FOURIER POSITIONAL EMBEDDINGS FOR VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Convolutional neural networks (CNNs) have long shaped progress in computer vision by incorporating strong inductive biases such as translation equivariance and locality. Vision Transformers (ViTs) now reach comparable performance across many tasks, but rely on weaker built-in spatial priors and instead depend on positional encodings to represent image structure. We introduce Fourier Positional Embedding for Vision Transformers (FViT), a positional encoding inspired by Fourier analysis. Rather than adding positional embeddings to token representations, FViT represents relative spatial structure through learned complex-valued phase factors that modulate attention, preserving the flexibility of self-attention while better aligning positional encoding with image geometry. Empirically, FViT substantially outperforms absolute, relative, and sinusoidal positional embeddings and matches Rotary Positional Embeddings (RoPE) on CIFAR-100 and ImageNet. On COCO object detection, FViT substantially improves over the standard YOLOS baseline under a matched training setup and achieves higher average recall than RoPE across all reported recall metrics, while RoPE retains an advantage in average precision. Finally, we argue that FViT and RoPE perform strongly because their phase-based positional operations reflect how translations act through Fourier characters.

1 INTRODUCTION

Inductive biases have been central to the success of convolutional neural networks (CNNs) in vision. By hard-coding locality and translation equivariance, CNNs learn features that generalize across spatial locations with relatively modest data, and they do so with a clear signal-processing interpretation: convolution aggregates local evidence while respecting the geometry of the image grid.

Vision Transformers (ViTs) replace convolution with global self-attention (1; 3), gaining flexible long-range interactions and strong scaling behavior. The price is that attention is inherently permutation-invariant: if token order is shuffled, the raw attention operator is unchanged. As a result, ViTs require an explicit mechanism to encode spatial structure. The most common solution is to add absolute positional embeddings to token representations (6; 3). While effective, additive position treats spatial information as auxiliary metadata, leaving the model to learn how geometry should influence pairwise interactions. This is fundamentally different from convolutional structure, where translation acts directly on the computation.

A large body of work has explored more structured alternatives, including relative position representations (2; 19) and rotary-style encodings (5; 20). These methods move beyond purely absolute indexing, but a common pattern remains: position is typically added to embeddings or applied as a real-valued transformation on queries/keys, rather than appearing as a multiplicative modulation that *acts inside* the attention interaction itself. From a geometric and signal-processing perspective, this is a missed opportunity. Images are naturally modeled on a 2D grid, and translations on this grid have a canonical representation in the frequency domain.

The Fourier shift theorem provides the algebraic motivation for our construction. If $f_\Delta(u) = f(u - \Delta)$, then

$$\mathcal{F}[f_\Delta](\omega) = e^{-i\omega^\top \Delta}\mathcal{F}[f](\omega).$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

For a fixed frequency ω, displacement therefore acts through a unit-modulus phase factor. FViT borrows this phase-rotation form while allowing its frequency to depend on the query representation. We do not assume that the value features are Fourier coefficients or that the complete attention operator is a fixed unitary representation of translations.

Motivated by this principle, we introduce *Fourier Positional Embedding for Vision Transformers (FViT)*, a positional encoding mechanism that injects spatial information directly into attention through multiplicative, frequency-based interactions. Instead of adding a position vector to each token, FViT attaches a complex-valued phase to each query–key interaction so that relative spatial displacement appears as a structured phase rotation. Concretely, standard attention forms real-valued similarity scores via inner products of projected queries and keys. FViT augments this computation by introducing a complex factor $\exp(i\varphi)$ whose phase φ depends on the relative displacement between tokens on the 2D patch grid. The absolute value of the interaction retains standard attention-based similarity, while the complex exponential encodes geometry through Fourier-style phase shifts.

