Extended Abstract Track

# Do traveling waves make good positional encodings?


## Abstract

Transformers rely on positional encoding to compensate for the inherent permutation invariance of self-attention. Traditional approaches use absolute sinusoidal embeddings or learned positional vectors, while more recent methods emphasize relative encodings to better capture translation equivariances. In this work, we propose RollPE, a novel positional encoding mechanism based on *traveling waves*, implemented by applying a circular roll operation to the query and key tensors in self-attention. This operation induces a relative shift in phase across positions, allowing the model to compute attention as a function of positional differences rather than absolute indices. We show this simple method significantly outperforms traditional absolute positional embeddings and is comparable to RoPE. We derive a continuous case of RollPE which implicitly imposes a topographic structure on the query and key space. We further derive a mathematical equivalence of RollPE to a particular configuration of RoPE. Viewing RollPE through the lens of traveling waves may allow us to simplify RoPE and relate it to processes of information flow in the brain.


## 1. Introduction

The transformer architecture has achieved state-of-the-art performance across natural language processing, vision, and multimodal tasks. A central feature enabling this success is the self-attention mechanism, which models pairwise dependencies between tokens. However, because attention is inherently permutation-invariant, positional encoding is required to inject sequence order information.

Absolute encodings – such as fixed sinusoidal embeddings (Vaswani et al., 2017) or learned vectors (Dosovitskiy et al., 2020) – assign each token a distinct position representation. While effective, these encodings lack *relative awareness*: the model must learn to infer positional differences implicitly. This motivated the widely praised Rotary Positional Embeddings (RoPE) (Su et al., 2024) which efficiently implements a relative positional encoding, or *shift-equivariance*, by rotating the query and key vectors. However, recent evidence suggests strict equivariance is not a necessity for RoPE bringing into question what makes a good positional encoding (Ostmeier et al., 2024; van de Geijn et al., 2025; Barbero et al., 2024).

We introduce *rolling positional encodings* (RollPEs), a deliberately simplistic approach where position is encoded by *rolling* query and key tensors before computing their dot product. This operation is rather trivially a relative positional encoding, but has a compelling interpretation as a traveling wave. These positional encodings induce a topographic arrangement in the query and key space as detailed by Keller and Welling (2021), and, by adding additional smoothness constraints over this topographic structure, lead to the spatial loss of TDANNs (Margalit et al., 2024), which have been shown to reproduce aspects of the hallmark behavior in the ventral stream in the visual systems of primates (Lee et al., 2020). Furthermore, within neuroscience there has been recent evidence that traveling waves play a significant role in the formation of long-term memory (Muller et al., 2018) and that humans encode visual events as multiplexed traveling waves (King and Wyart, 2021). While

we propose RollPE as a toy model, it links directly with these recent trends in theoretical neuroscience.

In our initial experiments, these simple embeddings significantly outperform classic encodings and, through Multiplexed RollPE, we show once again – as suggested by van de Geijn et al. (2025) – this behavior does not appear to be due to its relative inductive bias. These embeddings have a similar behavior to RoPE; in fact, we can mathematically derive a RollPE as a form of RoPE. We hypothesize that many of the properties and interpretations of RollPE apply transitively to RoPE. Thus, we hypothesize that RoPE's success may be derived from the implicit topographic structure and traveling waves that it and RollPE impose. Through this extended abstract, we motivate our on-going work into viewing the RoPE through the lens of traveling waves.

## 2. Roll Positional Encoding

Let a sequence of hidden states be represented as $X \in \mathbb{R}^{t \times n}$, with queries $Q = XW_Q$, keys $K = XW_K$, and values $V = XW_V$. In standard attention,

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \tag{1}$$

We define the circular roll operator $\text{Roll}_p : \mathbb{R}^n \to \mathbb{R}^n$ which maps $q_i \mapsto q_{i'}$, where $i' = (i + p) \bmod n$. We could represent this in matrix form by defining the permutation matrix $S \in \mathbb{R}^{n \times n}$ be the 1-step roll matrix,

$$S^1 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad \text{Roll}_p(\mathbf{q}) = S^p\mathbf{q}. \tag{2}$$

