

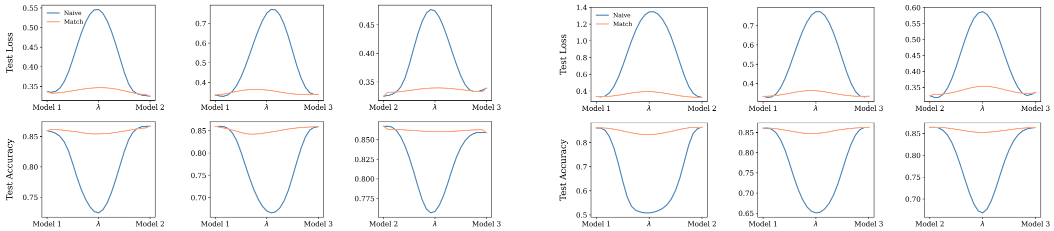
(b) 8 attention heads

Figure 30: Linear Mode Connectivity for BERT-RoPE on AGnews with 2 layers



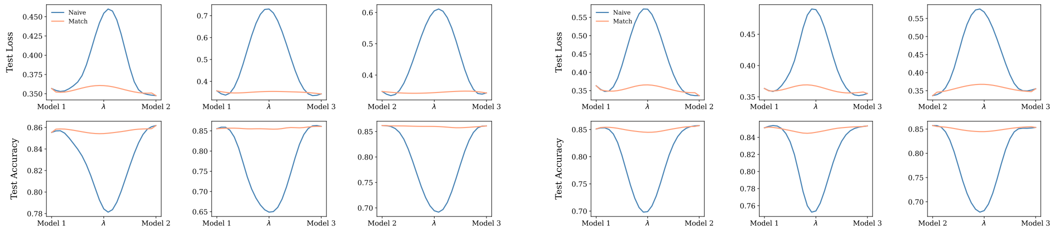
(a) 4 attention heads (b) 8 attention heads

Figure 31: Linear Mode Connectivity for BERT-RoPE on AGnews with 6 layers



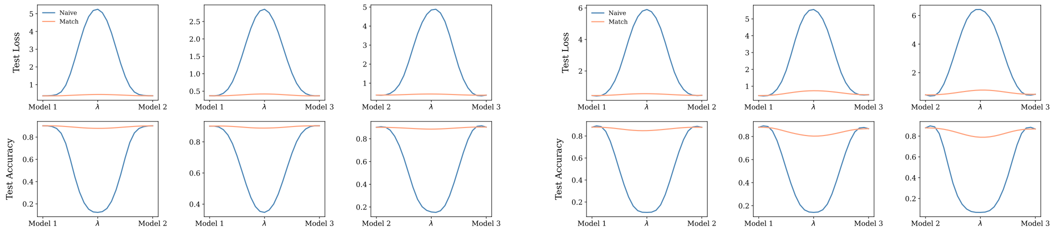
(a) 4 attention heads (b) 8 attention heads

Figure 32: Linear Mode Connectivity for BERT-RoPE on IMDBreview with 2 layers



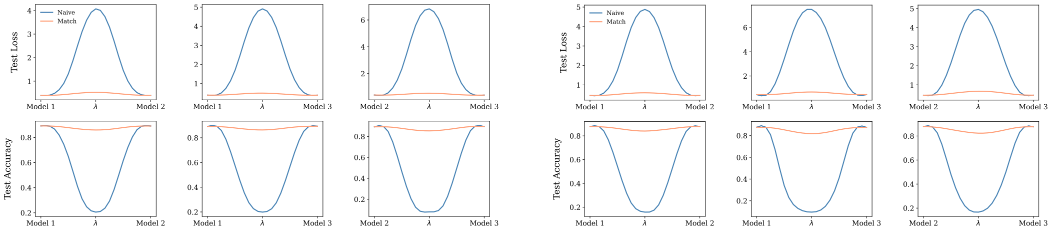
(a) 4 attention heads (b) 8 attention heads

Figure 33: Linear Mode Connectivity for BERT-RoPE on IMDBreview with 6 layers



(a) 4 attention heads (b) 8 attention heads

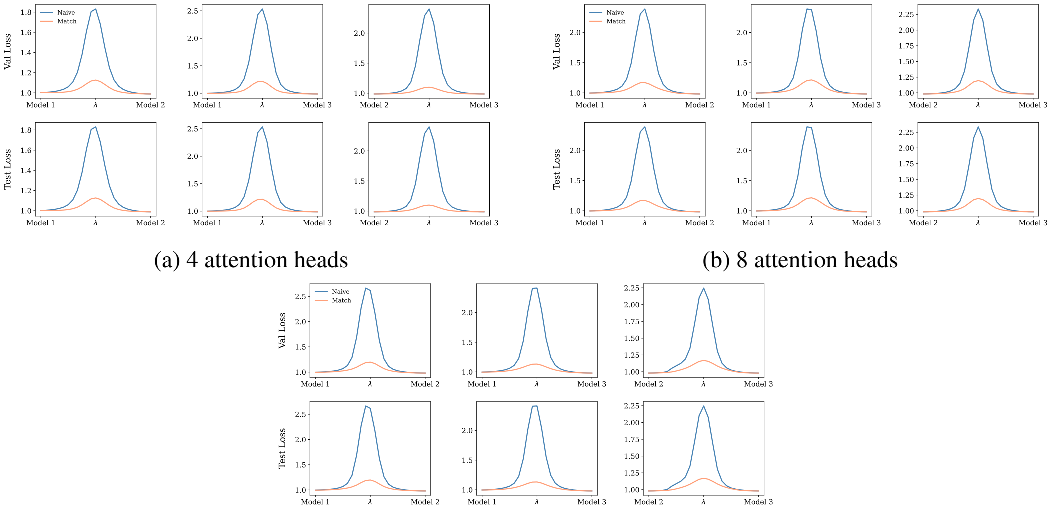
Figure 34: Linear Mode Connectivity for BERT-RoPE on DBPedia with 2 layers



(a) 4 attention heads

(b) 8 attention heads

Figure 35: Linear Mode Connectivity for BERT-RoPE on DBPedia with 6 layers

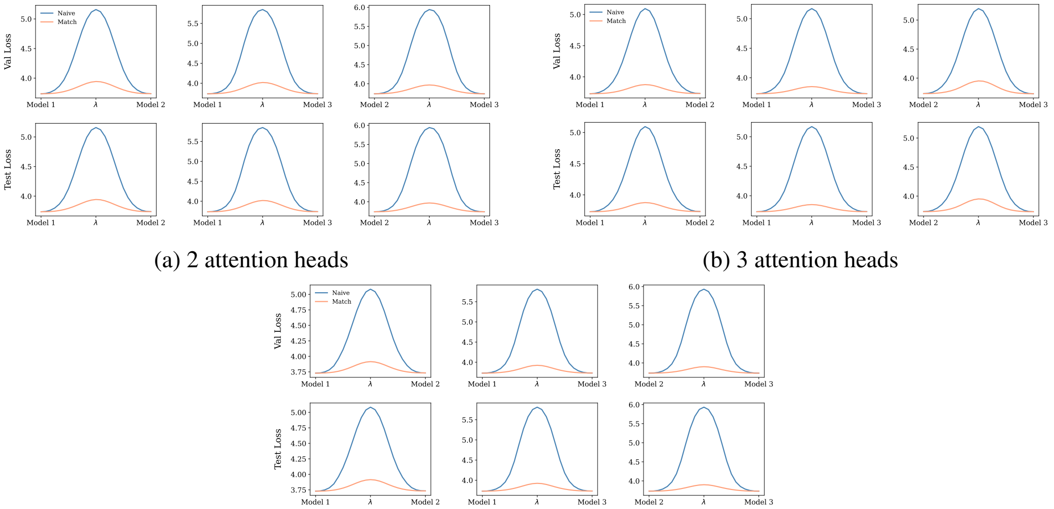


(a) 4 attention heads

(b) 8 attention heads

(c) 16 attention heads

Figure 36: Linear Mode Connectivity for GPT2-RoPE on Enwik8 with 12 layers.

