

A Broader Impact

This work is fundamental research. While this work could lead to the discovery of better positional encodings and higher performing visual foundation models, the positivity or negativity of this impact is determined by the downstream task and not this work.

B Limitations

While our results do not prove relative embeddings to be detrimental, we believe them to be evidence that equivariance is not the reason for RoPE’s success. However, our experiments were performed in Vision where the number of tokens is limited compared to the long context lengths of NLP. Moreover, the datasets are not what many believe to be “at scale”. While Spherical RoPE and LieRE would intuitively favored at scale over Axial RoPE, as they have less inductive bias, it is unclear whether inductive bias and equivariance is favored at scale [9].

C Notation

Symbol / Term	Dimension	Meaning	Notes
\mathbf{x}_i	\mathbb{R}^D	Patch/token/content vector of token i	Raw input embedding
x_i	\mathcal{X}	Abstract content of token i	Raw input embedding
p_i	\mathbb{R}^M or \mathcal{P}	Position of token i , can be M -D or abstract \mathcal{P}	Scalar (1D) or vector (2D)
m	\mathbb{Z}	Modality index	e.g., x, y , time
M	\mathbb{Z}	Number, or space, of Modalities	
D	\mathbb{Z}	Hidden dimension	Number of pairs/triples/quadruples
T	\mathbb{Z}	Number of Tokens	
$\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$	$\mathbb{R}^{N \times D}$	Query, Key, Value Matrices	
\mathbf{q}	\mathbb{R}^N	$\mathbf{q}_i = \mathbf{W}_q x_i$	Query vector
\mathbf{k}	\mathbb{R}^N	$\mathbf{k}_j = \mathbf{W}_k x_j$	Key vector
\mathbf{v}	\mathbb{R}^N	$\mathbf{v}_j = \mathbf{W}_v x_j$	Value vector
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	$\mathbb{R}^{T \times N}$	Query, Key, Values	T tokens, D latent dimensions
$\varphi(x, p)$	$\mathcal{X} \times \mathcal{P} \rightarrow \mathbb{R}^D$	Positional Encoding function	
\mathbf{Z}	$\mathbb{R}^{T \times N}$	Output of Attention	$\mathbf{Z} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$
$a(i, j)$	\mathbb{R}	Attention weight	Softmax of attention scores
$\alpha(\mathbf{q}, \mathbf{k})$	\mathbb{R}	Attention score	Inner product $\mathbf{q}^\top \mathbf{k}$
ω_d / λ_d	\mathbb{R}	Rotation frequency for dimension d	Equivalent to eigenvalue of generator
\mathbf{q}_d	$\mathbb{R}^{2/3/4}$	Query pair/triple/quadruple at dimension d	After RoPE or LieRE applied
$\mathbf{R}_{\omega_d p}$	$\mathbb{R}^{2 \times 2}$	2×2 rotation matrix	Rotation based on frequency and position

Table 3: Summary of Notations and Key Concepts

D Literature Review

D.1 Natural Language Processing

In natural language, positional encoding has been used to break the permutation, “bag of words”, symmetry [71]. Although this could be done by learning a vector per position, this is both memory-expensive for large context sizes making it practical to apply to only the first layer. Moreover, it does not allow for extrapolation at test time to context sizes beyond training. Thus, it is favorable to perform positional embeddings with a predictable deterministic function. One way of doing this is to make the attention relative with local receptive fields, as is done implicitly in convolutional neural networks [12]. Sinusoidal positional embeddings were proposed due to approximate local and shift-invariant properties of Random Fourier Features [55]. Since sinusoidal, other methods have been proposed to get guaranteed shift invariance by explicitly parameterizing based on distance [62, 51, 52]. However, these methods require a positional embedding for every pair of positions which is not supported by many of the efficient attention optimizations such as Flash attention [16] [3].

Rotary Positional Embeddings (RoPE) have become the staple in NLP having recently been adopted by many of the large language models [74, 20, 69, 40, 25]. However, these methods also use causal

masking, which has been shown to allow models with no positional embedding to recover absolute position [22, 80, 73, 30]. This has led to questions on the importance of relative position [5]. In language, there has also been extensions to RoPE proposed through NTKs and kernel methods [11]. However, these methods have not, to our knowledge, seen use in vision.

D.2 Vision and Video

Vision transformers were introduced in Dosovitskiy et al. [17] and, though they tried sinusoidal position encodings, found learnable position encodings to perform best. For convolution-esque models such as SWin transformers, relative positional encodings have been popular [44, 14]. More recently, RoPE has been shown to be an efficient and simple way to have relative embeddings and has been extended to 2D using Axial and Mixed RoPE. Going beyond 2D to Video data, Axial RoPE has become increasingly popular. The extension was first attributed to Wang et al. [74] as 3D-RoPE or M-RoPE, leading to two separate Video-RoPE papers from Wei et al. [75] and Liu et al. [45]. Both of these focus on the order of the position enumeration and interleaving positions. However, this should not be a problem if frequencies are not deterministic, *or* if frequencies are indexed by both d and modality m as done in Eq 13. We highly recommend using either Mixed RoPE or LieRE which extend naturally for videos.

