A Theoretical and Experimental Analysis of the Robustness of Similarity-based Positional Encoding Under Rotation

Anonymous authors
Paper under double-blind review

Abstract

Transformer architectures heavily rely on positional encodings (PEs) to inject spatial structure; however, standard approaches remain highly sensitive to spatial rotations. In this paper, we present a unified theoretical and empirical investigation into the rotational robustness of similarity-based positional encoding (SimPE). Theoretically, we prove that while SimPE is not strictly rotation-invariant, it exhibits provable stability under simple Lipschitz regularity assumptions, yielding explicit Frobenius norm perturbation bounds that cleanly decouple architectural features from rotational magnitude. Empirically, we evaluate models trained on canonical orientations against increasing rotation angles across five diverse datasets, by comparing SimPE with a comprehensive suite of baselines PEs. Across the tested datasets, SimPE provides better performance (measured in terms of classification accuracy and F1-score) in small-to-moderate rotation regimes ($\leq 15°$), with respect to competing PE methods. Crucially, on a real clinical benchmark (`Chest X-Ray`), SimPE outperforms all competing PEs across every evaluated angle (5°–30°), and allows the corresponding ViT architecture to reach significant levels of accuracy. Our findings bridge theoretical stability guarantees with practical spatial robustness, establishing SimPE as a highly effective encoding strategy for domains subject to routine geometric perturbations.

1 Introduction

Transformers have become a dominant architecture in both natural language processing and computer vision (Vaswani et al., 2017; Dosovitskiy et al., 2021). Since self-attention is permutation-invariant, Transformer models require explicit positional information to preserve the structure of the input domain. In vision applications, positional encoding (PE) is especially important because it allows models to represent the spatial arrangement of image patches or feature tokens (Parmar et al., 2018; Liu et al., 2021).

Classical PE schemes, such as sinusoidal encodings (Vaswani et al., 2017) or learned absolute embeddings (Gehring et al., 2017; Dosovitskiy et al., 2021), inject coordinate information directly. Similarity-based positional encoding (simPE) (Santomauro et al., 2026) belongs to a distinct family in which positional information is encoded through pairwise similarity operators acting on token representations, without relying on fixed coordinate systems. The study of SimPE is particularly motivated by medical imaging, which served as the original application domain for this encoding strategy. In clinical workflows, spatial rotations are far from rare, frequently arising from imaging instrument setups or minor variations in patient positioning (Chen et al., 2021). Crucially, angular variations up to 15° fall within a safe operating regime, as they do not significantly alter key anatomical landmarks or compromise diagnostic validity (Jud et al., 2023; Tannast et al., 2014; Siebenrock et al., 2003; Fisk & et al., 2021). As a result, maintaining model robustness against these routine geometric perturbations is not merely a theoretical exercise, but a practical necessity for real-world deployment; a positional encoding designed for this setting should therefore exhibit a controlled response to such perturbations.

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.

In this paper, we address both the theoretical and empirical aspects of this robustness. Our contribution addresses several issues: (i) we formalize simPE in terms of functional mapping and identify the regularity assumptions needed for stability analysis, (ii) we prove that simPE remains robust under rotations and derive perturbation bounds, including a global Frobenius-norm estimate and an explicit angle bound, and (iii) we provide experimental evidence of this robustness on five controlled datasets, including a real clinical chest X-ray benchmark, demonstrating that SimPE generally offers better stability compared to existing positional encodings under spatial rotation, particularly in small-to-moderate angular regimes.

2 Related Work

Standard Vision Transformers (ViTs) typically rely on learned absolute positional embeddings (Dosovitskiy et al., 2021); alternatively, to introduce dynamic spatial inductive biases, conditional positional encodings such as the Positional Encoding Generator (PEG) (Chu et al., 2023) derive positional information directly from visual features using local depthwise convolutions. Rotatory PE (RoPE) incorporates relative position via query-key rotations at every attention layer of the transformer (Su et al., 2023). On the other hand, SimPE, even tough it exploits convolution, is not based on explicit coordinates, and constructs spatial signals entirely from pairwise feature similarities. This represents an alternative paradigm that offer flexibility over rigid grid-based encoding, however its behavior under spatial transformations requires rigorous investigation.

In medical image analysis, accurate anatomical localization is critical, making models highly vulnerable to minor spatial variations, such as small patient or sensor rotations. While convolutional networks can achieve exact rotation equivariance through steerable or group-equivariant frameworks (Cohen & Welling, 2016; Weiler & Cesa, 2019), enforcing strict equivariance in Transformers is challenging because positional encodings explicitly break spatial symmetry.

However, in some practical scenarios (e.g. medical imaging) strict equivariance is often less crucial than *positional stability*: small rotations in input space should induce smooth, bounded changes in the positional representation without disrupting spatial relationships. Evaluating how different PE maintain this stability under rotation is a key to building geometrically robust vision models that may be safely applied in critical domains such a the medical one.

3 Preliminaries

3.1 Notation

Let $X = (x_1, \dots, x_N) \in \mathbb{R}^{N\times d}$ denote a collection of N token representations in dimension d. We denote by $\|\cdot\|$ the Euclidean norm for vectors and by $\|\cdot\|_F$ the Frobenius norm for matrices.