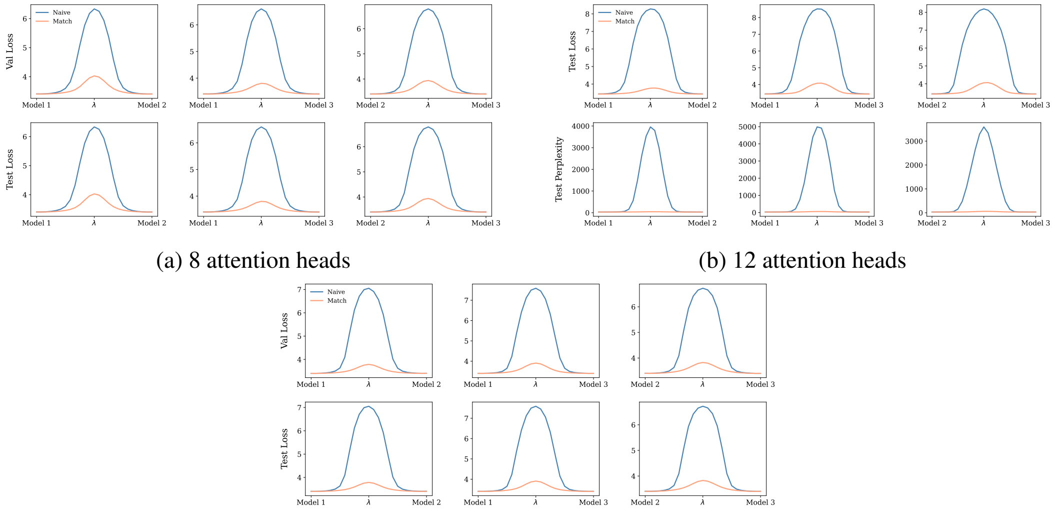


(a) 2 attention heads

(b) 3 attention heads

(c) 4 attention heads

Figure 37: Linear Mode Connectivity for GPT2-RoPE on Wikitext103 with 12 layers.



(a) 8 attention heads

(b) 12 attention heads

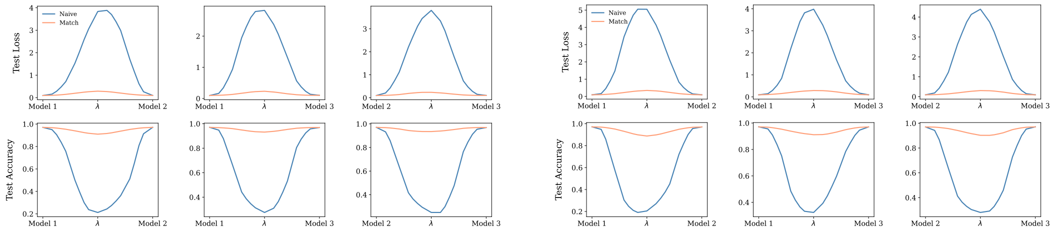
(c) 16 attention heads

Figure 38: Linear Mode Connectivity for GPT2-RoPE on OneBillionWord with 12 layers.

J.2 Linear Mode Connectivity for Attention at All Layers

Table 4: Experimental configurations for LMC evaluation under re-initialization of *all attention layers*. The table reports datasets, model depth, and attention head counts, with figure references showing interpolation curves for APE and RoPE variants. Entries of the form $A \rightarrow B$ denote models pretrained on A, fine-tuned on B, and assessed on B.

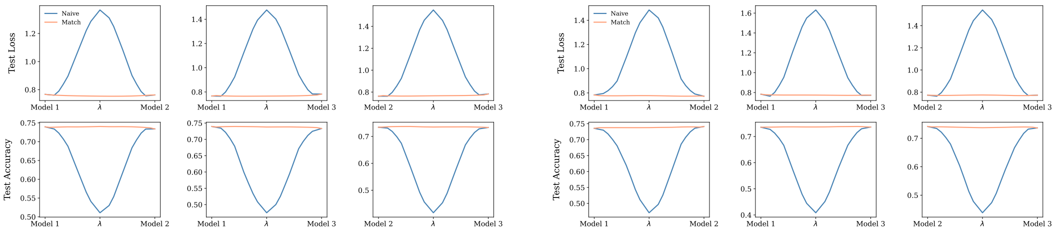
Dataset	Layers	Heads	APE	RoPE	Dataset	Layers	Heads	APE	RoPE
MNIST	2	[4, 8]	[39a, 39b]	[54a, 54b]	AGNews	2	[4, 8]	[46a, 46b]	[60a, 60b]
CIFAR-10	2	[4, 8]	[40a, 40b]	[55a, 55b]		6	[4, 8]	[47a, 47b]	[61a, 61b]
	4	[4, 8]	[41a, 41b]	[56a, 56b]	IMDB	2	[4, 8]	[48a, 48b]	[62a, 62b]
	6	[4, 8]	[42a, 42b]	[57a, 57b]		6	[4, 8]	[49a, 49b]	[63a, 63b]
CIFAR-100	6	[4, 8]	[43a, 43b]	[58a, 58b]	DBPedia	2	[4, 8]	[50a, 50b]	[64a, 64b]
ImageNet-21k→CIFAR-10	12	[6]	[44a]	[59a]		6	[4, 8]	[51a, 51b]	[65a, 65b]
ImageNet-21k→CIFAR-100	12	[6]	[44b]	[59b]	Enwik8	12	[8]	[52a]	[52b]
ImageNet-1k	12	[12]	[45a]	[45b]	WikiText103	12	[3]	[53a]	[53b]



(a) 4 attention heads

(b) 8 attention heads

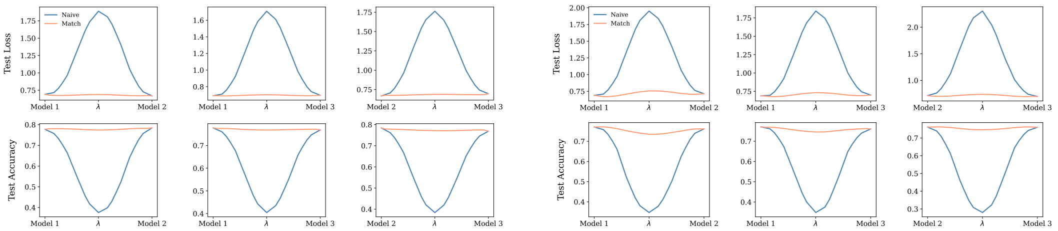
Figure 39: Linear Mode Connectivity for ViT on MNIST with 2 layers



(a) 4 attention heads

(b) 8 attention heads

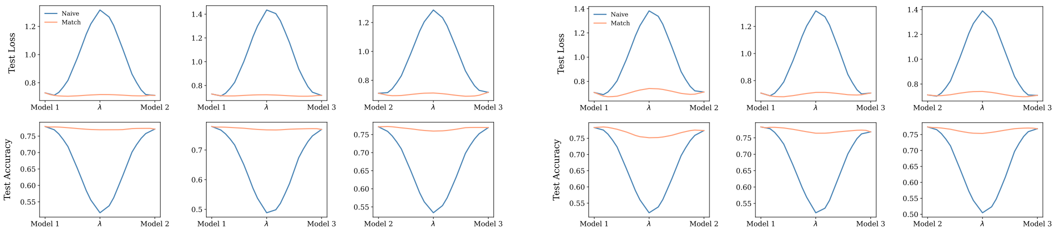
Figure 40: Linear Mode Connectivity for ViT on CIFAR-10 with 2 layers



(a) 4 attention heads

(b) 8 attention heads

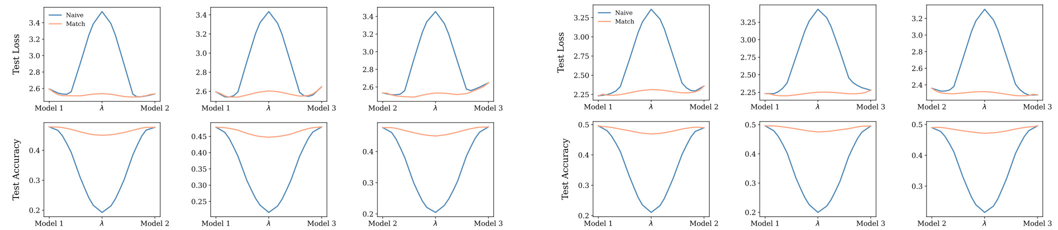
Figure 41: Linear Mode Connectivity for ViT on CIFAR-10 with 4 layers