LieRE embeddings have thus far been the most general form of RoPE to N -D. However, Schenck et al. [58] has claimed the method to have a large memory footprint and proposed STRING. This paper, a preprint released concurrently with the writing of this manuscript, follows much of the same math as this paper. However, they did not recognize that an orthogonal matrix is implicitly learned by the query and key matrix. Moreover, their method relies on commuting Lie algebras. From our insights in Section 3, their method can likely be viewed as a slower implementation of N -D Mixed-RoPE.

It is also worth noting that positional encodings have also been explored within vision through the area of Neural Fields [78]. Traditional coordinate MLPs have been found to be biased toward low-frequency functions [68] leading to more advanced positional encodings such as Random Fourier Features [55] or sinusoidal activation functions [63]. These implicit functions have been used to encode attention and message passing in graph neural networks with recent work being put in to make these functions equivariant to symmetry transformations [57, 8, 33].

D.3 Graphs and AI in Science

Positional encodings are well studied within graph neural networks [39, 50]. Graphs are limited in their expressivity up to the Weisfeiler-Lehman (WL) graph isomorphism test [79], so positional encodings can break the isomorphism symmetry [23, 81]. Within this community, they propose *spectral attention* and graph Laplacians for positional encoding [35]. These methods seem extremely close to our analysis of RoPE, but from a very different perspective. We show that the frequencies of RoPE can be interpreted as the eigenvalues of an orthogonal transformation by taking the spectral decomposition.

In an overlapping vein, relative position encodings have been studied in terms of equivariant graph neural networks, often for scientific disciplines such as molecular physics [8, 60] or drug discovery [26]. One method to achieve equivariance is through defining relative coordinate frames [34]. This corresponds to the learned relative positional method described in Shaw et al. [62], but can be generalized to higher dimensions and different transformation using bi-invariant distance functions [6, 33, 76]. The message-passing functions of these works correspond to a generalization of attention scores [19].

However, even in these tasks with physics-grounded symmetries, the need for equivariance is hotly debated. While AlphaFold [26] was originally touted as the example of the success of equivariant inductive biases in science, AlphaFold 3 [1] explicitly stated that they benefited from removing this inductive bias at scale. However, while the harm of inductive bias at scale is the prevalent zeitgeist, it is not an established fact [9].

632 D.4 Computational Neuroscience

633 Coupled oscillators have become a growing area of interest within computational neuroscience
634 [31, 32, 65]. By observing the projection of the RoPE circles onto the real axis, one can interpret
635 RoPE as time progression in D uncoupled, undamped harmonic oscillators. This perspective naturally
636 connects RoPE to Löwe et al. [46]’s series of papers on complex autoencoders and their extensions
637 [47, 48].

638 In another, vein of research, there has been some work in hyper-dimensional computing[27, 28] in
639 Phasor and Residue VSAs [37] which represent concepts as rotations around unit circles in high-
640 dimensional spaces. These representations have strong connections with RoPE. Additionally, progress
641 has been made in hypothesizing how biological neural networks encode positional knowledge with
642 hexagonal grid cells, which can be represented as a discrete sum of three periodic functions oriented
643 at the cubic roots of unity[64].

644 D.5 Generality of RoPE

645 The generality of RoPE has been found by others. Schenck et al. [58], Su [66], and Liu and
646 Zhou [41] all propose proofs similar to Proposition 1. However, Schenck et al. [58] miss that the
647 orthogonal transformation can be incorporated into key matrix. Liu and Zhou [41] and Su [66] take
648 the assumption of *reversibility*, which leads to the independent eigenvalue assumptions of Axial RoPE.
649 All three works take the assumption of an abelian subgroup –ie commutative generators, – but miss
650 the generality of Mixed RoPE. While Su [66] propose quaternions – i. e. spherical rotations – as a
651 direction, they immediately dismiss it as a *no-go* because they lack equivariance. This exemplifies the
652 “circular argument,” where equivariance is assumed to be necessary because work will not investigate
653 non-equivariant positional encodings because equivariance is necessary.

654 Because our derivation was found independently of these works and the previous works are, to our
655 knowledge, not published, we have left in Proposition 1. We would like to acknowledge their work,
656 but retain the flow of this paper.