A key aspect of FViT is that the phase is *head-specific* and *learned*, enabling different heads to specialize to different spatial-frequency responses. In our formulation, each head learns axis-specific projections that produce frequency coefficients from token representations; these coefficients then modulate attention according to the relative offset between the interacting tokens. Because the phase factor has unit magnitude, it alters the *structure* of interactions without changing their scale, and we keep the softmax normalization identical to standard attention by normalizing only with real-valued similarity scores. This design makes FViT a drop-in replacement for positional encodings in ViTs, integrating naturally with common attention patterns, including local windowed attention for scalability in high-resolution settings.

We evaluate FViT across both image classification and object detection. For image classification, we use a ViT backbone, with experiments on CIFAR-100 and ImageNet following the standard class-token protocol (3). On CIFAR-100, we perform a controlled comparison across learned absolute embeddings (3), sinusoidal encodings (1), decomposed relative embeddings (19), axial RoPE (20), and FViT. On ImageNet, we focus on a direct matched comparison between FViT and RoPE. Within each comparison, we keep the backbone architecture, training schedule, and optimization settings fixed and vary only the positional encoding module.

For object detection, we evaluate on the COCO 2017 benchmark (17) using the YOLOS architecture (18). We compare the standard YOLOS baseline with matched FViT and RoPE variants under the same training configuration.

In summary, this paper makes four contributions. First, we propose FViT, a Fourier-motivated positional encoding that represents relative spatial displacement as a complex phase acting multiplicatively inside attention. Second, we provide a practical attention formulation in which content similarity and positional phase cleanly separate: the amplitude is determined by standard real-valued dot products, while geometry is expressed as a unit-modulus rotation, leaving the softmax normalization unchanged. Third, we empirically evaluate FViT across both image classification and object detection under controlled conditions. The classification experiments establish parity with RoPE at scale, while the COCO experiments show that the benefits of structured multiplicative positional encoding transfer beyond classification: both FViT and RoPE substantially outperform the YOLOS baseline, with FViT favoring object recall and RoPE favoring average precision. Finally, we argue that FViT and RoPE perform strongly because their phase-based positional operations reflect how translations act through Fourier characters.

2 Related Work

Positional information in transformers Self-attention is permutation-invariant, so transformers require an explicit mechanism to represent order or spatial layout. The original Transformer introduced fixed sinusoidal embeddings and also discussed learned absolute embeddings as a drop-in alternative (1). In large language models, learned absolute embeddings became common (e.g., BERT (6)), but they hard-code a maximum sequence length and tend to extrapolate poorly when the input length or resolution changes. These issues carry over to vision: ViTs tokenize images

into patch sequences, and the same model may be evaluated at different resolutions or on dense prediction tasks where the effective token grid varies.

Absolute vs. relative position encodings A broad line of work augments attention by *adding* position-dependent vectors to token embeddings (absolute encodings), or by injecting relative position information directly into the attention computation (relative encodings). RoFormer provides a useful synthesis of this landscape: it contrasts additive absolute encodings (learned or sinusoidal) with a family of relative encodings that parameterize interactions as a function of token displacement (5). Classic relative position representations add learned embeddings that depend on the clipped offset between two tokens (2). Subsequent variants refine where and how relative position enters attention, including decompositions used in Transformer-XL (7) and XLNet (8), bias-only formulations as in T5 (9), and disentangled attention mechanisms such as DeBERTa (10). Other work analyzes and improves these designs (e.g., rethinking positional encoding choices during pretraining (11), and improved relative embedding parameterizations (12)).

In vision, relative encodings are typically adapted to 2D grids by factorizing horizontal and vertical offsets. Decomposed relative position embeddings for ViTs (19) exemplify this approach and have become a standard baseline when comparing positional encoding strategies.

Multiplicative encodings and rotations We use the term *multiplicative positional encoding* to describe methods that encode position through position-dependent transformations that multiply token embeddings.

RoFormer (RoPE) rotates keys and queries according to token position, mathematically equivalent to introducing a relative complex phase into their dot product, so that the resulting attention scores depend explicitly on relative displacement without changing the norms of the rotated key and query representations (5). RoFormer highlights several practical advantages of this construction, including improved length extrapolation, an implicit decay of interactions with distance, and compatibility with linear-attention variants, as the rotation can be applied without disrupting attention factorization.