(a) 4 attention heads

(b) 8 attention heads

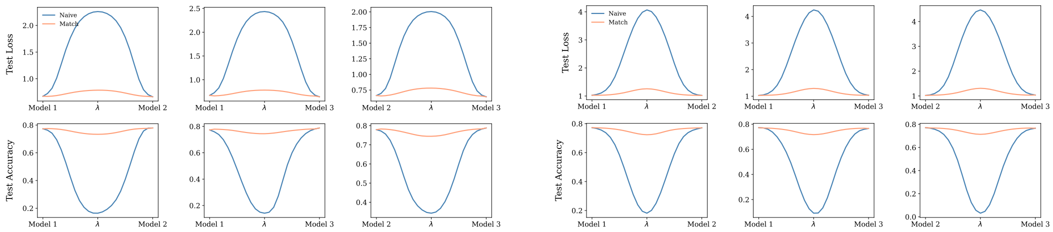
Figure 42: Linear Mode Connectivity for ViT on CIFAR-10 with 6 layers



(a) 4 attention heads

(b) 8 attention heads

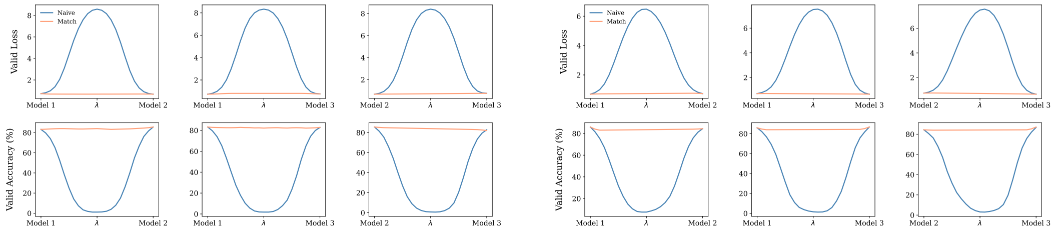
Figure 43: Linear Mode Connectivity for ViT on CIFAR-100 with 6 layers



(a) CIFAR-10

(b) CIFAR-100

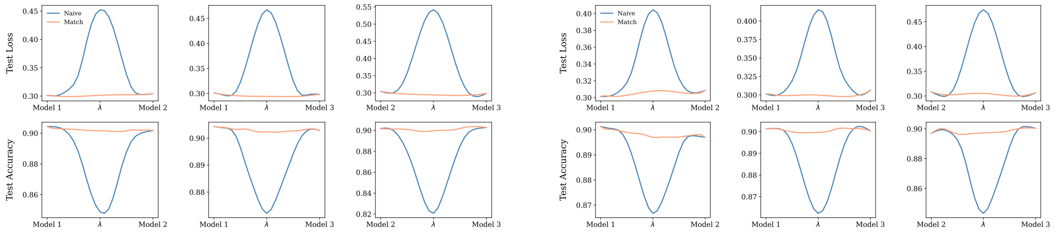
Figure 44: Linear Mode Connectivity for ViT on ImageNet21k→CIFAR-10/100 with 12 layers and 6 heads



(a) APE

(b) RoPE

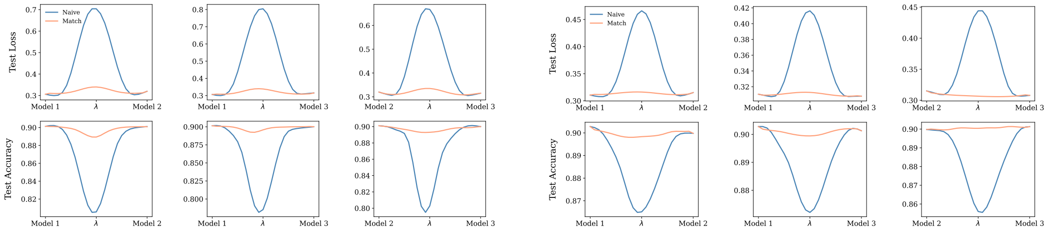
Figure 45: Linear Mode Connectivity for ViT with APE and RoPE on ImageNet-1k with 12 layers



(a) 4 attention heads

(b) 8 attention heads

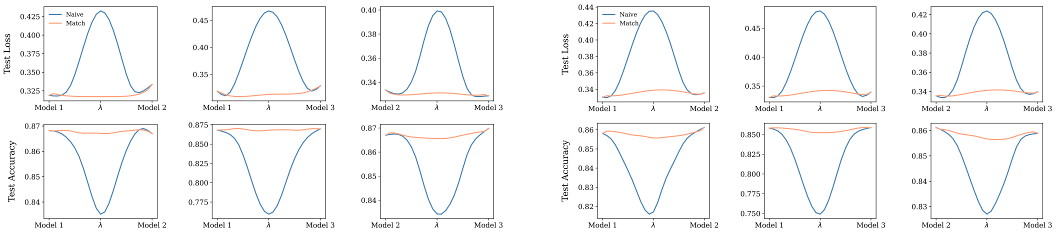
Figure 46: Linear Mode Connectivity for BERT on AGnews with 2 layers



(a) 4 attention heads

(b) 8 attention heads

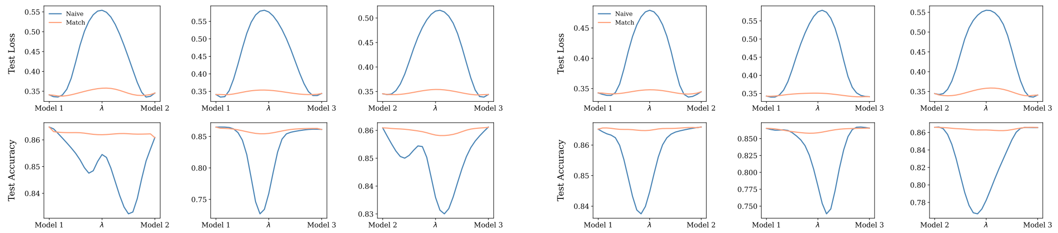
Figure 47: Linear Mode Connectivity for BERT on AGnews with 6 layers



(a) 4 attention heads

(b) 8 attention heads

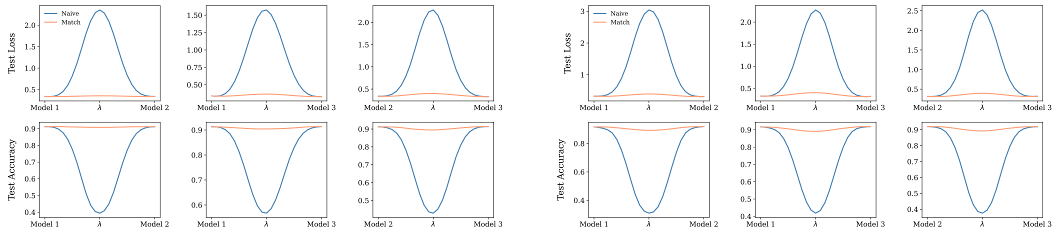
Figure 48: Linear Mode Connectivity for BERT on IMDBreview with 2 layers



(a) 4 attention heads

(b) 8 attention heads

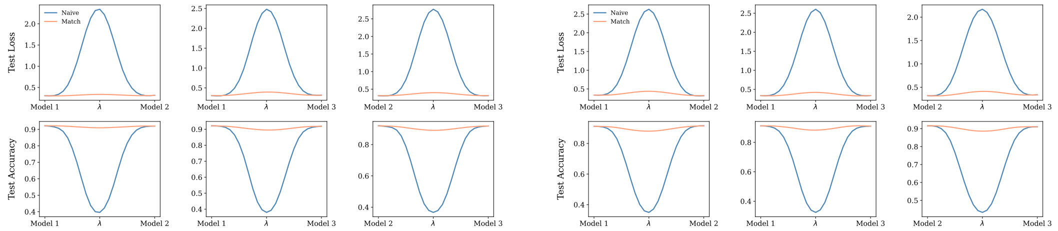
Figure 49: Linear Mode Connectivity for BERT on IMDBreview with 6 layers