Positional Encoding	Vision	Learned	Extrapolation	QK Separable	Relative	Linear Flow	Used In
Absolute (Sinusoidal)	✗	✓/✗	✓	✓	✓	✗	Transformer[71]
Absolute (Learned)	✓	✓	✗	✓	✓	✗	BERT, GPT, ViT[17]
Absolute (Random-Fourier)	✗	✗	✓	✓	✗	✓	FNet[38], Performer [13]
Relative (Learned)	✗	✓	✗	✗	✗	✗	Transformer-XL, T5 [53]
ALiBi	✗	✓/✗	✓	✓	✓	✓	LLaMA 2 [20], ALiBi [52]
NoPE	✗*	✗	✓*	✓*	✓*	✓*	LLaMA 4 [2]
Rotary (RoPE)	✗	✗	✓	✓	✓	✓	Contemporary LLMs [74, 20, 69, 25]
Axial RoPE	✓	✓/✗	✓	✓	✓	✓	VisionLLaMA[15], Qwen2[74], VideoRoPE[75]
Mixed RoPE	✓	✓	✓	✓	✓	✓	Heo et al. [24]
LieRE	✓	✓	✓	✓	✗	✓	[49]
Spherical RoPE	✓	✓/✗	✓	✓	✗	✓	Ours
Uniform RoPE	✓	✓/✗	✓	✓	✓	✓	Ours

Table 4: Comparison of positional encoding methods in transformer models

E Positional Encoding Properties

Rotary positional embeddings were derived in Su et al. [67] by drawing equations from assumed properties. While these appear as arithmetic assumptions and equations in their work, we formalize what properties these assumptions imply and why we may choose these assumptions in this section. In their paper, to derive their equations, they use equivariance (relativity), query-key separability of the positional encoding, linearity and incompressibility, locality, and query-key symmetry.

1. Equivariance/Relativity: Attention score should be affected only by the relative position of two tokens, i. e. have the form

$$\alpha(x_i, x_j, p_i, p_j) = \hat{\alpha}(x_i, x_j, p_i - p_j). \quad (26)$$

2. Key-query separability: The positional encoding, φ , of the query should not depend on the position of the key

$$\alpha(x_i, x_j, p_i, p_j) = \bar{\alpha}(\varphi(x_i, p_i), \varphi(x_j, p_j)) \quad (27)$$

3. Linearity: The positional encoding should be a linear flow, see Appendix E.3. Namely,

$$\varphi(\varphi(x, p_i), p_j) = \varphi(x, p_i + p_j). \quad (28)$$

4. Locality: The attention score between two tokens should decay with distance

$$\lim_{|p_i - p_j| \rightarrow \infty} \alpha(x_i, x_j, p_i, p_j) = 0 \quad (29)$$

E.1 Relativity and Equivariant

We use the term *equivariant* interchangeably with *relative*. Strictly speaking, one should specify the transformation or group you would like to be relative to, e. g. shift/rotation or $SO(2)$. As previous literature always refers to relative positional bias in terms of shifts/translations, in the main text, this is what we mean. We use the term equivariance to be the generalization of relativity beyond language because we would like to refrain from using the term "relativity" to describe the property of being a relative PE too often due to its connotation within theoretical physics. First, we define relative in the case of positional encodings in language as

$$\alpha(x_i, x_j, p_i, p_j) = \hat{\alpha}(x_i, x_j, p_i - p_j). \quad (30)$$

In the rest of this section, we mathematically explore where this equation comes from.

The behavior we are trying to capture is that if we renumber the words in the sentence, it should not affect the attentions score. Intuitively, if a text is padded with spaces at the beginning, that will not have a significant effect on the meaning of the sentences. We can ensure this by colloquially saying that the attention between two words should depend on the distance between them. Notice, that strictly speaking this is not a proper distance, since it can be negative; it is, instead, a *signed* distance function. Though this may seem pedantic in one dimension, in two dimensions defining a distance function is less unique. For example, one may choose \mathbb{L}_1 or \mathbb{L}_2 distance metrics. Because distance functions are more nebulous, it makes more sense to define relative in terms of the transformations that we would like our attention score to be independent of.

$$\alpha(x_i, x_j, p_i, p_j) = \alpha(x_i, x_j, T(p_i), T(p_j)). \quad (31)$$

687 These transformations can be combined to generate a set of transformations which leave the attention
688 score unchanged, or *symmetric*. This set has the mathematical properties of a group and is known as
689 a symmetry group. We can index transformations by elements in the symmetry group, $g \in G$, and let
690 the elements act on

$$\alpha(x_i, x_j, p_i, p_j) = \alpha(x_i, x_j, g.p_i, g.p_j). \quad (32)$$

691 As an example, g could represent an angle, θ , and it may act on a vector \mathbf{p} as a rotation $g.\mathbf{p} = \mathbf{R}_\theta \mathbf{p}$.