Table 1: Accuracy on CIFAR100

| PE | CIFAR100 |
|---|---|
| Baseline (APE) | 64.2±0.9 |
| Axial RoPE | 72.4±0.4 |
| RollPE | 72.1±0.1 |
| Multiplexed RollPE | 73.4±0.5 |

As a positional encoding, we apply this to both queries and keys when calculating the attention score as done in Su et al. (2024),

$$\alpha_{i,j} = \frac{\text{Roll}_{p_i}(\mathbf{q}_i) \cdot \text{Roll}_{p_j}(\mathbf{k}_j)}{\sqrt{d}}. \tag{3}$$

Note, the attention score between token $i$ and $j$ now depends on the relative displacement induced by the roll (proof in Appendix B). That is, RollPE is *equivariant* to position. One can extend this simple rolling positional encoding by multiplexing – i. e. representing queries as the superposition of multiple vectors which are rolled at different shift speeds. This gives Multiplexed RollPE (see Appendix C for details). While this yields better results, it requires more parameters and breaks equivariance.

From the results in Table 1, we see that RollPE outperforms classic ViT positional encodings and performs similarly to RoPE. We also observe that Multiplexed RollPE outperforms RollPE, which can suggest that strict shift-equivariance may not be necessary as also suggested in van de Geijn et al. (2025).

### 2.1. Beyond discrete positions

One obvious flaw in RollPE is the need for discrete positions. While this is often reasonable in vision and language, it limits applications to continuous data such as point clouds. The second flaw is that RollPE is inherently periodic with period $n$. While this is not a problem for low-context domains such as vision—where images are typically represented with on the order of 16 patches in each direction (Dosovitskiy et al., 2020) —this becomes very limiting for language, where desirable context length is on the order of millions. Both problems can be addressed by generalizing the shift operator $S$ using its Lie algebra.

While Roll is only defined for integer shifts, we can write

$$\mathcal{A} := \log S, \qquad S = \exp(\mathcal{A}), \tag{4}$$

where exp and log denote the matrix exponential and logarithm. Since $S$ is a permutation (and hence orthogonal) matrix, $\mathcal{A}$ belongs to the Lie algebra $\mathfrak{so}(n)$, i. e. it is skew-symmetric: $\mathcal{A}^\top = -\mathcal{A}$. This lets us define a continuous shift operator

$$\mathrm{Roll}_p(\mathbf{q}) \;=\; \exp\Big(\tfrac{p}{\lambda}\mathcal{A}\Big)\,\mathbf{q}, \tag{5}$$

which is now well-defined for all $p \in \mathbb{R}$ and allows the period to be stretched by a wavelength parameter $\lambda$. By changing the wavelength parameter, one can modulate the periodicity of RollPE.

## 3. RollPE is RoPE

Because $\mathcal{A}$ is guaranteed to be skew-symmetric, it can be decomposed into

$$\mathcal{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\dagger, \tag{6}$$

where $\mathbf{\Lambda}$ is purely imaginary, $\mathbf{U}$ is unitary, and $\dagger$ is the Hermitian transpose. Eq. 5 can then be written $\mathbf{U}\exp(\frac{p}{\lambda}\mathbf{\Lambda})\mathbf{U}^\top$. In the case of the circular roll matrix, it is well known that $\mathbf{U}$ and $\mathbf{\Lambda}$ correspond to the discrete Fourier transform (DFT) matrices, $F$ and $\mathrm{diag}\big(2\pi i\frac{pk}{\lambda n}\big)_{k=0}^{n-1}$, respectively. Thus, the continuous Roll operator is given by

$$S(p) := F^* \,\mathrm{diag}\big(e^{2\pi i\frac{pk}{\lambda n}}\big)_{k=0}^{n-1} F, \tag{7}$$

where $n$ is the dimension of the query vector and $k$ is the enumeration of its eigenvalues, and $F^*$ is the inverse Fourier matrix. Notice, $\mathrm{diag}\big(e^{2\pi i\frac{pk}{\lambda n}}\big)_{k=0}^{n-1}$ is equivalent to a RoPE matrix.