Let $SO(d)$ denote the special orthogonal group of $d \times d$ matrices,

$$SO(d) := \big\{R \in \mathbb{R}^{d\times d} : R^\top R = I,\ \det(R) = 1\big\}.$$

The orthogonality condition $R^\top R = I$ makes R a genuine rotation rather than an arbitrary linear map: it is equivalent to R preserving inner products, and hence norms and pairwise distances:

$$\|Rx\| = \|x\|, \qquad \|Rx - Ry\| = \|x - y\| \qquad \text{for all } x, y \in \mathbb{R}^d.$$

This property is essential for the present analysis, since it avoids to confuse pure rotation with a change of scale: any perturbation we measure in the encoding is ascribable to the reorientation of the input alone. Among the possible orthogonal matrices, we are interested in the special group, i.e., those matrices having $\det(R) = 1$, since this rules out reflections and other improper rotations (orientation-reversing isometries).

For $R \in SO(d)$, the rotated configuration of X is written as

$$RX := (Rx_1, \dots, Rx_N).$$

Let $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a similarity function. We define the similarity map $S : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times N}$, which maps a collection of tokens $X = (x_1, \ldots, x_N)$ to its corresponding matrix $S(X)$ with entries:

$$S(X)_{ij} = \kappa(x_i, x_j).$$

In a ViT, similarity-based positional encoding is modeled as a map

$$\text{simPE}(X) = \Phi(S(X)), \tag{1}$$

where $X \in \mathbb{R}^{N\times d}$ is the matrix resulting from a patch-wise convolution of the input image divided into N patches[1], and $\Phi : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times m}$ denotes a post-processing transformation that turns the patch similarity matrix into a usable per-token positional feature of dimension m[2].

In practice, Φ is implemented as a learned linear map applied to $S(X)$, reducing or reshaping the N-dimensional similarity profile of each token into an m-dimensional feature, analogous to the role of a learned embedding matrix acting on raw coordinates in absolute positional encoding. In other words, a projection weight matrix $W \in \mathbb{R}^{N\times m}$ is learned such that $\Phi = SW$.

Once the (m-dimensional) PE representation Φ has been computed, it can be added to the (m-dimensional) embedding of the original patches as presented in Santomauro et al. (2026).

3.2 Lipschitz Continuity

A map $f : \mathcal{X} \to \mathcal{Y}$ between normed spaces is Lipschitz continuous if there exists $L \geq 0$ such that

$$\|f(u) - f(v)\|_{\mathcal{Y}} \leq L\|u - v\|_{\mathcal{X}} \qquad \forall u, v \in \mathcal{X}.$$

In the following, we will base our theoretical analysis about the robustness to rotation of simPE on the notion of Lispschitz continuity.

4 Assumptions

In order to analyze the robustness property of SimPE we assume some standard regularity conditions on the input domain and the constituent operators as follows.

Assumption 1 (Bounded feature domain)**.** *There exists a constant $M > 0$ such that the embedding representations satisfy*

$$\|x_i\| \leq M \qquad \text{for all } i = 1, \ldots, N.$$

Equivalently, the admissible set of inputs is contained in a bounded subset of $\mathbb{R}^d$.

Assumption 1 directly follows from the standard ViT architecture where image patches are represented as vectors of flattened and bounded pixel values (Appendix A).

Assumption 2 (Lipschitz similarity on bounded sets)**.** *The similarity function $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is Lipschitz on bounded subsets of $\mathbb{R}^d \times \mathbb{R}^d$. In particular, under Assumption 1, there exists a constant $L_\kappa > 0$ such that for all $u, v, u', v' \in \mathbb{R}^d$ with*

$$\|u\|, \|v\|, \|u'\|, \|v'\| \leq M,$$

it follows that

$$|\kappa(u, v) - \kappa(u', v')| \leq L_\kappa\big(\|u - u'\| + \|v - v'\|\big).$$

Assumption 2 is valid for the similarities typically used in practice. For example, bounded bilinear similarities such as cosine similarity are Lipschitz on bounded sets, and smooth similarity maps (such as RBF kernel) are Lipschitz on compact subsets (Kreyszig, 1978;

[1] This is usually realized as a strided 2D convolution with d filters with spatial dimensions matching the patch size, and stride equal to filter dimensions.

[2] In the original definition presented in Santomauro et al. (2026), it is assumed $d = m$ even if, in general, d and m do not need to be equal.

Rockafellar & Wets, 1998). Moreover, a closed form solution for the Lipschitz constant L_κ can be obtained by computing the supremum of the norm of the gradient of the similarity function with respect to one parameter, yielding a value $L_\kappa = \frac{1}{\sigma\sqrt{e}}$ for an RBF similarity $k(x,y) = exp(-\frac{\|x-y\|^2}{2\sigma})$ (Vural & C.Guillemot, 2018), and a value of $L_\kappa = \frac{1}{\lambda}$ in case of cosine similarity $k(x,y) = \frac{x \cdot y}{\|x\|\|y\|}$, for bounded vectors $\|x\|, \|y\| \geq \lambda > 0$ (Qi et al., 2023).