(a) 4 attention heads

(b) 8 attention heads

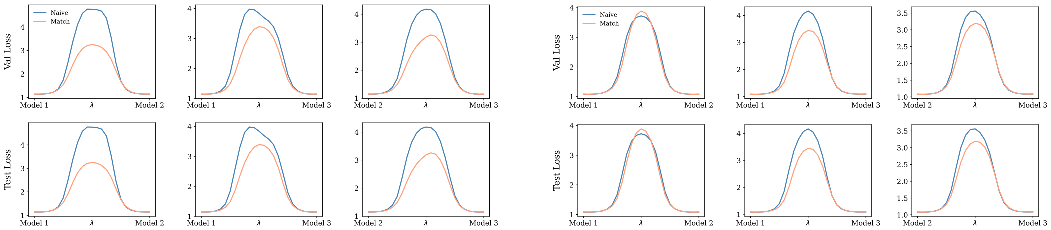
Figure 50: Linear Mode Connectivity for BERT on DBPedia with 2 layers



(a) 4 attention heads

(b) 8 attention heads

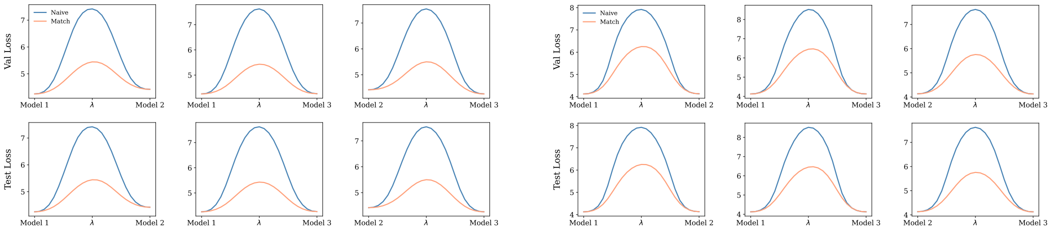
Figure 51: Linear Mode Connectivity for BERT on DBPedia with 6 layers



(a) APE

(b) RoPE

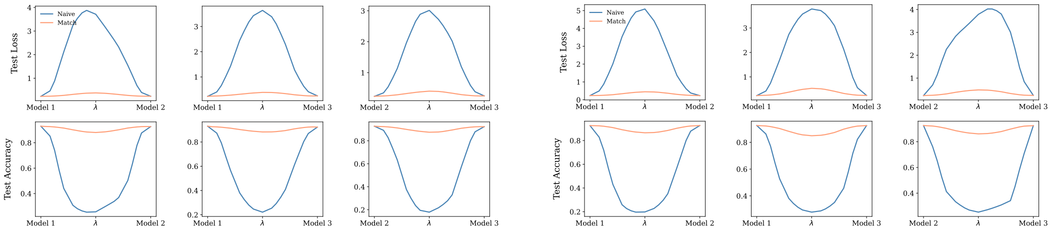
Figure 52: Linear Mode Connectivity for GPT2 with APE and RoPE on Enwik8 with 12 layers and 8 heads



(a) APE

(b) RoPE

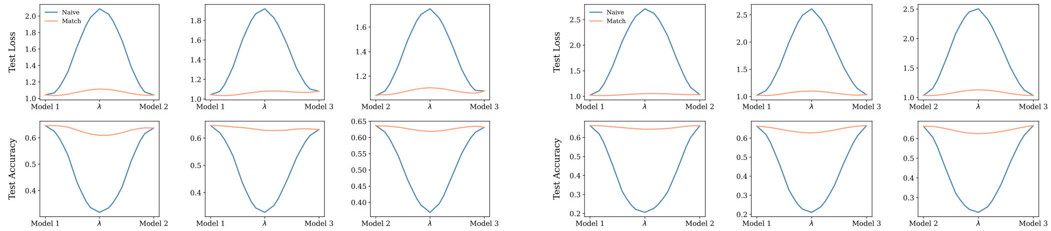
Figure 53: Linear Mode Connectivity for GPT2 with APE and RoPE on Wikitext103 with 12 layers and 3 heads



(a) 4 attention heads

(b) 8 attention heads

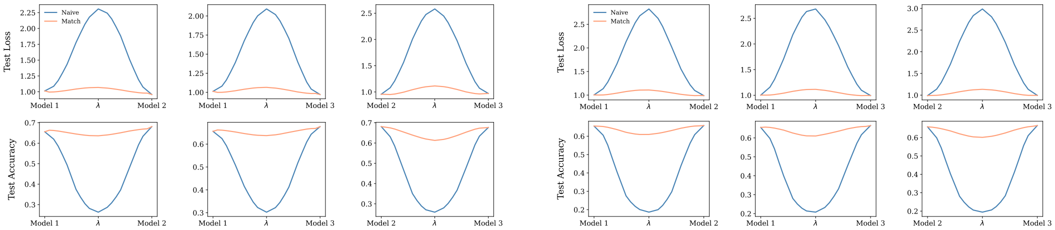
Figure 54: Linear Mode Connectivity for ViT-RoPE on MNIST with 2 layers



(a) 4 attention heads

(b) 8 attention heads

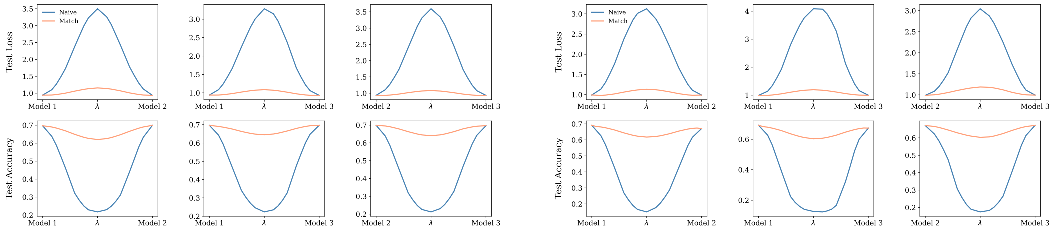
Figure 55: Linear Mode Connectivity for ViT-RoPE on CIFAR-10 with 2 layers



(a) 4 attention heads

(b) 8 attention heads

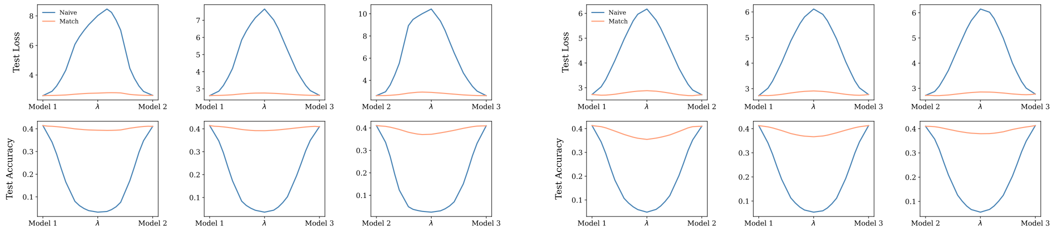
Figure 56: Linear Mode Connectivity for ViT-RoPE on CIFAR-10 with 4 layers



(a) 4 attention heads

(b) 8 attention heads

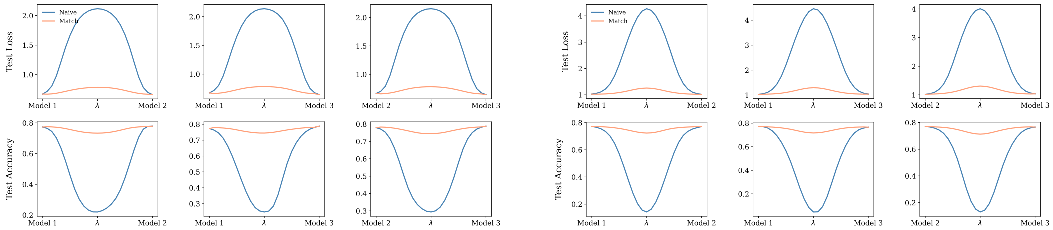
Figure 57: Linear Mode Connectivity for ViT-RoPE on CIFAR-10 with 6 layers



(a) 4 attention heads

(b) 8 attention heads

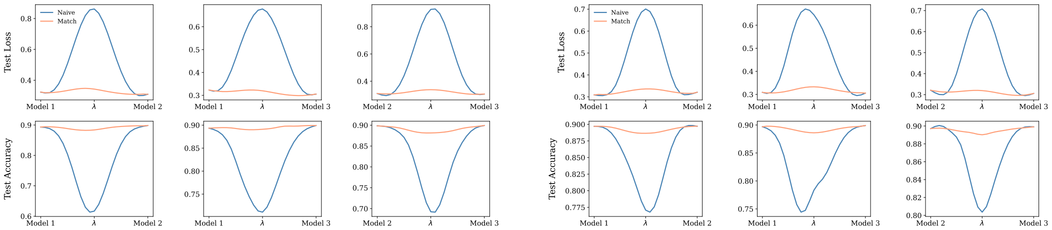
Figure 58: Linear Mode Connectivity for ViT-RoPE on CIFAR-100 with 6 layers