In vision, RoPE must be extended from a one-dimensional sequence to two-dimensional patch coordinates. Axial RoPE represents the two spatial axes through separate axial frequency components, providing a direct 2D adaptation of the rotary construction. Our experiments use this axial formulation as the primary rotary baseline.

FViT belongs to the multiplicative family but applies its positional transformation at a different stage. RoPE rotates queries and keys using fixed frequencies, directly changing attention logits. FViT rotates the values using learned frequencies by introducing a query-dependent phase term into attention. Its defining feature is therefore encoding position via query-dependent phases, rather than Fourier features or learnable frequencies alone (24; 25).

Spectral alternatives and efficient attention Spectral ideas have also been used more aggressively by replacing attention with Fourier mixing, as in FNet (15). Our goal differs: we retain adaptive content-based attention and introduce Fourier structure only through positional modulation.

In our experiments, we adopt windowed attention patterns for high-resolution vision (16).

3 Proposed Approach

We introduce Fourier Positional Embedding for Vision Transformers (FViT), a positional encoding mechanism that incorporates spatial information directly into the attention computation through multiplicative, frequency-based interactions. FViT is motivated by the observation that spatial translations are most naturally represented as phase shifts in the frequency domain, corresponding to unitary representations of the translation group. By embedding position as a phase, FViT aligns positional encoding with the harmonic structure underlying image data.

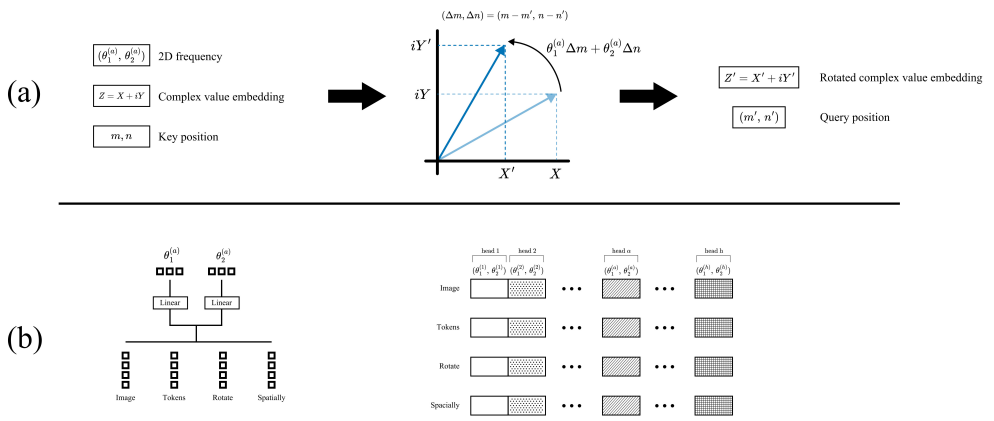

Figure 1: Conceptual illustration of Fourier positional embedding (FViT).
(a) FViT encodes relative displacement $(\Delta m, \Delta n)$ as a complex-valued exponential phase (equivalently, a planar rotation) with angle $(\theta_1^{(a)}\Delta m + \theta_2^{(a)}\Delta n)$. This transforms a complex-valued token embedding $Z = X + iY$ into $Z' = X' + iY'$.
(b) The head-specific frequencies $(\theta_1^{(a)}, \theta_2^{(a)})$ are produced by learned linear projections, so each head learns a distinct spatial frequency response.

3.1 Conceptual Overview

Standard self-attention computes pairwise interactions between query and key embeddings using real-valued inner products. Positional information is typically injected by adding learned or fixed embeddings to token representations prior to attention, treating position as auxiliary metadata.