Lemma 1. *Under Assumptions 1 and 2, the similarity matrix map $S : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times N}$ is Lipschitz with respect to the Frobenius norm. More precisely, for all admissible $X, Y \in \mathbb{R}^{N\times d}$,*

$$\|S(X) - S(Y)\|_F \leq 2L_\kappa\sqrt{N}\,\|X - Y\|_F.$$

Proof. For each pair (i,j), Assumption 2 yields

$$|S(X)_{ij} - S(Y)_{ij}| = |\kappa(x_i, x_j) - \kappa(y_i, y_j)| \leq L_\kappa\big(\|x_i - y_i\| + \|x_j - y_j\|\big).$$

Squaring both sides and using $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain

$$|S(X)_{ij} - S(Y)_{ij}|^2 \leq 2L_\kappa^2\|x_i - y_i\|^2 + 2L_\kappa^2\|x_j - y_j\|^2.$$

Summing over all i, j,

$$\|S(X) - S(Y)\|_F^2 \leq 4L_\kappa^2 N \sum_{i=1}^{N} \|x_i - y_i\|^2 = 4L_\kappa^2 N\|X - Y\|_F^2.$$

Taking square roots concludes the proof. □

5 Robustness to Rotations

Given the regularity assumptions on the input domain and model components previously introduced, we now prove that simPE is Lipschitz continuous, establishing its stability under rotational perturbations.

Theorem 1. $\text{simPE}(X) = \Phi(S(X))$ *is Lipschitz continuous. In particular, for all admissible inputs $X, Y \in \mathbb{R}^{N\times d}$,*

$$\|\text{simPE}(X) - \text{simPE}(Y)\|_F \leq 2L_\Phi L_\kappa\sqrt{N}\,\|X - Y\|_F.$$

Proof. Let us consider Φ as a parametric (on W) map of S: $\Phi_W(S) = SW$. Consider now two similarity matrices $S_1, S_2 \in \mathbb{R}^{N\times N}$, we have that

$$\|\Phi_W(S_1) - \Phi_W(S_2)\|_F = \|(S_1 - S_2)W\|_F$$

From the operator norm inequality we know that for any pair of matrices A, B,

$$\|AB\|_F \leq \|A\|_F\|B\|_2$$

where $\|\cdot\|_2$ is the spectral norm of a matrix. It follows that

$$\|(S_1 - S_2)W\|_F \leq \|W\|_2\|S_1 - S_2\|_F = \sigma_{\max}(W)\|S_1 - S_2\|_F$$

where $\sigma_{max}(W)$ is the largest singular value of W. By considering now S_1 and S_2 as similarity maps from the input X and Y respectively, using Lemma 1 we obtain

$$\|(S_1(X) - S_2(Y))W\|_F \leq 2\sigma_{max}(W)L_\kappa\sqrt{N}\,\|X - Y\|_F.$$

that proves the theorem by considering $L_\Phi = \sigma_{max}(W)$[3]. Constant L_κ depends on the similarity function as discussed when introducing Assunmption 2. □

[3]Since the Frobenius norm is an upper bound for the spectral norm, the use of the latter determines a stricter lower bound for the Lipschitz constant.

The theorem allows us to conclude that variations in the postional encoding representation are bounded by values that depends on the square root of the number of patches N, on the differences (distances) of the token representations (convolutional patch embeddings), on the maximum geometric deformation allowed by the learned weight matrix W ($\sigma_{max}(W)$), and on the chosen similarity function. In particular, we can observe that:

- increasing the number of patches can increase (in a sublinear way) the variation bound, meaning that using a finer grid in patching the image may decrease the robustness of the enconding, but in a limited way;
- enforcing a strict similarity in case of an RBF kernel (low bandwidth σ) is going to decrease the robustness, as well as a low value of the lower bound on the vector norms in case of a cosine similarity; if $\|x\|, \|y\| \geq \lambda > 0$, a small value of λ means that (as expecrted) token representations becomes more sensitive to perturbations because they have low magnitude;
- the way in which weights W are learned is going to determined how much small shifts in patch similarity may impact the final positional encoding representation (given that this impact is however bounded and cannot be amplified infinitely): a small maximum singular value of W will reduce the impact of such shifts increasing robustness[4].

Corollary 1 makes this observations explicit for rotations:

Corollary 1. *Let $R \in SO(d)$.*

$$\|\text{simPE}(RX) - \text{simPE}(X)\|_F \leq 2L_\Phi L_\kappa \sqrt{N}\, \|RX - X\|_F.$$

Proof. Apply Theorem 1 with $Y = RX$. □

Therefore, the closer R is to the identity, the smaller is the variation of simPE, all other things being equal (similarity function, number of patches and stretching factor of the learned weights).

The previous results are quite general. We now make the dependence on the rotation angle explicit.

Let $R_\theta \in SO(d)$ be a rotation of angle θ. Since R_θ is orthogonal,

$$\|R_\theta X - X\|_F \leq \|R_\theta - I\|_2 \|X\|_F.$$

It can then be proved (Appendix C) that:

$$\|R_\theta - I\|_2 = O(|\theta|).$$

Hence there exists a constant $c > 0$ such that

$$\|R_\theta X - X\|_F \leq c|\theta| \|X\|_F$$

for a rotation angle $|\theta|$.