(a) CIFAR-10

(b) CIFAR-100

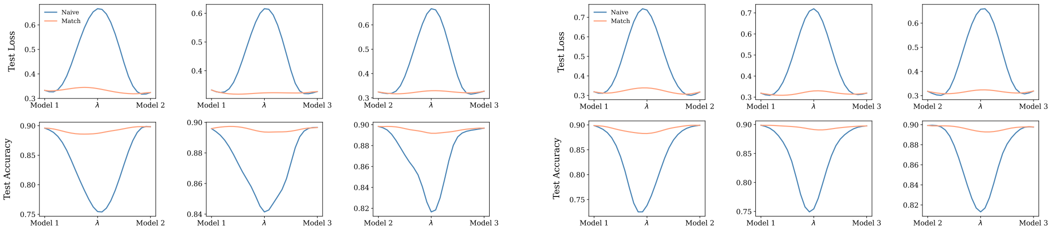
Figure 59: Linear Mode Connectivity for ViT-RoPE on ImageNet21k→CIFAR-10/100 with 12 layers and 6 heads



(a) 4 attention heads

(b) 8 attention heads

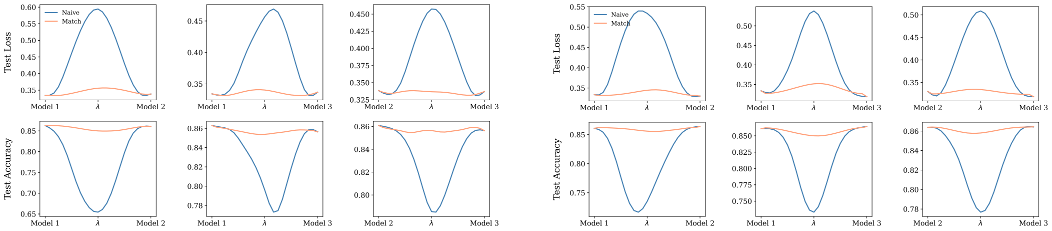
Figure 60: Linear Mode Connectivity for BERT-RoPE on AGnews with 2 layers



(a) 4 attention heads

(b) 8 attention heads

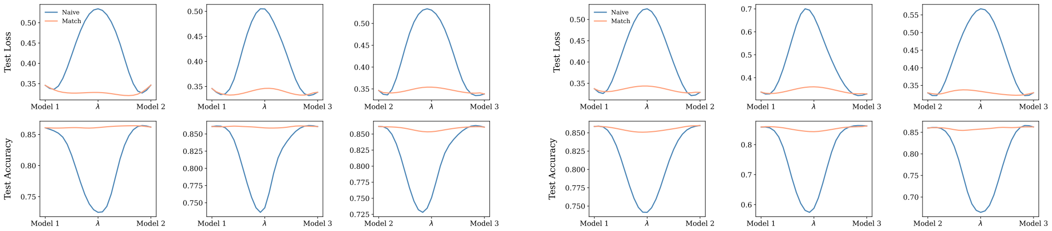
Figure 61: Linear Mode Connectivity for BERT-RoPE on AGnews with 6 layers



(a) 4 attention heads

(b) 8 attention heads

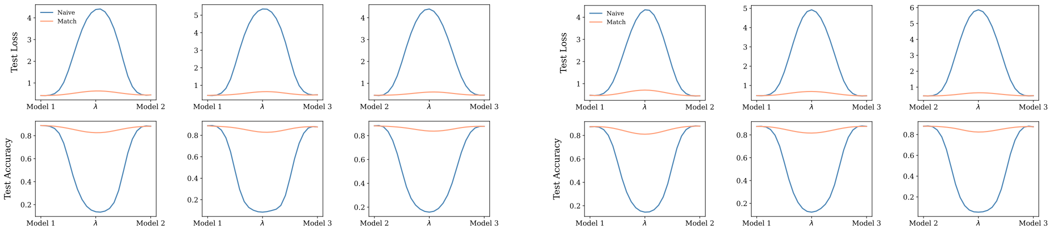
Figure 62: Linear Mode Connectivity for BERT-RoPE on IMDBreview with 2 layers



(a) 4 attention heads

(b) 8 attention heads

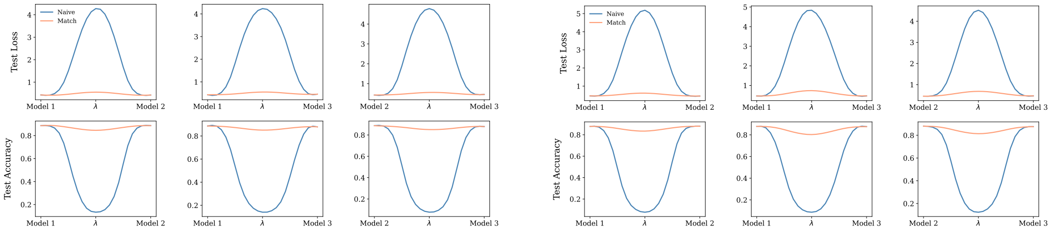
Figure 63: Linear Mode Connectivity for BERT-RoPE on IMDBreview with 6 layers



(a) 4 attention heads

(b) 8 attention heads

Figure 64: Linear Mode Connectivity for BERT-RoPE on DBPedia with 2 layers



(a) 4 attention heads

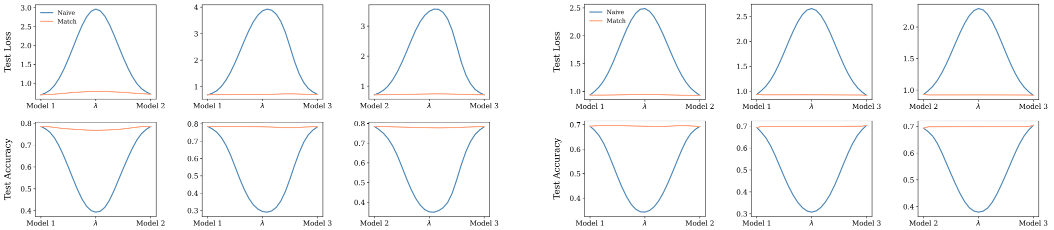
(b) 8 attention heads

Figure 65: Linear Mode Connectivity for BERT-RoPE on DBPedia with 6 layers

J.3 Linear Mode Connectivity for Transformer First Layer

Table 5: Experimental setups for LMC under first Transformer layer re-initialization. The table lists datasets, model depths, and attention head counts, PE type, along with references to figures.

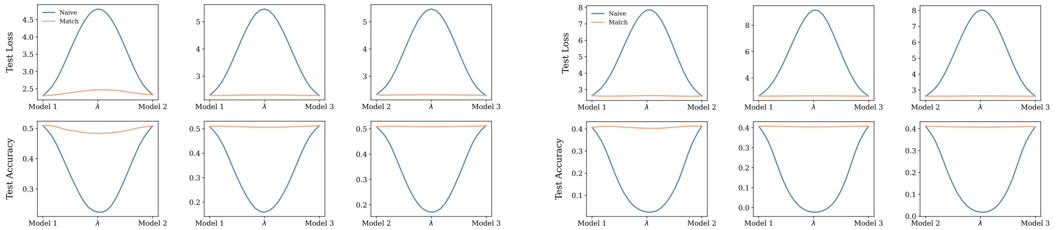
Dataset	Layers	Heads	APE	RoPE	Dataset	Layers	Heads	APE	RoPE
CIFAR-10	6	[8]	[66a]	[66b]	AGNews	6	[8]	[69a]	[69b]
CIFAR-100	6	[8]	[67a]	[67b]	DBPedia	6	[8]	[70a]	[70b]
ImageNet-1k	12	[12]	[68a]	[68b]	Wikitext103	12	[12]	[71a]	[71b]