In contrast, FViT incorporates spatial position directly into the attention operation through a complex-valued, multiplicative phase. The standard real-valued attention weights determine the magnitude of each value contribution, while a query-dependent phase, computed from the relative spatial displacement between tokens using learned frequencies, rotates that contribution during aggregation. Position therefore acts on the attended representations rather than being added to the token embeddings beforehand. This construction is inspired by Fourier analysis, where translations correspond to phase shifts.

3.2 Frequency-Based Attention Modulation

To apply a complex phase shift to real-valued token embeddings, we identify each vector $x \in \mathbb{R}^d$ with a complex counterpart $x^C \in \mathbb{C}^{d/2}$ via

$$x_r^C = x_{2r-1} + ix_{2r}, \qquad r = 1, \ldots, d/2. \tag{1}$$

In particular, a complex phase shift is identified with the 2D real rotation

$$\exp(i\varphi)x_r^C = \big(\cos\varphi\, x_{2r-1} - \sin\varphi\, x_{2r}\big) + i\big(\sin\varphi\, x_{2r-1} + \cos\varphi\, x_{2r}\big), \qquad r = 1, \ldots, d/2. \tag{2}$$

FViT operates on a 2D grid of image patches. Let $i = (m, n)$ and $j = (m', n')$ denote query and key positions, respectively, and let $v_j^{(a)} = V^{(a)}x_j \in \mathbb{R}^{d_h}$. For each attention head $a \in \{1, \ldots, h\}$, two learned axis-specific projections $U_1^{(a)}, U_2^{(a)} \in \mathbb{R}^d$ map the query embedding $x_i \in \mathbb{R}^d$ to real-valued frequency coefficients

$$\theta_{1,i}^{(a)} = \frac{\big(U_1^{(a)}\big)^T x_i}{\sqrt{d_h}}, \qquad \theta_{2,i}^{(a)} = \frac{\big(U_2^{(a)}\big)^T x_i}{\sqrt{d_h}}, \tag{3}$$

where $d_h = d/h$ is the per-head embedding dimension. These coefficients are computed independently for each head, so the resulting phases are added head-wise in the attention computation.

Let $\Delta m = m' - m$ and $\Delta n = n' - n$ denote the relative displacement from i to j. FViT converts the frequency coefficients into a rotation in the complex plane

$$Z \mapsto Z' = \exp(i\varphi^{(a)})Z, \tag{4}$$

$$\varphi_{ij}^{(a)} \;=\; \frac{2\pi}{N}\Big(\Delta m\,\theta_{1,i}^{(a)} + \Delta n\,\theta_{2,i}^{(a)}\Big), \tag{5}$$

where N is the number of tokens along the corresponding spatial axis (e.g., width or height). This phase modulates an otherwise standard real-valued similarity score. Concretely, let $K^{(a)}, Q^{(a)} \in \mathbb{R}^{d_h \times d}$ be the key and query matrices respectively, and let

$$s_{ij}^{(a)} \;=\; \frac{\big(Q^{(a)}x_i\big)^T\big(K^{(a)}x_j\big)}{\sqrt{d_h}}. \tag{6}$$

We assume that the per-head dimension d_h is even. The real-valued attention probabilities are

$$P_{ij}^{(a)} = \frac{\exp(s_{ij}^{(a)})}{\sum_{t\in\mathcal{N}(i)}\exp(s_{it}^{(a)})}. \tag{7}$$

Here $\mathcal{N}(i)$ is the set of keys permitted by the attention mask, and $s_{ij}^{(a)}$ is the similarity score defined above.