Combining this with Corollary 1 provides the following result.

Corollary 2. *For a rotation angle θ,*

$$\|\text{simPE}(R_\theta X) - \text{simPE}(X)\|_F \leq 2cL_\Phi L_\kappa \sqrt{N}\, |\theta|\, \|X\|_F.$$

Proof. Substitute the estimate

$$\|R_\theta X - X\|_F \leq c|\theta| \|X\|_F$$

into Corollary 1. □

[4]This opens also the possibility of focusing the learning process in this direction, possibly incorporating a term related to the magnitude of singular values into the loss function.

Corollary 2 formally states that response to rotations in simPE are directly bounded in a linear way, to the rotation angle. In addition to what we have previously discussed (dependence from the similarity function, from the learned projection matrix and from the image patching grid), we also observe that the presence in the bound of the Frobenius norm of the convolutional embeddings X, relates the robustness to the quality of such embeddings as well. In fact, a large value of such norm represents the presence of embeddings of large magnitude, that is embeddings carrying signficant information; as expected, when the information is very significant, changes in the representation may change the encoding in a more relevant way.

6 Datasets and Experimental Protocol

To cleanly evaluate intrinsic geometric stability without the confounding effects of data-augmentation, all experiments follow a controlled protocol: models are trained and validated exclusively on unrotated images, while test sets are evaluated across discrete rotation angles. We use five datasets spanning synthetic, benchmark, and clinical domains, summarized in Table 1.

Table 1: Summary of experimental datasets and test evaluation protocol.

Dataset	Domain	Classes	Resolution	Train / Val	Test (per angle)	Test Rotation Angles
`Arrow`	Synthetic	4	224×224	20,000 / 4,000	4,000	$0°, 5°, 10°, 15°, 20°, 25°$
Digits	Synthetic	10	224×224	50,000 / 10,000	10,000	$0°, 15°, 30°, 45°, 60°, 75°, 90°$
Shapes	Synthetic	4	224×224	20,000 / 4,000	4,000	$0°, 15°, 30°, 45°, 60°, 75°, 90°$
FashionMNIST	Benchmark	10	28×28	50,000 / 10,000	10,000	$0°, 5°, 10°, 15°, 30°$
Chest X-Ray	Clinical	2	224×224	4,048 / 1,000	624	$0°, 5°, 10°, 15°, 20°, 25°, 30°$

Synthetic and Standard Benchmarks. To probe specific geometric ambiguities, we generated three synthetic datasets: *`Arrow`* (4-class pointing direction: up, down, left and right), *Digits* (10-class rendered digits with serif font and jitter), and *Shapes* (4 asymmetric primitives: arrow, L, T, and F-shapes). Instances feature randomized scale, thickness, and hue. Additionally, we evaluate on *FashionMNIST* (Xiao et al., 2017) as a standard benchmark for image classification.

Clinical Chest X-Ray Audit. For real-world evaluation, we use the *Chest X-Ray* dataset (Kermany et al., 2018) (binary classes *normal* vs. *pneumonia*). To verify that this clinical corpus satisfies our canonical-orientation requirement, we audited the dataset's intrinsic tilt using anatomical bilateral self-symmetry (minimizing MSE between rotated and horizontally mirrored images over a circular foreground mask). The estimated tilt across 5,031 valid training images yielded a median absolute deviation of 0.80° (mean 2.64°, 0.16° signed bias), confirming that the training data is effectively upright and valid for controlled evaluation.

Architecture and Positional Encodings. All models share an identical standard Vision Transformer (ViT) backbone (Dosovitskiy et al., 2021), taking as input 224×224 images which are split into 16×16 patches. The encoder consists of 4 layers, 4 attention heads, and a feed-forward network with a hidden dimension of 512, with a linear classification head on the `[CLS]` token. Keeping the backbone fixed, we compare five positional encoding strategies: **simPE** (Santomauro et al., 2026), **Learned PE** (Dosovitskiy et al., 2021), **Sinusoidal PE** (Vaswani et al., 2017), **PEG** (Chu et al., 2023), **RoPE** (Su et al., 2023) and a baseline omitting positional encodings (**no-PE**). The similarity function adopted for **SimPE** is the cosine similarity, while the projection weight matrix W is learned through a simple linear layer.

Training and Hyperparameters. Models are trained using AdamW (learning rate 10^{-4}) with a cosine annealing schedule for up to 200 epochs (batch size 128, early stopping patience of 20 epochs). To strictly evaluate intrinsic geometric robustness without confounding factors, no rotation data augmentation is applied during training or validation.

7 Experimental Results

Table 2 summarizes the classification accuracy and F1-score across varying rotation angles for all evaluated datasets and PEs. We also reports results at no rotation to evaluate the degree of performance loss once a rotation occurs.