(a) APE

(b) RoPE

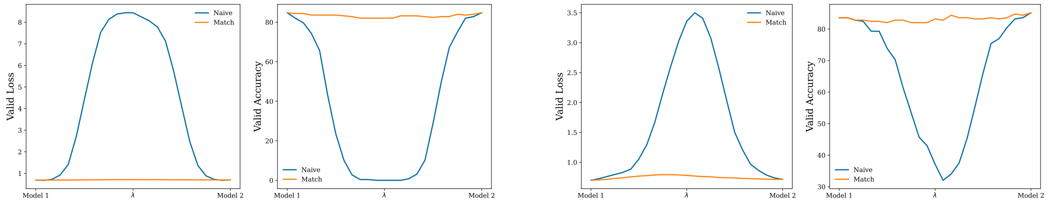
Figure 66: Linear Mode Connectivity for ViT with APE and RoPE on CIFAR-10 with 6 layers and 8 heads



(a) APE

(b) RoPE

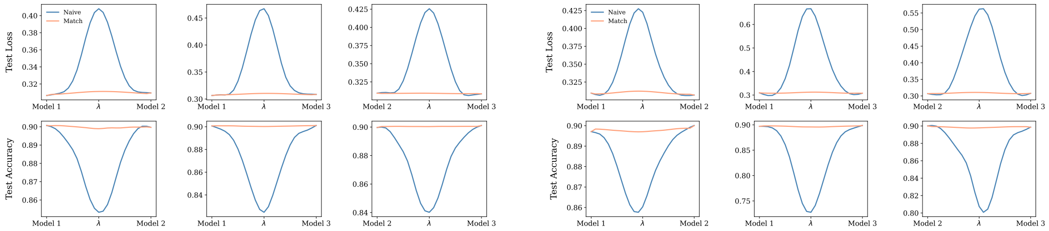
Figure 67: Linear Mode Connectivity for ViT with APE and RoPE on CIFAR-100 with 6 layers and 8 heads



(a) APE

(b) RoPE

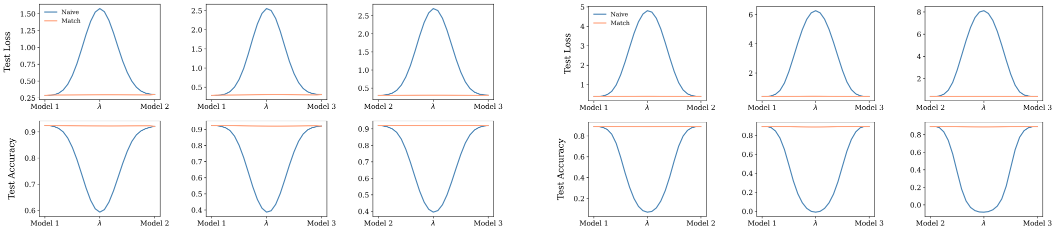
Figure 68: Linear Mode Connectivity for ViT with APE and RoPE on ImageNet-1k with 12 layers



(a) APE

(b) RoPE

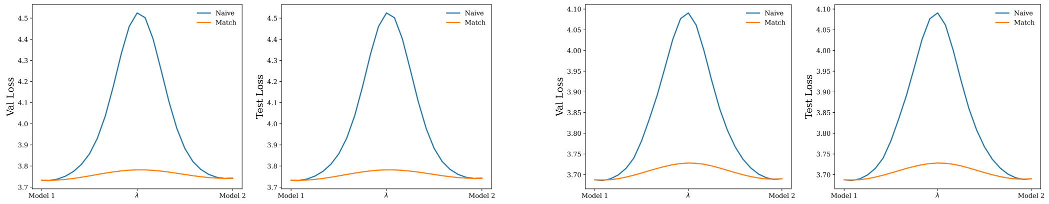
Figure 69: Linear Mode Connectivity for BERT with APE and RoPE on AGNews with 6 layers and 8 heads



(a) APE

(b) RoPE

Figure 70: Linear Mode Connectivity for BERT with APE and RoPE on DBPedia with 6 layers and 8 heads



(a) APE

(b) RoPE

Figure 71: Linear Mode Connectivity for GPT2 with APE and RoPE on Wikitext103 with 12 layers

J.4 Linear Mode Connectivity for Full Model

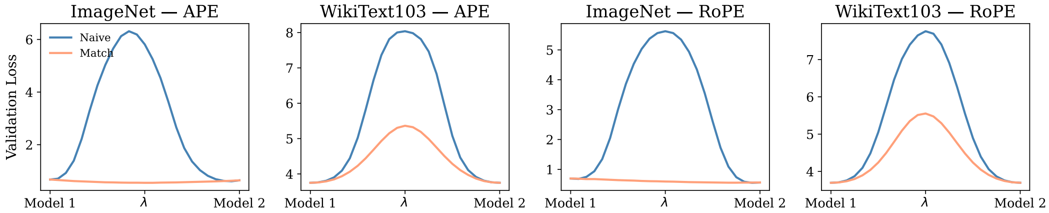
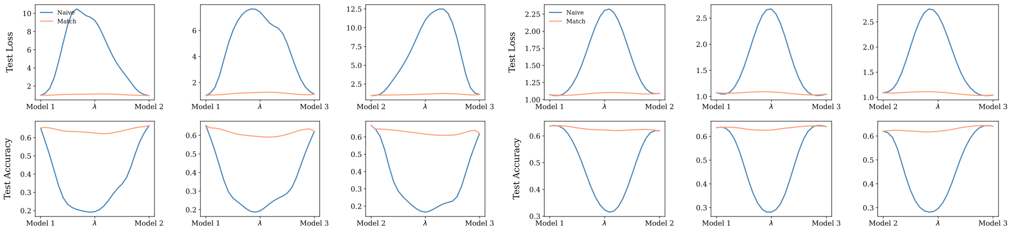


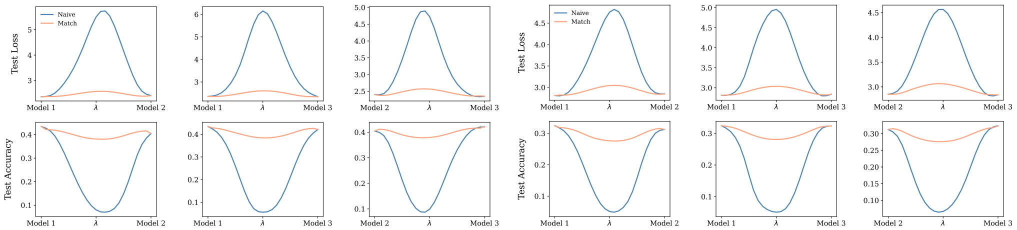
Figure 72: LMC interpolation plots for ViT on ImageNet-1K (subplots 1 and 3) and GPT-2 on WikiText103 (subplots 2 and 4), with APE and RoPE under full Transformer re-initialization.



(a) APE

(b) RoPE

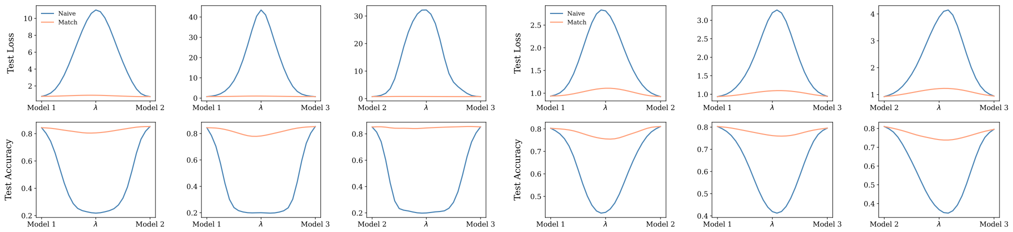
Figure 73: Linear Mode Connectivity for ViT with APE and RoPE on CIFAR-10 with 6 layers and 8 heads



(a) APE

(b) RoPE

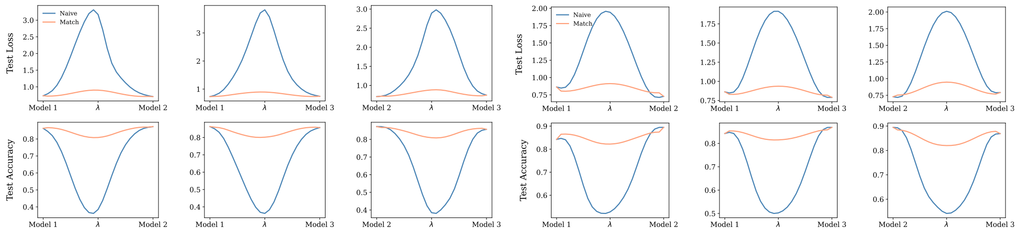
Figure 74: Linear Mode Connectivity for ViT with APE and RoPE on CIFAR-100 with 6 layers and 8 heads