692 Connecting everything back to Eq. 30, Noether’s theorem states that any continuous symmetry can
693 be expressed as a conservation law. This allows us to introduce bi-invariant function [33, 76], or
694 “Noether charge”, $\beta(p_i, p_j)$, that is invariant under the group action,

$$\beta(p_i, p_j) = \beta(g.p_i, g.p_j) \implies \beta(p_i, p_j) - \beta(g.p_i, g.p_j) = 0. \quad (33)$$

695 Thus, we can express our symmetry group through isodistances of β ,

$$\alpha(x_i, x_j, p_i, p_j) := \hat{\alpha}(x_i, x_j, \beta(p_i, p_j)). \quad (34)$$

696 For example, we can pick the function

$$\beta(p_i, p_j) = p_i - p_j = (p_i - p_0) - (p_j - p_0) = \beta(p_i - o, p_j - p_0) \quad (35)$$

697 If we were to define $\beta(p_i, p_j) = |p_i - p_j|$, then we would additionally be equivariant to reflection of
698 the order of tokens in a sentence. If we trivially define $\beta(p_i, p_j) = C$, then we arrive at bag of words,
699 or no positional encoding (NoPE). For a list of common transformations and their corresponding
700 bi-invariants see Theorem 1 of Bekkers et al. [6].

701 E.2 Query-Key Separability

702 Query and key separability is important for efficiency reasons. If we can decompose our positional
703 encoded attention score as,

$$\alpha(x_i, x_j, p_i, p_j) = \alpha(\varphi(x_i, p_i), \varphi(x_j, p_j)) \quad (36)$$

704 then we can pre-compute the positional encoding for the queries and keys on time making the
705 computation $O(T)$. If the positional encoding is not separable, then it will need to be computed for
706 every pair, (i, j) [44, 54, 62]. Although there are many symmetries that can be exploited to make this
707 not a quadratic computation, it removes the symmetries exploited by efficient attention mechanisms
708 [7, 13, 29].

709 E.3 Linear Flow Property

710 The property of being a “flow” was first proposed in Liu et al. [42], however it is not often discussed.
711 It is a property inherently present in RoPE[67], LieRE[49] and ALiBi [52] embeddings, specifically
712 as a *linear flow*.

713 We use the term *linear flow* for this property because the embedding can be found by repeated
714 application of a linear function. However, the term “linear” this is a small misnomer because it is
715 only *locally* linear. We define a *flow* as function

$$\varphi : \mathbb{R}^N \times \mathbb{R} \rightarrow \mathbb{R}^N \quad (37)$$

716 such that for all $x \in X$ and $p_1, p_2 \in \mathbb{R}$, the following conditions hold:

717 1. Initial condition (identity at time zero):

$$\varphi(0, x) = x \quad (38)$$

718 2. Group property (flow property):

$$\varphi(\varphi(\mathbf{x}, p_1), p_2) = \varphi(\mathbf{x}, p_1 + p_2) \quad (39)$$

719 3. Continuity (or differentiability): φ is continuous with respect to its variables, depending on
720 the context

Strictly speaking, continuity is not necessary for positional encodings as positions tend to be integer values. What we really wish to capture with this property is for the positional encoding to be recursively defined. It may be strange to wish to apply the positional encoding multiple times; however, by having the positional encoding as an endomorphism it can allow for more predictable behavior when extrapolating to larger contexts, which we suspect helps the model train.

We define a position embedding to be a *linear flow* if the flow has the form:

$$\varphi(\mathbf{x}, \Delta p) = \mathbf{A}\mathbf{x}, \quad (40)$$

for $\mathbf{A} \in \mathbb{R}^{N \times N}$ and $\mathbf{x} \in \mathbb{R}^N$, where Δp is the increment rate for position. By Eq. 39, any position $p := p_0 \Delta p$ can then be attained by,

$$\varphi(\mathbf{x}, p) = \mathbf{A}^{p_0} \mathbf{x}. \quad (41)$$

This can be seen as a *geometric series* if \mathbf{A} is a scalar as seen in Press et al. [52]. If we let Δt become infinitesimal, then we can express the recurrence relationship as the ODE,

$$\frac{\partial \varphi}{\partial t} = \mathcal{A} \varphi \quad (42)$$

which we can integrate to get,

$$\varphi(\mathbf{x}, p) = \exp(\mathcal{A}p) \mathbf{x} \quad (43)$$

This \mathcal{A} is our *generator* of the flow, which is also a generator for a *matrix Lie algebra*, which we focus on in the main text. The matrix exponential, $\exp : \mathbb{R}^{N \times N} \rightarrow \mathbb{R}^{N \times N}$, can be unstable for long contexts; similar to the scalar exponential function e^{xp} , the function can quickly become large for high values of x . However, this can be stable value $x = 0$, since it always results in one. Similarly, the matrix exponential can be stable if the divergence of the flow – trace of the generator – is zero. We call flow “incompressible” or “divergence-free” if the trace of \mathcal{A} is zero, making the determinant of \mathbf{A} unit. If fluid dynamics, this is called *incompressibility*. For fluids, this implies that the flow conserves mass.