On the `Arrow` dataset, both no-PE and SLPE experience early degradation at 10°, with a 18.50% and 23.50% drop in accuracy (and losing 0.19 and 0.20 in F1-score) respectively. While PEG maintains solid performance up to 15°, its accuracy drops sharply to 39.00% (F1: 0.45) at 20°. In contrast, SimPE performs better than Sin at 5° (99.50% vs. 98.50%) and 10° (99.50% vs. 97.00%), keeping a stable performance across both initial angles. Although Sin yields higher accuracy and F1-score at 15° (95.00% and 0.94, respectively) and performs on par with SimPE at 20° in terms of F1-score (0.88), SimPE achieves the highest overall robustness under severe rotation (25°), outperforming all other methods with an accuracy of 82.50% and an F1-score of 0.83.

On the `Digits` dataset, SimPE outperforms every other method at each angle, in terms of both accuracy and F1-Score. An exception occurs at 45°, where no-PE exhibits a slightly higher accuracy (32.60% vs. 31.50%), whereas SimPE's F1-Score is higher. We can notice that all methods experience severe degradation beyond 30°, with the only exception being RoPE, which degrades more slowly.

On the `FashionMNIST` dataset, no-PE consistently yields the lowest performance across all evaluation angles. At mild rotations, (up to 15°), SimPE shows the best performance, with RoPE and PEG that remain competitive. SimPE exhibits some problems when rotation becomes much more significant (30°) where however, no competing PE method is able to provide significant results.

On the `Shapes` dataset, SimPE shows again the best accuracy at a mild rotation of 15°,while PEG performes better at larger rotation angles (up to 60°).

Finally, on the medical dataset (`Chest X-Ray`), we have considered a more fine grained grid of rotation angles (to model a set of possible and realistic perturbations of the images); in this case, SimPE consistently achieves the highest Accuracy and F1-Score across all tested rotation angles (5° to 30° with a 5° step). Notably, SimPE demonstrates high stability under all the tested rotations, with significative values of both accuracy and F1-score. In this dataset, PEG emerges as the closest competitor, also exhibiting steady behavior with accuracy dropping from 84.13% at 5° to 79.33% at 30°; RoPE performance also maintains a quite steady accuracy. In contrast, Sin and SLPE suffer more pronounced degradations as the angle increases, while no-PE remains ineffective throughout, hovering between 50.48% and 56.57% accuracy regardless of rotation.

Overall, evaluating the performance drop between 0° and the initial rotation angle reveals that SimPE is consistently the most robust method against mild tilts, suffering minimal accuracy losses across all datasets (e.g., losing only 0.05% on `Shapes` and 0.80% on `Chest X-Ray`). While RoPE and PEG also demonstrate strong initial resistance, baseline methods like no-PE and SLPE experience immediate, severe degradation (dropping by up to 40.23% and 32.50% on `Digits` at just 15°).

8 Discussion

The analysis and experiments developed above highlight an important distinction between invariance and robustness. Exact invariance is a strong property requiring that the positional encoding remain unchanged under rotation; simPE (as well as the other PE methods) generally does not satisfy this requirement (Appendix B). However, in many practical scenarios, exact invariance is neither necessary nor desirable: orientation often carries useful discriminative information, and discarding it entirely would reduce the encoding's expressiveness (Cohen & Welling, 2016). What is needed instead is a weaker form of stability ensuring that small geometric perturbations induce only small changes in the encoding. This is captured through Theorem 1 and Corollary 2.

Table 2: Experimental results for Accuracy (%) and F1-Score as a function of the rotation angle.The best value in each column is highlighted in **bold**.