We implement the complex phase using real rotations on adjacent pairs of projected value channels. For $r = 1, \ldots, d_h/2$, the output is

$$\begin{bmatrix} y_{i,2r-1}^{(a)} \\ y_{i,2r}^{(a)} \end{bmatrix} = \sum_{j\in\mathcal{N}(i)} P_{ij}^{(a)} \begin{bmatrix} \cos\varphi_{ij}^{(a)} & -\sin\varphi_{ij}^{(a)} \\ \sin\varphi_{ij}^{(a)} & \cos\varphi_{ij}^{(a)} \end{bmatrix} \begin{bmatrix} v_{j,2r-1}^{(a)} \\ v_{j,2r}^{(a)} \end{bmatrix}. \tag{8}$$

Equivalently, define the complex-coordinate map $\mathcal{C}(v)_r = v_{2r-1} + iv_{2r}$. Then $y_i^{(a)} = \mathcal{C}^{-1}\left(\sum_{j\in\mathcal{N}(i)} P_{ij}^{(a)} e^{i\varphi_{ij}^{(a)}}\mathcal{C}(v_j^{(a)})\right)$, where $\mathcal{C}^{-1}$ interleaves the real and imaginary components. Both components are retained, so $y_i^{(a)} \in \mathbb{R}^{d_h}$ has the original head dimension. The head outputs are concatenated and projected as $o_i = W\,\mathrm{Concat}_{a=1}^{h}(y_i^{(a)}) \in \mathbb{R}^d$, with $W \in \mathbb{R}^{d\times d}$.

Each pairwise rotation preserves the norm of the projected value it acts on. The aggregate norm can nevertheless change because rotated contributions may reinforce or cancel. The phase leaves the scalar softmax probabilities unchanged.

FViT is motivated by the phase form of Fourier translations. Its frequencies are predicted from query representations, producing a content-conditioned family of relative phase rotations rather than a single input-independent transformation of all token features.

3.3 Base Architecture and Integration

All experiments are built on a Vision Transformer (ViT) backbone, which provides a baseline for comparing positional encoding strategies. Images are partitioned into fixed-size patches and embedded into token sequences processed by transformer layers augmented with FViT.

To ensure scalability to high-resolution inputs and dense prediction tasks, we adopt the windowed self-attention pattern introduced in SAM (16), which interleaves local windowed attention with global attention layers. This design preserves computational efficiency while allowing the model to integrate global contextual information.

For image classification tasks, including CIFAR-100 and ImageNet, we append a learned class token to the token sequence and use its final representation for prediction. For object detection, we use the YOLOS architecture, which augments the Vision Transformer token sequence with detection tokens and an object-detection head. The detector architecture is kept fixed across the YOLOS, RoPE, and FViT variants so that the positional encoding mechanism remains the primary experimental variable.

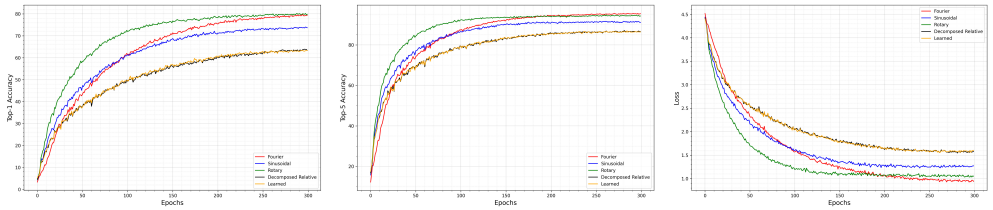

Figure 2: Top-1 accuracy (left), Top-5 accuracy (center), and test loss (right) on CIFAR-100 over training epochs for different positional encoding methods.

4 Experiments

We evaluate FViT across both image classification and object detection. Within each task, we isolate the effect of positional encoding by holding the architecture, training schedule, and optimization settings fixed across the compared variants.

We conduct classification experiments on CIFAR-100 and ImageNet using a DeiT-style Vision Transformer. Positional encodings are the only varying component. We compare learned absolute embeddings (3), sinusoidal embeddings (1), decomposed relative embeddings (19), axial Rotary Positional Embeddings (RoPE) (20), and FViT.

For object detection, we evaluate on the COCO 2017 benchmark (17) using the YOLOS architecture (18). We compare the standard YOLOS baseline with matched FViT and RoPE variants under the same training configuration.

4.1 CIFAR-100

All CIFAR-100 experiments were run on four NVIDIA V100 GPUs (32GB) for 300 epochs with a batch size of 64. Models share identical architectures and training schedules.