**Theorem 1** *RollPE can be represented as RoPE.*

**Proof** So long as Lie algebra given by Eq. 4 is skew-symmetric RollPE is a special case of LieRE (Ostmeier et al., 2024), which has been shown to be equivalent to a configuration of RoPE for one-dimensional positions (van de Geijn et al., 2025). □

To be exact, RollPE corresponds to RoPE if the rotation frequency per dimension is $\omega_k = \frac{2\pi k}{\lambda n}$. The traditional formulation of RoPE uses fixed frequencies given by $\omega_k = 10000^{-2k/n} = e^{-2k/n\ln 10000}$. Note, one could define $\lambda = -\frac{\pi}{\ln 10000}$ to get these expressions close to each other – however, the RoPE frequencies require an additional exponentiation. The behavior of this extra exponentiation is currently unknown, but to our understanding it is included by tradition rather than necessity.

## 4. Discussion

RollPE is an interesting case to study because other architectures have shown the Roll operation to induce equivariant capsules and reproduce topographic structuring similar to that observed in the visual systems of primates (Keller and Welling, 2021; Keller et al., 2021). One can abstract the query and key vectors of continuous RollPE as continuous signals over a continuous circle that are sampled at discrete "sensors". By imposing structure on these signals one recovers topographics methods on a simplified one-dimensional cortical surface (Lee and DiCarlo, 2019; Margalit et al., 2024).

### 4.1. Topological Regularization

In many settings, small variations in token position should not result in disproportionate changes in semantic representation. This means a token with a small shift to position should remain highly correlated with itself, i. e., if

$$\frac{\mathrm{Roll}_p(\mathbf{q}) \cdot \mathrm{Roll}_{p+\Delta p}(\mathbf{q})}{\|\mathbf{q}\|^2} \geq 1 - \epsilon, \tag{8}$$

then the representations are close in Euclidean space:

$$\|\mathbf{q} - \mathrm{Roll}_{\Delta p}(\mathbf{q})\| < \epsilon. \tag{9}$$

From a topological perspective, this constraint enforces that the signal defined over the underlying manifold varies smoothly with respect to positional shifts. In other words, we are encouraging the representation to be *Lipschitz continuous* with respect to positional perturbations.

In practice, such smoothness can be promoted through an auxiliary regularization loss that penalizes differences between neighboring latent dimensions. This formulation directly parallels Laplacian regularization in graph-based learning (Kipf, 2016), where our graph is simply a directed cyclic graph.

**Topographic methods** In parallel, Topographic Deep Artificial NN (TDANN) models have sought to impose biologically inspired constraints to deep learning models by imposing a wiring length loss (Margalit et al., 2024). As a proxy for wiring length, they impose a spatial loss over the activations which are reshaped to a "cortical sheet" – which corresponds directly to the above regularization, but with a more complicated mesh than a cyclic graph. These models have been said to reproduce the neuronal clustering patterns associated with face recognition found in the visual cortex in primates (Lee et al., 2020). While their method can be seen as imposing a structure as a 2D square lattice – as opposed to RollPEs cycle graph – one can imagine generalizing traveling waves over more complex topographies. Pang et al. (2023) and Horibe et al. (2019) have suggested that the topographic and geometric structure of the brain's connectome plays an important role in how waves propagate and the types of waveforms – suggesting a benefit to extending the topological structuring of RollPE.

**Conclusion** RoPE is strongly connected to traveling wave dynamics and we believe that it can be tied to many biological motivations. Because of its connection to RollPE, we hypothesize RoPE may perform well because it has similar properties to RollPE. That is, we hypothesize traveling wave dynamics is what makes RoPE a "good" positional encoding.

## References

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024.

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.

Reinhard Eckhorn, Roman Bauer, Wolfgang Jordan, Michael Brosch, Wolfgang Kruse, Matthias Munk, and Herbert J Reitboeck. Coherent oscillations: a mechanism of feature linking in the visual cortex? multiple electrode and correlation analyses in the cat. *Biological cybernetics*, 60(2):121–130, 1988.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

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

Kazuya Horibe, Ken-ichi Hironaka, Katsuyoshi Matsushita, and Koichi Fujimoto. Curved surface geometry-induced topological change of an excitable planar wavefront. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(9), 2019.

Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive computation*, 1:139–159, 2009.

T Anderson Keller and Max Welling. Topographic vaes learn equivariant capsules. *Advances in Neural Information Processing Systems*, 34:28585–28597, 2021.

T Anderson Keller and Max Welling. Neural wave machines: learning spatiotemporally structured representations with locally coupled oscillatory recurrent neural networks. In *International Conference on Machine Learning*, pages 16168–16189. PMLR, 2023.