If there are more than one generator of the Lie group, \mathcal{A}_1 and \mathcal{A}_2 , then Eq. 39 must be modified to,

$$\varphi(\varphi(\mathbf{x}, \mathbf{p}_1), \mathbf{p}_2) = \varphi(\mathbf{x}, \mathbf{p}_1 \circ \mathbf{p}_2), \quad (44)$$

where \circ is the group product. By the Baker–Campbell–Hausdorff formula, $\exp \mathcal{A}_1 p_1 \exp \mathcal{A}_2 p_2 = \exp \mathcal{A}_1 p_1 + \mathcal{A}_2 p_2$ iff the commutator of $\mathcal{A}_1 p_1$ and $\mathcal{A}_2 p_2$ is zero, i. e. the matrices commute. If they do commute, then

$$\varphi(\varphi(\mathbf{x}, \mathbf{p}_1), \mathbf{p}_2) = \varphi(\varphi(\mathbf{x}, \mathbf{p}_2), \mathbf{p}_1) \implies \varphi(\mathbf{x}, \mathbf{p}_1 \circ \mathbf{p}_2) = \varphi(\mathbf{x}, \mathbf{p}_2 \circ \mathbf{p}_1) \quad (45)$$

thus making \circ commutative and having the same properties as addition, $\circ := “+”$, and Eq. 39 will hold. In this case, the group/flow is known as an *abelian* Lie group, or *abelian flow*. However, if they do not commute, then \circ will not commute and they are known as *non-abelian*. This also makes the flow *non-integrable*.

E.4 Locality

Locality is often conflated with relativity. The general idea is that tokens far from each other should be independent of one another – i. e. attention should decay as distance grows. This often motivates the definition

$$\lim_{|p_i - p_j| \rightarrow \infty} \alpha(x_i, x_j, p_i, p_j) = 0 \quad (46)$$

for $p_i, p_j \in \mathbb{R}$ and $x_i, x_j \in \mathbb{R}^D$. However, this definition is *both* relative and local. We instead define local as,

$$\lim_{|p_i - p_0| \rightarrow \infty} \alpha(x_i, x_j, p_i, p_0) = 0. \quad (47)$$

The difference being that p_0 is the *origin* position. If an embedding is relative, then the origin is arbitrary and can be defined as p_i or p_j . In Press et al. [52], they define the origin vector as the next word. However, they can only do this because of the causal mask.

In general, the most natural way to measure locality is through the concept of the quantum mechanical concept of the *variance of an operator*. We will simply use exponential decay, but we point interested readers to Chapter 3 of Griffiths [21]. This formalism works for RoPE as it is a linear transformation and the attention mechanism defines a Hilbert space.

To be clear, RoPE and LieRE are *not* relative embeddings. This was shown for RoPE in Barbero et al. [5]. Because they are orthogonal matrices, they have unit determinant, which naturally precludes locality.

764 E.5 Other properties

765 For completeness, there are two additional assumptions that are common.

766 **Adjoint symmetry of the Positional Encoding** We implicitly assume that the positional encoding
 767 is symmetric for the query and key. That is, we assume that the query and key are from the same
 768 domain, so the positional encoding has the same representation. More generally, the positional
 769 encoding can act differently on the query and key,

$$\alpha(\bar{\varphi}(x_i, p_i), \varphi(x_j, p_j)) = \alpha(\varphi(x_i, p_i), \varphi(x_j, p_j)), \quad (48)$$

770 where $\bar{\varphi}$ is the positional encoding function for queries. More generally, we can have a relative
 771 embedding by letting $\bar{\varphi}$ act on queries differently from the keys. For example, if we let

$$\varphi(x, p) = \exp(\Lambda p) \quad \bar{\varphi}(x, p) = \exp(-\Lambda p), \quad (49)$$

772 where Λ is a diagonal matrix. We end up with,

$$\alpha(\bar{\varphi}(x_i, p_i), \varphi(x_j, p_j)) = \mathbf{q}_i^\top \exp(\Lambda(p_j - p_i)) \mathbf{k}_j, \quad (50)$$

773 where RoPE can be interpreted as a simple harmonic oscillator, by weakening the symmetry require-
 774 ment, one could incorporate damping. This can also be used to incorporate graph Laplacian positional
 775 encodings into the framework.

776 **Reversibility** Reversibility means that the positional encoding is an injective map – that is, every
 777 coordinate is mapped to a unique rotation, thus position can be recovered. This property is important
 778 in Liu and Zhou [41] and Su [66] to derive Axial RoPE. While it prevents Eq. 11, it is necessary only
 779 for the $D = 1$ case. More generally, Mixed RoPE can learn an injective map for large D . Moreover,
 780 while having a “lossless” positional encoding is nice mathematically, its practical utility has yet to be
 781 soundly justified, especially if the positional encoding is learnable.

F Fast Implementation

Because GPUs are not well optimized for batch batch matrix multiplications with small matrices, we follow a vectorized implementation for Spherical RoPE similar to the “fast implementation” proposed in Su et al. [67].