	Accuracy (%)							F1-Score						
`Arrow`														
PE	**0°**	**5°**	**10°**	**15°**	**20°**	**25°**	-	**0°**	**5°**	**10°**	**15°**	**20°**	**25°**	-
no-PE	99.50	99.00	80.50	73.00	26.50	3.50	-	0.99	**0.99**	0.80	0.72	0.26	0.02	-
Sin	99.75	98.50	97.00	**95.00**	**89.00**	81.00	-	0.99	0.98	0.97	**0.94**	**0.88**	0.79	-
SLPE	99.50	98.00	74.50	65.50	44.50	33.50	-	0.99	0.98	0.78	0.69	0.40	0.25	-
RoPE	99.75	99.00	96.50	90.50	80.50	79.50	-	0.99	**0.99**	0.96	0.90	0.80	0.79	-
SimPE	99.75	**99.50**	**99.50**	89.50	88.00	**82.50**	-	0.99	**0.99**	**0.99**	0.90	**0.88**	**0.83**	-
PEG	99.75	99.00	96.50	84.00	39.00	23.50	-	0.99	**0.99**	0.96	0.83	0.45	0.22	-
`Digits`														
PE	**0°**	**15°**	**30°**	**45°**	**60°**	**75°**	**90°**	**0°**	**15°**	**30°**	**45°**	**60°**	**75°**	**90°**
no-PE	97.43	57.20	38.80	32.60	13.40	6.15	4.10	0.97	0.66	0.34	0.32	0.16	0.08	0.03
Sin	99.50	55.80	30.40	24.20	18.80	14.40	7.40	0.99	0.64	0.29	0.23	0.16	0.15	0.08
SLPE	98.80	66.30	34.40	12.00	11.50	7.90	3.00	0.97	0.73	0.32	0.13	0.06	0.04	0.02
RoPE	97.00	58.80	54.30	**50.60**	**43.60**	**31.40**	**29.00**	0.95	0.54	0.53	**0.44**	**0.37**	**0.24**	**0.22**
SimPE	99.93	**86.20**	**56.90**	31.50	19.50	15.10	18.70	0.99	**0.88**	**0.62**	0.35	0.25	0.17	**0.22**
PEG	99.00	81.60	41.40	17.00	13.80	9.60	12.60	0.99	0.81	0.58	0.21	0.18	0.11	0.13
`FashionMNIST`														
PE	**0°**	**5°**	**10°**	**15°**	**30°**	-	-	**0°**	**5°**	**10°**	**15°**	**30°**	-	-
no-PE	69.43	60.30	54.17	42.99	29.40	-	-	0.69	0.63	0.59	0.40	0.22	-	-
Sin	84.02	81.38	77.01	64.46	**42.00**	-	-	0.84	0.81	0.77	0.67	**0.33**	-	-
SLPE	83.71	76.69	70.82	57.21	32.84	-	-	0.82	0.77	0.71	0.62	**0.33**	-	-
RoPE	87.23	84.14	76.83	68.36	32.82	-	-	0.87	0.84	0.78	0.67	0.25	-	-
SimPE	87.31	**84.49**	**77.61**	**69.39**	23.84	-	-	0.87	**0.85**	**0.79**	**0.69**	0.25	-	-
PEG	83.05	80.16	76.03	66.66	27.76	-	-	0.82	0.79	0.75	0.65	0.19	-	-
`Shapes`														
PE	**0°**	**15°**	**30°**	**45°**	**60°**	**75°**	**90°**	**0°**	**15°**	**30°**	**45°**	**60°**	**75°**	**90°**
no-PE	77.25	63.75	57.75	41.00	43.50	**52.25**	**51.75**	0.68	0.56	0.50	0.36	0.38	0.45	**0.46**
Sin	99.50	95.00	89.25	65.75	37.50	27.75	19.25	0.98	0.94	0.89	0.65	0.30	0.21	0.19
SLPE	84.87	76.75	72.25	69.75	48.50	42.50	40.34	0.87	0.79	0.73	0.71	0.49	**0.47**	0.42
RoPE	99.65	93.50	69.25	50.25	32.25	20.50	8.75	0.99	0.93	0.64	0.43	0.24	0.14	0.09
SimPE	99.05	**99.00**	81.00	63.50	53.00	41.00	39.50	0.98	**0.97**	0.82	0.65	**0.56**	0.44	0.43
PEG	99.75	96.50	**92.75**	**76.50**	**54.50**	48.75	49.00	0.99	0.96	**0.92**	**0.73**	0.46	0.37	0.38
`Chest X-Ray`														
PE	**0°**	**5°**	**10°**	**15°**	**20°**	**25°**	**30°**	**0°**	**5°**	**10°**	**15°**	**20°**	**25°**	**30°**
no-PE	64.01	54.01	53.53	54.01	56.57	50.48	54.01	0.58	0.42	0.40	0.41	0.47	0.34	0.42
Sin	81.24	78.21	72.44	66.51	63.94	65.87	68.59	0.81	0.80	0.74	0.65	0.62	0.66	0.72
SLPE	73.98	70.83	69.87	66.35	63.62	61.54	56.73	0.72	0.71	0.70	0.65	0.60	0.56	0.47
RoPE	83.49	82.69	82.37	81.25	81.19	79.17	79.00	0.87	0.86	0.86	0.85	0.83	0.82	0.80
SimPE	86.70	**85.90**	**86.06**	**85.74**	**84.74**	**84.46**	**83.97**	0.87	**0.87**	**0.87**	**0.85**	**0.83**	**0.83**	**0.81**
PEG	86.52	84.13	83.65	83.10	81.40	80.93	79.33	0.86	0.86	0.84	0.80	0.79	0.78	0.77

The theoreticsl mechanism underlying the study of robustness is the Lipschitz continuity of the constituent components of simPE. Under mild regularity assumptions (Assumptions 1–2) simPE is Lipschitz continuous, and its response to rotations is bounded by the product of the specific Lipschitz constants (depending on the similarity function used, and on the learned projection matrix weights) and the magnitude of the rotation.

From the experimental point of view, across all evaluated datasets, omitting positional encoding (no-PE) leads to severe performance degradation as rotation increases. On the other hand SimPE consistently exhibits better robustness when considering the small-to-moderate rotation angle regimes. This is particularly interesting, since in practical applications we either would not expect severe rotations to occur (as in the case of medical imaging (Sun et al., 2024)) or, if such large rotations occur, they are actually modifying in a significative way the image and more specific ways (which are not based on the change in the positional encoding method) have to be devised to deal with the issue (Jaderberg et al., 2016; Hutchinson et al., 2021).