T Anderson Keller, Qinghe Gao, and Max Welling. Modeling category-selective cortical regions with topographic variational autoencoders. *arXiv preprint arXiv:2110.13911*, 2021.

Jean-Rémi King and Valentin Wyart. The human brain encodes a chronicle of visual events at each instant of time through the multiplexing of traveling waves. *Journal of Neuroscience*, 41(34):7224–7233, 2021.

TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

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

Christopher J Kymn, Denis Kleyko, E Paxon Frady, Connor Bybee, Pentti Kanerva, Friedrich T Sommer, and Bruno A Olshausen. Computing with residue numbers in high-dimensional representation. *Neural Computation*, 37(1):1–37, 2024.

Hyodong Lee and James J DiCarlo. Topographic deep artificial neural networks (tdanns) predict face selectivity topography in primate inferior temporal (it) cortex. *arXiv preprint arXiv:1909.09847*, 2019.

Hyodong Lee, Eshed Margalit, Kamila M Jozwik, Michael A Cohen, Nancy Kanwisher, Daniel LK Yamins, and James J DiCarlo. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *BioRxiv*, pages 2020–07, 2020.

Sindy Löwe, Phillip Lippe, Maja Rudolph, and Max Welling. Complex-valued autoencoders for object discovery. *arXiv preprint arXiv:2204.02075*, 2022.

Sindy Löwe, Phillip Lippe, Francesco Locatello, and Max Welling. Rotating features for object discovery. *Advances in Neural Information Processing Systems*, 36:59606–59635, 2023.

Eshed Margalit, Hyodong Lee, Dawn Finzi, James J DiCarlo, Kalanit Grill-Spector, and Daniel LK Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451, 2024.

Takeru Miyato, Sindy Löwe, Andreas Geiger, and Max Welling. Artificial kuramoto oscillatory neurons. *arXiv preprint arXiv:2410.13821*, 2024.

Lyle Muller, Frédéric Chavane, John Reynolds, and Terrence J Sejnowski. Cortical travelling waves: mechanisms and computational principles. *Nature Reviews Neuroscience*, 19(5): 255–268, 2018.

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

James C Pang, Kevin M Aquino, Marianne Oldehinkel, Peter A Robinson, Ben D Fulcher, Michael Breakspear, and Alex Fornito. Geometric constraints on human brain function. *Nature*, 618(7965):566–574, 2023.

David P Reichert and Thomas Serre. Neuronal synchrony in complex-valued deep networks. *arXiv preprint arXiv:1312.6115*, 2013.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216, 1990.

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

Chase van de Geijn, Timo Lüddecke, Polina Turishcheva, and Alexander Ecker. A circular argument: Does rope *need* to be equivariant for vision?, 2025. Manuscript.

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

## Appendix A. Related Work

**Neuro-Symbolic AI** Using the Roll operation within deep learning to encode position has been proposed in Keller and Welling (2021) to encode time for sequential data, such as rotating digits or color changes. They modeled a generation process that was motivated by predictive coding where the next frame of the sequence should be attained by rolling the latent vector. They found that these models, like the TDANNs (Lee et al., 2020), learn spatial organization and selectivity toward categories such as faces, similar to the Fusiform Face Area (Keller et al., 2021). This naturally led to more recent on Neural Wave Machines (Keller and Welling, 2023) and Kuramoto models Miyato et al. (2024).

The idea of Kuramoto models is to phase align with waves, so that "neurons that wire together fire together". In many ways, the classic attention mechanism in transformers also encourages phase alignment between related tokens, meaning there is likely a direct relationship of RollPE to AKoRN (Miyato et al., 2024). Similarly, the line of papers (Löwe et al., 2022, 2023) that led to AKoRN began with complex autoencoders, which are strikingly similar to RoPE. These "synchrony models" (Reichert and Serre, 2013) try to answer the "perceptual binding problem" of neuroscience of integrating features such as shape, color, and location into one object representation (Greff et al., 2020). One answer for this problem is through hyperdimensional computing (Kanerva, 2009) which has trended toward the direction of *phasor* representations (Smolensky, 1990; Kymn et al., 2024) which, once again, are strikingly similar to the form or rotary encodings. Tracing these ideas back further, one arrives, once again, at traveling wave behavior in the brain (Eckhorn et al., 1988).