Results. Table 1 reports CIFAR-100 classification performance. FViT outperforms additive positional embeddings and matches RoPE in Top-1 accuracy, while achieving the highest Top-5 accuracy and lowest test loss. FViT also exhibits more stable late-stage training dynamics, overtaking RoPE in Top-5 accuracy after approximately 190 epochs.

Table 1: CIFAR-100 Performance

Embedding	**Top-1 (%)**	**Top-5 (%)**	**Test Loss**	**# Params**
Learned	63.33	86.39	1.602	55.5M
Sinusoidal	73.70	91.14	1.271	55.3M
Decomposed Relative	63.31	86.55	1.575	55.3M
RoPE	79.86	94.28	1.052	55.3M
FViT	79.41	95.39	0.941	55.3M

Fig. 2 illustrates CIFAR-100 training dynamics over 300 epochs. Fourier positional embedding matches Rotary in Top-1 accuracy, overtakes it in Top-5 accuracy after approximately 190 epochs, and achieves the lowest test loss.

4.2 ImageNet

For ImageNet, training was scaled to 16 NVIDIA A100 GPUs for 300 epochs with a batch size of 32. Based on CIFAR-100 results, we focus on a direct comparison between FViT and RoPE.

Results. Table 2 reports ImageNet classification results. FViT and RoPE exhibit nearly identical performance across all metrics, confirming that Fourier-based positional encoding scales effectively to large datasets and high-resolution inputs while maintaining parity with state-of-the-art positional embeddings. Fig. 3 further illustrates this behavior through their closely matched training dynamics.

Table 2: ImageNet Performance

Embedding	Top-1 (%)	Top-5 (%)	Test Loss	# Params
RoPE	81.01	95.18	0.888	56.0M
FViT	80.81	95.19	0.902	56.0M

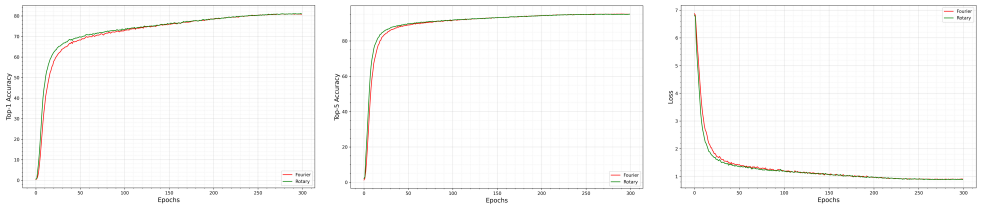

Figure 3: Top-1 accuracy (left), Top-5 accuracy (center), and test loss (right) on ImageNet over training epochs for FViT and RoPE.

4.3 COCO Object Detection

We next evaluate whether the benefits of Fourier positional encoding transfer beyond image classification using the COCO 2017 object detection benchmark. The YOLOS baseline and its FViT and RoPE variants are evaluated under the same training configuration, with the backbone, optimization schedule, batch configuration, evaluation resolution, and training horizon held fixed across models. The models are trained from scratch, allowing the comparison to isolate the effect of the positional encoding mechanism within the matched detection setup. Table 3 reports COCO bounding-box average precision and average recall.

Both structured positional encodings substantially improve object detection relative to the YOLOS baseline. FViT increases AP from 10.0 to 15.9 and AR_{100} from 27.8 to 36.8. RoPE achieves the strongest precision-oriented performance, reaching 17.3 AP and leading all six reported AP metrics. In contrast, FViT leads all six reported average-recall metrics, including AR_{100} as well as recall for small, medium, and large objects. This reveals a distinction that is not apparent from classification accuracy alone: RoPE retains an advantage in average precision, whereas FViT consistently retrieves a larger fraction of objects under the matched detection setup.

Table 3: COCO 2017 object detection performance under the matched training setup. Best results in each row are shown in bold.