Performance crossovers at extreme angles reflect domain-specific structural priors rather than a failure of the theoretical stability guarantees. This distinction is especially important in medical imaging, the original domain for which simPE was conceived (Santomauro

et al., 2026). In this setting, small rotations or artificial perturbations are not rare, and may lead to significant errors in image classification (Finlayson et al., 2019; Tsai et al., 2023). They naturally arise from the acquisition process itself; instrument configuration, patient placement, and subtle differences in imaging setup can all introduce small rotational discrepancies across scans. From the practical point of view, our results on the `Chest X-Ray` dataset confirm that simPE is significantly more robust that alternative PE methods. Notably, the validation process has been designed in order to avoid information leakege in the training set; indeed, we checked, via the self-symmetry tilt analysis, that the training images used to instantiate the canonical-orientation protocol did not themselves carry a confounding amount of intrinsic rotation. Had the training set exhibited substantial pre-existing tilt, the observed test-time degradation could not have been cleanly attributed to the applied rotation alone. This methodological step is likely to be necessary whenever this type of controlled rotation-robustness protocol is applied to real, uncurated clinical corpora rather than to synthetically generated data.

9 Conclusions and Future Works

This paper presented a unified theoretical and empirical investigation into the rotational robustness of similarity-based positional encoding (SimPE). Theoretical analysis demonstrates that SimPE exhibits provable stability against rotational perturbations under Lipschitz regularity assumptions. Specifically, we derived explicit perturbation bounds and a global Frobenius norm estimate that decouples the architectural Lipschitz constants from the magnitude of geometric perturbation. Empirically, we validated these theoretical findings across five controlled datasets spanning synthetic benchmarks (`Arrow`, `Shapes`, `Digits`), standard computer vision (`FashionMNIST`), and a real-world clinical imaging (`Chest X-Ray`). Evaluating models trained on canonical orientations against increasing rotation angles revealed that SimPE generally provides better stability and performance compared to the other evaluated PEs. This advantage is most pronounced in the small-to-moderate angular regime ($\leq 15°$), which corresponds to the clinically safe operating range in medical imaging. For the clinical benchmark, the canonical-orientation premise was explicitly verified via a self-symmetry tilt analysis, confirming a median intrinsic tilt below $1°$ across the training set. We believe that, collectively, these results establish a mathematical and empirical foundation for SimPE as a robust, structure-preserving positional encoding.

Concering limitations and future directions, while this work proposes a theoretical and empirical framework for SimPE under rotation, several promising avenues for future exploration remain. First, even if our theoretical formulation generally applies to n-dimensional spatial fields, our empirical benchmarks deliberately prioritize representative 2D visual domains; an experimental analysis concerning 3D-imaging domains (such as CT or MRI scans in a medical setting) is then planned to consolidate the results presented here. Second, our stability guarantees are derived to hold without restricting the similarity function; specializing these bounds with closed-form Lipschitz constants (L_Φ, L_κ) for specific kernels (as done for cosine and RBF) offers a pathway for fine-grained, domain-tailored bounds. Third, the influence of the learned projection matrix and of the quality of the embeddings obtained trough convolution on the robustness, suggests that specific learning procedures may improve the latter; in particular, designing a specific loss function for the learning of the projection matrix, penalizing large singular values may be promising. Finally, as SimPE is designed as a modular, model-agnostic layer, studying its theoretical interactions when paired with approximate equivariant backbones provides an interesting direction for end-to-end architectural modeling.

Ethics Statement

This work relies on a combination of standard, publicly available benchmark datasets and synthetic data. The public benchmarks were used strictly in compliance with their respective open-source licenses. Synthetic data was generated solely to meet the requirements of our controlled experimental setup. Furthermore, to ensure research integrity and full repro-

ducibility, the code used to generate all synthetic datasets is provided in the supplementary material. To the best of our knowledge, this work introduces no ethical or safety concerns.

AI Use Statement

Large Language Models (LLMs) were used solely as an supportive tool to assist in discovering, searching, and verifying relevant literature and references related to this work. All cited sources were subsequently checked and manually verified by the authors.

Reproducibility Statement

To ensure full reproducibility of our work, all source code, dependencies, and execution scripts required to generate the synthetic datasets and run the experimental pipeline are included in the anonymized supplementary material (`.zip` archive). A detailed `README.md` file with setup instructions, environment specifications, and commands to reproduce all main results is provided within the archive.

References

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=4PV8nJ6Y0Y.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2990–2999. PMLR, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane, and Andrew L. Beam. Adversarial attacks against medical deep learning systems, 2019. URL https://arxiv.org/abs/1804.05296.

Eric Fisk and et al. The hip center rule can be used to decide if measured pelvic incidence is accurate. *Spine*, 46(18):E1005–E1012, 2021. doi: 10.1097/BRS.0000000000004012.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1243–1252. PMLR, 2017.

Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*. Springer, Cham, 2nd edition, 2015. ISBN 978-3-319-13466-6. doi: 10.1007/978-3-319-13467-3.

Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4533–4543. PMLR, 2021. URL https://proceedings.mlr.press/v139/hutchinson21a.html.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks, 2016. URL https://arxiv.org/abs/1506.02025.

Lukas Jud, Tobias Roth, Silvan Beeler, Lazaros Vlachopoulos, Reto Sutter, and Sandro F. Fucentese. Significant changes in lower limb alignment due to flexion and rotation. *BMC Musculoskeletal Disorders*, 24(1):233, 2023. doi: 10.1186/s12891-023-06338-x.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y. L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A. N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. doi: 10.1016/j.cell.2018.02.010.