First, apply the rotation directly on after the other:

$$z_d[1] = \cos(\omega_y p_y) z_d[1] - \sin(\omega_y p_y) z_d[3] \quad (51)$$

$$z_d[3] = \sin(\omega_y p_y) z_d[1] + \cos(\omega_y p_y) z_d[3], \quad (52)$$

then

$$z_d[2] = \cos(\omega_x p_x) z_d[2] - \sin(\omega_x p_x) z_d[3] \quad (53)$$

$$z_d[3] = \sin(\omega_x p_x) z_d[2] + \cos(\omega_x p_x) z_d[3]. \quad (54)$$

Where steps 51 and 52 happen simultaneously, and steps 53 and 54 occur at the same time.

G Experimental Setup

Models We use the ViT-S backbone from the timm library [77]. The network always has a depth of 12. We keep N as close to constant across models as we can. For CIFAR100, the embedding dimensions are changed from $64 \times N_{\text{heads}}$ to $60 \times N_{\text{heads}}$ to be compatible with pairs, triplets and quadruples. For ImageNet, we make the embedding dimension $63 \times N_{\text{heads}}$ for Spherical RoPE and $64 \times N_{\text{heads}}$ for other methods. For classification, we use a class token to pool the tokens and predict. Unlike the patch tokens, the class token is not affected by any positional encoding.

CIFAR100 All experiments on CIFAR100 were performed on one A100 GPUs with a batch size 256. We use a patch size of 4×4 on the original image size 32×32 . The training uses heavy regularization and augmentations including dropout, MixUp[84] and CutMix [83]. The models are trained for 400 epochs, taking ~ 40 seconds per training loop.

ImageNet All experiments on ImageNet1k were performed on four A100 GPUs with a batch size 256. We used cosine learning rate with a learning rate of $3e-3$ for 200 epochs with 5 epochs of linear warm-up. We used a patch size of 16×16 on the cropped and resized 224×224 image after applying 3-Augment [70]. We use the LAMB [82] optimizer. All experiments took ~ 20 hrs with ~ 5 to 8 minutes to complete a training loop depending on method.

Positional Encodings For testing with different resolutions, the images from ImageNet’s validation set were normalized, resized and cropped. On training, the patches were assigned position $[-\pi, \pi]$ and for evaluation, the patch positions were extrapolated to the range $[-\frac{P}{P_0}\pi, \frac{P}{P_0}\pi]$. For Learned APE, the positional embeddings are instead interpolated. The fixed frequencies were given by $\omega_d = 1/100^{2d/D}$, where d is the index of the pair/tuple/quadruple. One frequency is shared between both x and y in our implementation of Axial RoPE.

Table 5: Hyperparameters for ImageNet-1K Training

Category	Setting
Model Architecture	
Patch Size	16x16
Heads	6
Latent Dimension	64 (63 for Spherical) \times Heads
Depth	12
Pooling	[CLS]
Stochastic Depth	No
Dropout	No
LayerScale	1
Optimization	
Optimizer	LAMB [82]
Base Learning Rate	4e-3
Weight Decay	0.05
Learning Rate Schedule	Cosine Decay
Warmup Schedule	Linear
Warmup Epochs	5
Epochs	200
Batch Size	512
Gradient Clipping	✓
Precision and Backend	
Precision	Mixed (bfloat16)
Backend	torch.autocast
Data Augmentation - Train	
Crop	RandomResizedCrop (192 \rightarrow 224)
Flip	✓
3-Augment	✓
Color Jitter	(0.3, 0.3, 0.3, 0.0)
Mixup [84]	✗
Cutmix [83]	✗
Normalization	ImageNet-1K Statistics
Data Augmentation - Test	
Resize	Resize \rightarrow Resolution
Crop	CenterCrop
Normalize	ImageNet-1K Statistics

Table 6: Hyperparameters for CIFAR100 Training

Category	Setting
Model Architecture	
Patch Size	16x16
Heads	12
Latent Dimension	$60 \times \text{Heads}$
Depth	12
Pooling	[CLS]
Stochastic Depth	0.1
Dropout	0.1
LayerScale	✓
Optimization	
Optimizer	LAMB [82]
Base Learning Rate	4e-3
Weight Decay	0.05
Learning Rate Schedule	Cosine Decay
Warmup Schedule	Linear
Warmup Epochs	5
Epochs	400
Batch Size	1024
Gradient Clipping	✓
Precision and Backend	
Precision	Mixed (bfloat16)
Backend	torch.autocast
Data Augmentation - Train	
Crop	RandomResizedCrop (32)
Flip	✓
3-Augment	✓
Color Jitter	(0.3, 0.3, 0.3, 0.0)
Mixup [84]	0.8
Cutmix [83]	1.0
Normalization	CIFAR Statistics
Data Augmentation - Test	
Normalize	CIFAR Statistics