Metric	YOLOS	FViT	RoPE
AP	10.0	15.9	**17.3**
AP_{50}	20.1	27.2	**31.3**
AP_{75}	8.8	15.6	**16.7**
AP_S	2.7	5.2	**6.3**
AP_M	8.6	15.4	**17.3**
AP_L	17.5	26.7	**27.9**
AR_1	14.6	**20.3**	19.9
AR_{10}	25.2	**33.4**	32.5
AR_{100}	27.8	**36.8**	35.4
AR_S	7.0	**11.9**	11.8
AR_M	25.4	**37.4**	35.9
AR_L	50.6	**61.7**	59.4

5 Discussion

On CIFAR-100, where a broader comparison of positional encoding schemes was computationally feasible, we observe the following ordering of positional embedding performance: FViT $\approx$ RoPE $>$ Absolute Sinusoidal $>$ Absolute Learned $\approx$ Relative This ordering is stable across the reported metrics and motivates examining which structural properties of positional encodings are associated with stronger vision transformer performance.

Due to computational constraints, the full ordering of positional embeddings was established only on CIFAR-100. On ImageNet, we focus on the FViT–RoPE comparison and use published ImageNet baselines to benchmark other encodings.

Controlled ImageNet ablations reported in Chu et al. (21) show that absolute baselines remain strong for classification: under a DeiT-Tiny (224×224) comparison, learned absolute embeddings reach 72.2% top-1 accuracy and fixed 2D sinusoidal embeddings reach 72.3%, while a standard 2D relative positional encoding obtains 70.5. In a separate same-model study, van de Geijn et al. (22) reports that RoPE variants consistently outperform learned absolute embeddings on ImageNet: learned absolute positional embeddings reach 68.8% top-1 accuracy, while Uniform RoPE, Axial RoPE, and Spherical RoPE achieve 70.0%, 70.7%, and 70.9%, respectively. In our ImageNet experiments (Table 2), FViT matches RoPE at scale. Taken together, these results support the following ImageNet ordering as a useful empirical summary:

$$\text{FViT} \approx \text{RoPE} > \text{Absolute Sinusoidal} \approx \text{Absolute Learned} > \text{Relative}.$$

The COCO results extend this comparison beyond image classification. Both FViT and RoPE substantially outperform the YOLOS baseline, indicating that the advantages of structured multiplicative positional encoding transfer to object detection. At the same time, the two methods exhibit distinct behavior: RoPE leads every reported average-precision metric, whereas FViT leads every reported average-recall metric. Thus, positional encodings that appear nearly interchangeable under classification accuracy can produce meaningfully different detection behavior.

A notable result is that relative positional embeddings do not outperform absolute embeddings, despite explicitly encoding translation equivariance. This is unexpected from the perspective of geometric deep learning, where equivariance is considered a primary inductive bias for vision (23). We also observe a dataset-dependent shift among absolute encodings: on CIFAR-100, fixed sinusoidal absolute embeddings outperform learned absolute embeddings, while on ImageNet learned absolute embeddings close the gap and match the sinusoidal baseline. A plausible explanation is that ImageNet scale provides enough data for learned absolute embeddings to internalize regular spatial regularities that resemble the structured patterns supplied by sinusoidal encodings. Overall, these results suggest that equivariance in the positional encoding alone is not sufficient to yield gains in the transformer variants studied here, and that the benefit of structured absolute priors diminishes as learned absolute embeddings are trained at scale.

At the top of the hierarchy are RoPE and FViT. Their shared advantage is not simply that they encode relative position, but that they introduce positional information multiplicatively. This distinction has a natural Fourier interpretation. For a fixed frequency vector ω, the map

$$\chi_\omega(\Delta) = e^{i\omega^\top \Delta}$$

is a character of the translation group: translations compose by multiplying their phase factors. RoPE applies fixed characters of this form to keys and queries, making the attention logits depend multiplicatively on relative displacement.