(a) APE

(b) RoPE

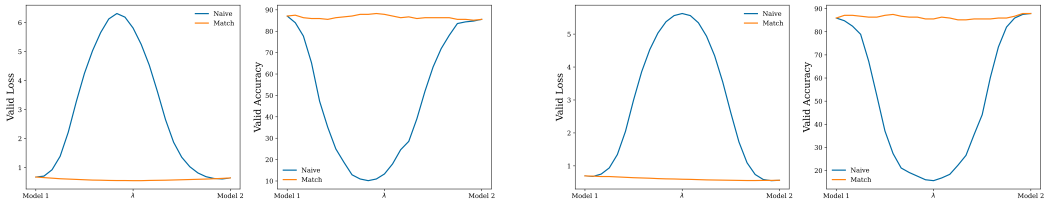
Figure 75: Linear Mode Connectivity for BERT with APE and RoPE on AGNews with 6 layers and 8 heads



(a) APE

(b) RoPE

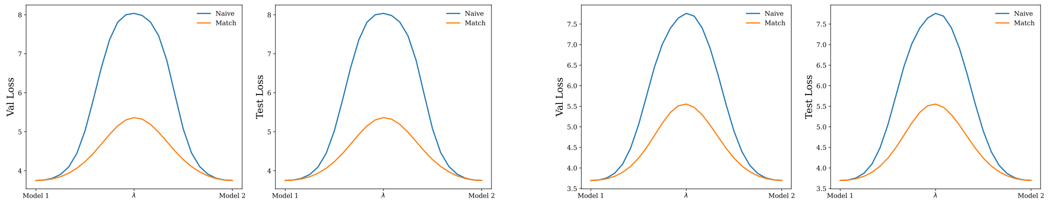
Figure 76: Linear Mode Connectivity for BERT with APE and RoPE on DBPedia with 6 layers and 8 heads



(a) APE

(b) RoPE

Figure 77: Linear Mode Connectivity for ViT with APE and RoPE on ImageNet-1k with 12 layers



(a) APE

(b) RoPE

Figure 78: Linear Mode Connectivity for GPT2 with APE and RoPE on Wikitext103 with 12 layers

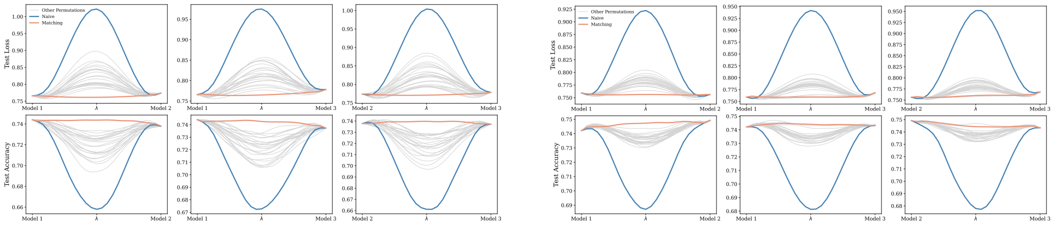
J.5 Ablation study on Head Permutation

We plot 24 head permutations, including the one selected by Stage 1 our method, with Stage 2 applied post-reordering for all permutation. For the 4-head case, this encompasses all possible

Table 6: Ablation study on head permutation

Dataset	No. layers	No. heads	APE Figure	RoPE Figure
CIFAR-10	2	[4, 8]	[79a, 79b]	[86a, 86b]
	6	[4, 8]	[80a, 80b]	[87a, 87b]
CIFAR-100	6	[4, 8]	[81a, 81b]	[88a, 88b]
IMDBreview	2	[4, 8]	[82a, 82b]	[89a, 89b]
	6	[4, 8]	[83a, 83b]	[90a, 90b]
DBPedia	2	[4, 8]	[84a, 84b]	[91a, 91b]
	6	[4, 8]	[85a, 85b]	[92a, 92b]

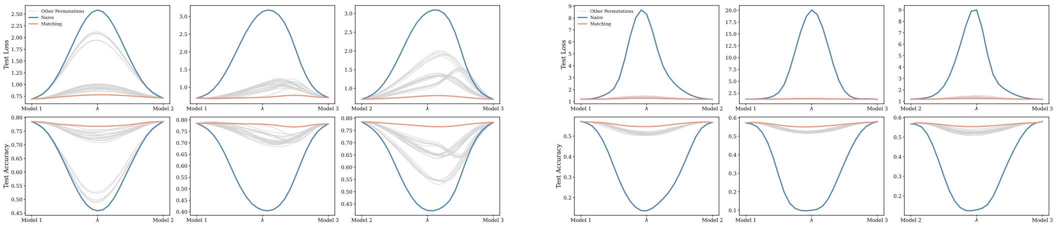
permutations (4! = 24). For the 8-head case, it includes 23 randomly sampled permutations along with the one chosen by our method.



(a) 4 attention heads

(b) 8 attention heads

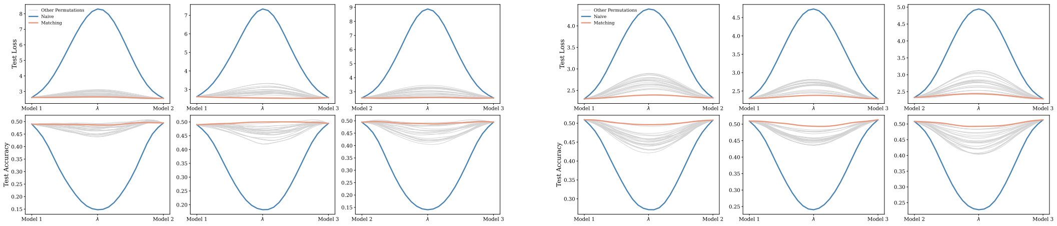
Figure 79: Linear Mode Connectivity for ViT on CIFAR-10 with 2 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

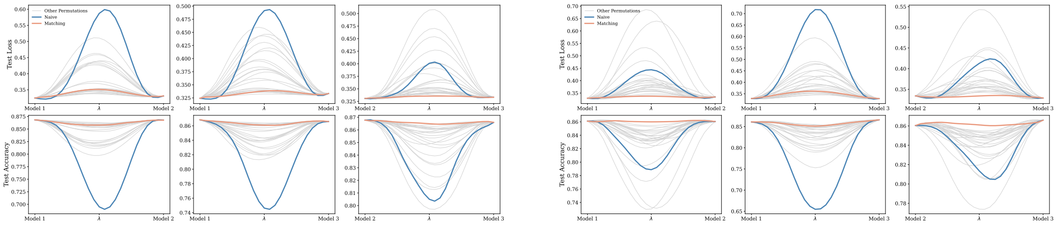
Figure 80: Linear Mode Connectivity for ViT on CIFAR-10 with 6 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

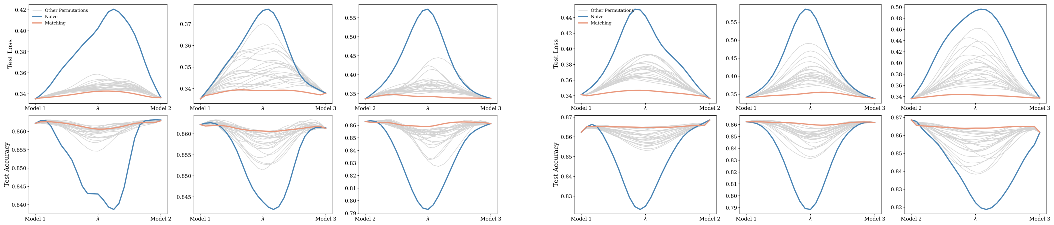
Figure 81: Linear Mode Connectivity for ViT on CIFAR-100 with 6 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