812 I Proofs and Lemmas

813 Axial RoPE Separability

Proposition 3. *Axial RoPE is separable in x and y , that is, the attention score can be decomposed into,*

$$\alpha(\mathbf{x}_i, \mathbf{x}_j, \mathbf{p}_i, \mathbf{p}_j) = \alpha_{ij}^{(x)} + \alpha_{ij}^{(y)}$$

814 **Proof.** Suppose we define the dot-product attention score as

$$\alpha(\mathbf{q}, \mathbf{k}) = \mathbf{q}^\top \mathbf{k}.$$

815 We incorporate *Axial Rotary Positional Embeddings* by rotating each 2-dimensional subvector of the
816 query (and likewise the key). Concretely, if the hidden dimension is $2n$, we partition

$$\mathbf{q} = [\mathbf{q}_{x,1}, \mathbf{q}_{y,1}, \dots, \mathbf{q}_{x,n}, \mathbf{q}_{y,n}]^\top, \quad \mathbf{k} = [\mathbf{k}_{x,1}, \mathbf{k}_{y,1}, \dots, \mathbf{k}_{x,n}, \mathbf{k}_{y,n}]^\top,$$

817 where each $\mathbf{q}_{x,d}, \mathbf{q}_{y,d}, \mathbf{k}_{x,d}, \mathbf{k}_{y,d} \in \mathbb{R}^2$. At spatial location $\mathbf{p} = (p_x, p_y)$, we apply rotations

$$\mathbf{q}'_{x,d} = \mathbf{R}(\omega_d p_x) \mathbf{q}_{x,d}, \quad \mathbf{q}'_{y,d} = \mathbf{R}(\omega_d p_y) \mathbf{q}_{y,d},$$

818 and similarly for \mathbf{k} . Here $\mathbf{R}(\theta) \in \mathbb{R}^{2 \times 2}$ is the planar rotation by angle θ .

819 For tokens at positions $\mathbf{p}_i = (p_{i,x}, p_{i,y})$ and $\mathbf{p}_j = (p_{j,x}, p_{j,y})$, their rotated queries and keys yield

$$\alpha_{ij} = \sum_{d=1}^n \left[(\mathbf{q}_{x,d})^\top \mathbf{R}(\omega_d (p_{j,x} - p_{i,x})) \mathbf{k}_{x,d} + (\mathbf{q}_{y,d})^\top \mathbf{R}(\omega_d (p_{j,y} - p_{i,y})) \mathbf{k}_{y,d} \right].$$

820 Define the horizontal and vertical components by

$$\alpha_{ij}^{(x)} := \sum_{d=1}^n (\mathbf{q}_{x,d})^\top \mathbf{R}(\omega_d (p_{j,x} - p_{i,x})) \mathbf{k}_{x,d}, \quad \alpha_{ij}^{(y)} := \sum_{d=1}^n (\mathbf{q}_{y,d})^\top \mathbf{R}(\omega_d (p_{j,y} - p_{i,y})) \mathbf{k}_{y,d}.$$

821 Hence the total attention decomposes additively:

$$\alpha_{ij} = \alpha_{ij}^{(x)} + \alpha_{ij}^{(y)},$$

822 demonstrating that *axial* rotary embeddings factorize the positional dependence along each axis. \square

823 **Matrix Exponentiation** Computing the matrix exponential by exponentiating the eigenvalues is a
824 common result in linear algebra and numerics, however we provide it here for those unfamiliar.

Lemma 1. *Let \mathbf{A} be a diagonalizable matrix $\mathbf{A} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^{-1}$, then the matrix exponential of \mathbf{A} is given by*

$$\exp(\mathbf{A}) = \mathbf{U} \exp(\mathbf{\Lambda}) \mathbf{U}^{-1}$$

825 **Proof.**

826 Recall the power-series definition of the matrix exponential:

$$\exp(\mathbf{A}) = \sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{A}^k. \quad (55)$$

827 Since \mathbf{A} is diagonalizable,

$$\mathbf{A}^k = (\mathbf{U} \mathbf{\Lambda} \mathbf{U}^{-1})^k = \mathbf{U} \mathbf{\Lambda}^k \mathbf{U}^{-1}. \quad (56)$$

828 Substituting into the series gives

$$\exp(\mathbf{A}) = \sum_{k=0}^{\infty} \frac{1}{k!} (\mathbf{U} \mathbf{\Lambda}^k \mathbf{U}^{-1}) = \mathbf{U} \left(\sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{\Lambda}^k \right) \mathbf{U}^{-1}. \quad (57)$$

829 Because $\mathbf{\Lambda}$ is diagonal, the series $\sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{\Lambda}^k$ is itself the diagonal matrix of scalar exponentials,

$$\exp(\mathbf{\Lambda}) = \text{diag}(e^{\lambda_1}, \dots, e^{\lambda_n}). \quad (58)$$