## Appendix B. RollPE is Relative

One can write $\text{Roll}_p(\mathbf{q})$ in matrix form as

$$\mathbf{q}' = \text{Roll}_p(\mathbf{q}) = S^p\mathbf{q}, \tag{10}$$

now let the attention score be given by $\alpha(\mathbf{q}', \mathbf{k}') = \mathbf{q}'^\top\mathbf{k}'$ where $\mathbf{q}'$ and $\mathbf{k}'$ are the positionally encoded query and key vectors. The attention score can be expanded as,

$$\alpha(\mathbf{q}', \mathbf{k}') = (S^{p_q}\mathbf{q})^\top S^{p_k}\mathbf{k}$$

where $p_q$ and $p_k$ are the positions of the query and key token. Because $S$ is a permutation – and thus orthonormal – we can write,

$$\alpha(\mathbf{q}', \mathbf{k}') = \mathbf{q}^\top S^{p_k - p_q}\mathbf{k}.$$

Thus, the positional encoding only depends on the (signed) relative distance $p_k - p_q$.

## Appendix C. Multiplexed RollPE

Multiplexing refers to the combination of multiple signals into a single composite signal—in this case, the superposition of several traveling waves. It is common in signal processing for communication devices, as well in neuroscience, e. g . in spiking neural network literature.

In *Multiplexed RollPE*, the query/key representation is defined as a superposition of multiple components, each rolling at a different speed.

Concretely, instead of a single query/key matrix, we introduce $\mathcal{W}$ distinct projections indexed by $w \in \{1, \dots, \mathcal{W}\}$:

$$\mathbf{q}^{(w)} := W_Q^{(w)} X. \tag{11}$$

The multiplexed positional encoding is then given by

$$\mathrm{MPRoll}_p(\mathbf{q}) := \sum_{w=1}^{\mathcal{W}} \mathrm{Roll}_{(wp)}\big(\mathbf{q}^{(w)}\big). \tag{12}$$

## Appendix D. Experimental Setup

We evaluate the positional encodings on CIFAR100 (Krizhevsky et al., 2009) using ViT-S (Dosovitskiy et al., 2020). The two positions for RollPE and Multiplexed RollPE are encoded axially where each coordinate affects a different sub-vector of the query/key analogously to what is done for Axial RoPE in (Heo et al., 2024).

## Appendix E. Exact Form of $\log S$

The exact logarithm of the cyclic shift matrix $S \in \mathbb{R}^{n \times n}$ is most naturally expressed in the Fourier basis. Let $F$ be the discrete Fourier transform matrix. Then

$$S^1 = F^* \operatorname{diag}\big(e^{2\pi i k/n}\big)_{k=0}^{n-1} F, \tag{13}$$

so its logarithm is

$$\mathcal{A} = \log S \;=\; F^* \operatorname{diag}\Big(\tfrac{2\pi i k}{n}\Big)_{k=0}^{n-1} F. \tag{14}$$

This operator is circulant and skew-Hermitian, not tridiagonal. Its action corresponds exactly to multiplication by $2\pi i k/n$ in the frequency domain, i.e. a discrete Fourier differentiation operator.

The tridiagonal matrix shown in the main text should therefore be viewed as a *local finite-difference approximation* to $\mathcal{A}$, which captures the intuition of a derivative with periodic boundary conditions, but is not the literal matrix logarithm. In the continuum limit, both agree with the true derivative, but at finite $n$, the Fourier-based logarithm preserves the entire spectrum exactly, while the tridiagonal stencil sacrifices spectral accuracy for sparsity and locality.

## Appendix F. RollPE is RoPE

In van de Geijn et al. (2025), they show that arbitrary dimensional special orthogonal transformations applied to queries and keys, $\mathbf{q}' = \exp(\mathcal{A}p)\mathbf{q}$, can be decomposed into RoPE

by taking the spectral decomposition of $\mathcal{A}$, where the rotation frequencies correspond to the eigenvalues. Since $\mathcal{A}$ can be seen as the generator $\mathcal{A}$, RollPE can be decomposed into RoPE. To be exact, RollPE is a particular case of RoPE where the rotation frequencies correspond to the frequencies of the discrete fourier differential operator given in Eq. 7