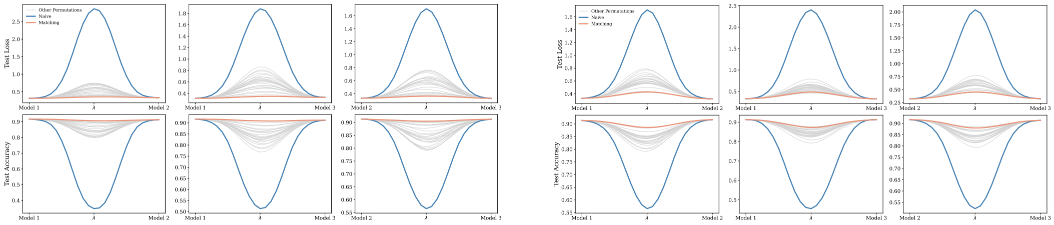
Figure 82: Linear Mode Connectivity for BERT on IMDBreview with 2 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

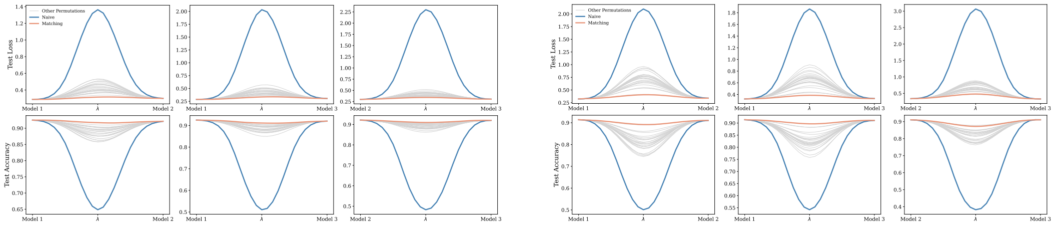
Figure 83: Linear Mode Connectivity for BERT on IMDBreview with 6 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

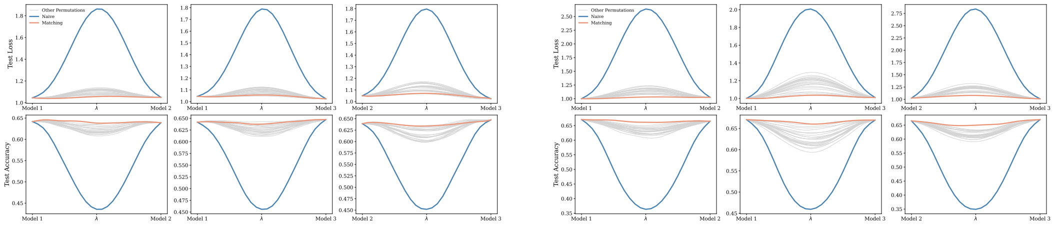
Figure 84: Linear Mode Connectivity for BERT on DBPedia with 2 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

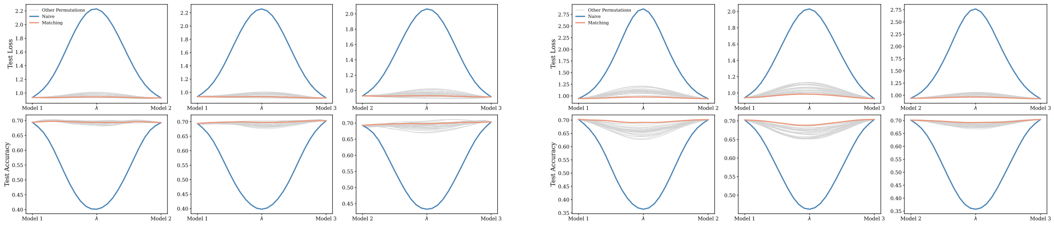
Figure 85: Linear Mode Connectivity for BERT on DBPedia with 6 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

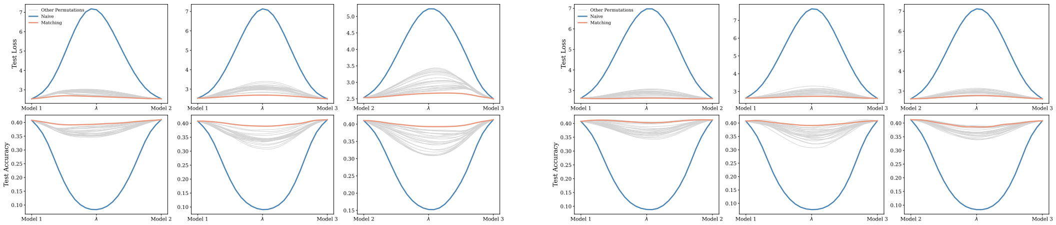
Figure 86: Linear Mode Connectivity for ViT-RoPE on CIFAR-10 with 2 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

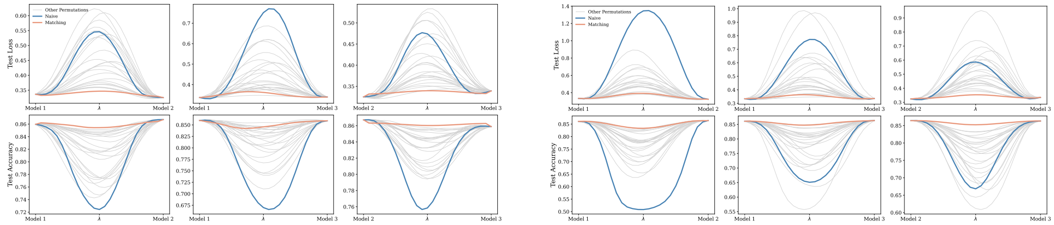
Figure 87: Linear Mode Connectivity for ViT-RoPE on CIFAR-10 with 6 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

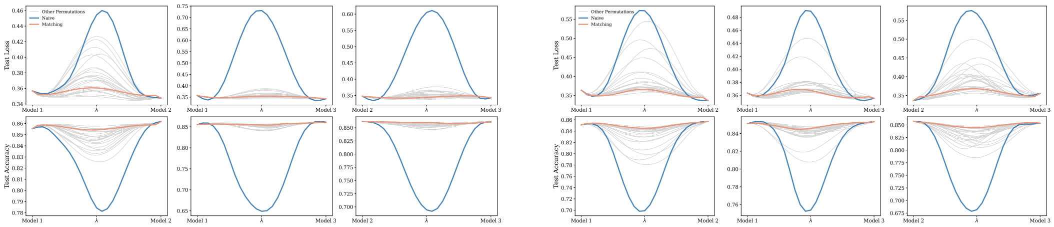
Figure 88: Linear Mode Connectivity for ViT-RoPE on CIFAR-100 with 6 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

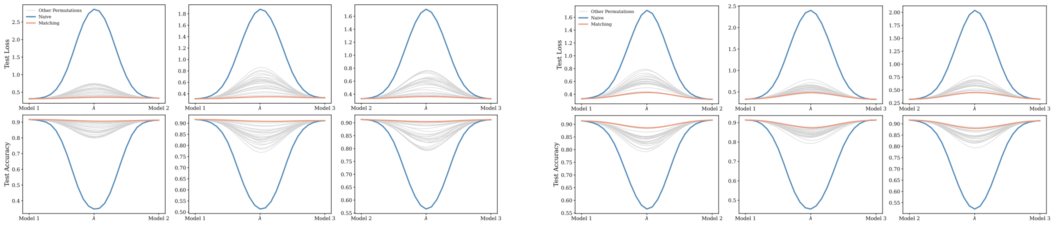
Figure 89: Linear Mode Connectivity for BERT-RoPE on IMDBreview with 2 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

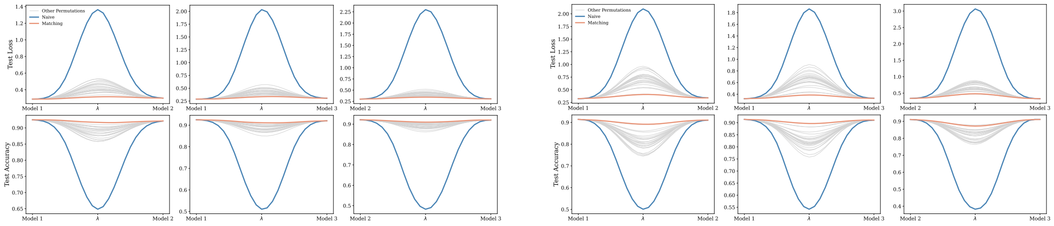
Figure 90: Linear Mode Connectivity for BERT-RoPE on IMDBreview with 6 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

Figure 91: Linear Mode Connectivity for BERT-RoPE on DBPedia with 2 layers (all head permutations)



(a) 4 attention heads

(b) 8 attention heads

Figure 92: Linear Mode Connectivity for BERT-RoPE on DBPedia with 6 layers (all head permutations)