830 Hence is well defined, and

$$\exp(\mathbf{A}) = \mathbf{U} \exp(\mathbf{\Lambda}) \mathbf{U}^{-1}. \quad (59)$$

831

□

832 **Simultaneous-Diagonalizability Proof** The proof that two (diagonalizable) matrixes are
 833 simultaneous-diagonalizability if and only if they are commutative is also a standard result. However,
 834 we provide it here:

Lemma 2. *Let \mathcal{A}_x and \mathcal{A}_y be skew-symmetric. Then \mathcal{A}_x and \mathcal{A}_y are simultaneously diagonalizable if and only if $AB = BA$.*

835 **Proof.**

836 Suppose \mathcal{A}_x and \mathcal{A}_y are simultaneously diagonalizable. Then, because they are skew-symmetric,
 837 there exists a unitary matrix \mathbf{U} such that

$$\mathbf{U} \mathbf{\Lambda}_x \mathbf{U}^\top = \mathcal{A}_x \quad \text{and} \quad \mathbf{U} \mathbf{\Lambda}_y \mathbf{U}^\top = \mathcal{A}_y, \quad (60)$$

838 where $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ are diagonal matrices.

839 Then,

$$\mathcal{A}_x \mathcal{A}_y = \mathbf{U} \mathbf{\Lambda}_x \mathbf{U}^\top \mathbf{U} \mathbf{\Lambda}_y \mathbf{U}^\top = \mathbf{U} \mathbf{\Lambda}_x \mathbf{\Lambda}_y \mathbf{U}^\top = \mathbf{U} \mathbf{\Lambda}_y \mathbf{\Lambda}_x \mathbf{U}^\top = \mathcal{A}_y \mathcal{A}_x \quad (61)$$

840 Hence, \mathcal{A}_x and \mathcal{A}_y commute.

841 Now suppose \mathcal{A}_x and \mathcal{A}_y commute, $\mathcal{A}_x \mathcal{A}_y = \mathcal{A}_y \mathcal{A}_x$. Since \mathcal{A}_x and \mathcal{A}_y are skew-symmetric, they
 842 are diagonalizable in $\mathbb{C}^{D \times D}$, thus there exists a basis of eigenvectors of \mathcal{A}_x . Because \mathcal{A}_y commutes
 843 with \mathcal{A}_x , the eigenspaces of \mathcal{A}_x are invariant under \mathcal{A}_y . That is, for any eigenvalue λ of \mathcal{A}_x , the
 844 corresponding eigenspace

$$E_\lambda = \{v \in \mathbb{C}^D : \mathcal{A}_x v = \lambda v\} \quad (62)$$

845 is \mathcal{A}_y -invariant: if $v \in E_\lambda$, then

$$\mathcal{A}_x(\mathcal{A}_y v) = \mathcal{A}_y(\mathcal{A}_x v) = \mathcal{A}_y(\lambda v) = \lambda \mathcal{A}_y v \Rightarrow \mathcal{A}_y v \in E_\lambda. \quad (63)$$

846 Now, restrict \mathcal{A}_x to each eigenspace E_λ . Since \mathbb{C} is algebraically closed and $\mathcal{A}_y|_{E_\lambda}$ is a linear
 847 operator on a finite-dimensional space, \mathcal{A}_y is diagonalizable on E_λ . Thus, we can choose a basis of
 848 eigenvectors for \mathcal{A}_y in each E_λ .

849 Putting these together, we get a basis for \mathbb{C}^N consisting of vectors that are eigenvectors for both \mathcal{A}_x
 850 and \mathcal{A}_y . Therefore, \mathcal{A}_x and \mathcal{A}_y are simultaneously diagonalizable.

851

□

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We provide proofs and theoretical evidence on benchmarks.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We emphasize that our conclusions are limited to vision and have a limitations section in the Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: While we do not formally list the assumptions, we implicitly make assumptions on positional encoding through assuming N -D LieRE.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We do our best to provide hyper-parameters for reproducing our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We intend to make the code public.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide them to the best of our ability.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We did not wish to train many large models. If reviewers deem it a necessity for our results, we will comply.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide basic information about the GPUs used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We do not believe there is any violations.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We include a section in the appendix, however, it is mostly not applicable for our paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper is more theoretical.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We cite libraries used and datasets, however they are standard libraries and benchmarks. There are no other specialized assets used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

1164 Question: Does the paper describe the usage of LLMs if it is an important, original, or
1165 non-standard component of the core methods in this research? Note that if the LLM is used
1166 only for writing, editing, or formatting purposes and does not impact the core methodology,
1167 scientific rigorousness, or originality of the research, declaration is not required.

1168 Answer: [NA]

1169 Justification:

1170 Guidelines:

- 1171 • The answer NA means that the core method development in this research does not
1172 involve LLMs as any important, original, or non-standard components.
- 1173 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1174 for what should or should not be described.