FViT applies the same character structure to the values, but its frequencies are conditioned on the queries. For a fixed query representation x, define

$$\omega(x) = \frac{2\pi}{N}\big(\theta_1(x), \theta_2(x)\big)$$

under the normalization of Equation (5). The corresponding phase map

$$\chi_x(\Delta) = e^{i\omega(x)^\top \Delta}$$

satisfies

$$\chi_x(\Delta_1 + \Delta_2) = \chi_x(\Delta_1)\chi_x(\Delta_2).$$

Thus, each fixed query defines a unitary translation character. Changing the query representation changes the frequency and therefore the character; consequently, this operation is not a single fixed representation of the translation group.

This perspective suggests that the multiplicative introduction of positional information is an important common feature of RoPE and FViT. Both methods make relative displacement act directly through a structured phase operation: RoPE modulates query–key interactions using fixed characters, whereas FViT modulates value contributions using query-conditioned characters. Additive positional encodings do not impose this compositional structure directly. Although our experiments do not establish multiplicative modulation as the sole cause of the observed gains, their results are consistent with the hypothesis that this structure contributes to the strong performance of RoPE and FViT.

AI Use Statement

Generative AI tools were used to assist with drafting and editing the manuscript. All original FViT code was written by the authors. AI tools were also used to facilitate the integration of FViT into the YOLOS architecture for the COCO object-detection experiments. The authors reviewed and validated all AI-assisted text and code and take full responsibility for the content and results presented in this paper.

Reproducibility Statement

We support reproducibility by describing the FViT formulation and its integration into the base architectures in Section 3.2 and the surrounding methodology sections. The experimental section reports the datasets, model variants, training durations, batch sizes, hardware, evaluation metrics, and controlled comparison settings used for the CIFAR-100, ImageNet, and COCO experiments. Across each task, the architecture and training configuration are held fixed between positional-encoding variants wherever applicable, so that the reported comparisons isolate the effect of the positional encoding.

References

[1] Vaswani et al. Attention Is All You Need. *NeurIPS*, 2017.

[2] Shaw et al. Self-Attention with Relative Position Representations. *NAACL*, 2018.

[3] Dosovitskiy et al. An Image Is Worth 16x16 Words. *ICLR*, 2021.

[4] Li et al. MViTv2: Improved Multiscale Vision Transformers. *CVPR*, 2022.

[5] Su et al. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 2024.

[6] Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*, 2019.

[7] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *ACL*, 2019.

[8] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *NeurIPS*, 2019.

[9] 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. *JMLR*, 2020.

[10] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. *ICLR*, 2021.

[11] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking Positional Encoding in Language Pre-training. *arXiv:2006.15595*, 2020.

[12] Zhiheng Huang, Davis Liang, Peng Xu, and Bing Xiang. Improve Transformer Models with Better Relative Position Embeddings. *Findings of EMNLP*, 2020.

[13] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. *ICML*, 2020.

[14] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking Attention with Performers. *ICLR*, 2021.

[15] Lee-Thorp et al. FNet: Mixing Tokens with Fourier Transforms. *NAACL*, 2021.

[16] Kirillov et al. Segment Anything. *ICCV*, 2023.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. *ECCV*, 2014.

[18] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You Only Look at One Sequence: Rethinking Transformer-based Object Detection. *NeurIPS*, 2021.

[19] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and Improving Relative Position Encoding for Vision Transformer. *ICCV*, 2021.

[20] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary Position Embedding for Vision Transformer. *ECCV*, 2024.

[21] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional Positional Encodings for Vision Transformers. *ICLR*, 2023.

[22] Chase van de Geijn, Timo Lüddecke, Polina Turishcheva, and Alexander S. Ecker. A Circular Argument: Does RoPE need to be Equivariant for Vision? *arXiv:2511.08368*, 2025.

[23] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478*, 2021.

[24] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *NeurIPS*, 2020.

[25] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable Fourier Features for Multi-Dimensional Spatial Positional Encoding. *NeurIPS*, 2021.