Erwin Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, New York, 1978. ISBN 978-0-471-50459-7.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. doi: 10.1109/ICCV48922.2021.00986.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4055–4064. PMLR, 2018.

Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, and Lei Zhang. Lipsformer: Introducing lipschitz continuity to vision transformers. In *Proceedings 11th International Conference on Learning Representations (ICLR 23)*, 2023. URL https://openreview.net/pdf?id=cHf1DcCwcH3.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1998. ISBN 978-3-540-62772-2. doi: 10.1007/978-3-642-02431-3.

Andrea Santomauro, Giorgio Leonardi, and Luigi Portinale. Comparing different positional encodings for the interpretation of medical images. In *Communications in Computer and Information Science*, volume 2830, pp. 78–89. Springer Nature Switzerland, 2026. ISBN 978-3-032-16708-8.

Klaus A. Siebenrock, Daniel F. Kalbermatten, and Reinhold Ganz. Effect of pelvic tilt on acetabular retroversion: a study of pelvises from cadavers. *Clinical Orthopaedics and Related Research*, 407:241–248, 2003. doi: 10.1097/01.blo.0000049697.72887.05.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Jiuai Sun, Pengfei Hou, Kai Li, Ling Wei, Ruifeng Zhao, and Zhonghang Wu. Automated estimation of thoracic rotation in chest x-ray radiographs: a deep learning approach for enhanced technical assessment. *British Journal of Radiology*, 97(1162):1690–1695, 2024. URL https://doi.org/10.1093/bjr/tqae149.

Moritz Tannast, Stefan Fritsch, Guoyan Zheng, Klaus A. Siebenrock, and Simon D. Steppacher. Which radiological parameters of acetabular retroversion are dependent on pelvic tilt and rotation? *Clinical Orthopaedics and Related Research*, 472(5):1461–1474, 2014. doi: 10.1007/s11999-013-3331-x.

Min-Jen Tsai, Ping-Yi Lin, and Ming-En Lee. Adversarial attacks on medical image classification. *Cancers*, 15(17), 2023. URL https://www.mdpi.com/2072-6694/15/17/4228.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

E. Vural and C.Guillemot. A study of the classification of low-dimensional data with supervised manifold learning. *Journal of Machine Learning Research*, 18:1–55, 2018.

Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

A Validity of Assumption 1

Consider a ViT architecture and an image patch $x \in \mathbb{R}^{P^2 \cdot C}$ (where P is the patch size and C is the number of channels); the patch x is projected into an embedding space of size d via a learned linear matrix $W_p \in \mathbb{R}^{d \times (P^2 C)}$ and a bias b_p:

$$E(x) = W_p x + b_p$$

Pixel values are naturally constrained to a closed, bounded interval which is typically normalized to $x_i \in [0, 1]$; it follows from the definition that the patch norm $\|x\|$ is bounded by $\sqrt{P^2 C}$. Let $k = P^2 C$:

$$\|x\| = \sqrt{\sum_{i=1}^{k} x_i^2} \leq \sqrt{\sum_{i=1}^{k} 1^2} = \sqrt{P^2 C}$$

By the operator norm inequality:

$$\|E(x)\| \leq \|W_p\| \|x\| + \|b_p\| \leq \|W_p\| \left(\sqrt{P^2 C}\right) + \|b_p\| = C_{\mathrm{p}}$$

Here, C_{p} is a fixed, finite constant determined entirely by the model weights and the image chosen patch size and number of channels.

B simPE is not Rotation-Invariant

We prove it by determine an example for which rotation invariance is not true.

Proposition 1. *Let f denote the patch feature extraction mapping $f : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^{N \times d}$, and let $\mathcal{R}_\theta$ be a spatial image rotation by an angle $\theta \notin 2\pi\mathbb{Z}$ (i.e. θ does not belong to the set of integer multiples of 2π). Then, there exists an image I such that* $\mathrm{simPE}(\mathcal{R}_\theta I) \neq \mathrm{simPE}(I)$.

Proof. Let I be an image consisting of a uniform background with a small bright spot located near the image center where four patches intersect. Choose the spot's angle such that a rotation by θ moves it into a different patch among these four; this is always possible since $\theta \notin 2\pi\mathbb{Z}$. Let p denote the index of the patch containing the spot prior to rotation.

All patches containing exclusively background produce an identical token representation x_b, whereas patch p produces a token $x_p \neq x_b$. Let k be one of the four central patches that contains the spot neither before nor after the rotation.

Assuming a normalized similarity function κ (such as cosine similarity, where $\kappa(u, u) = 1$), the similarity matrix entries evaluate to:

$$S(f(I))_{pk} = \kappa(x_p, x_b) < 1, \qquad S(f(\mathcal{R}_\theta I))_{pk} = \kappa(x_b, x_b) = 1,$$

since after rotation both patch p and patch k contain only background.

Thus, $S(f(\mathcal{R}_\theta I)) \neq S(f(I))$. Consequently, $\mathrm{simPE}(\mathcal{R}_\theta I) \neq \mathrm{simPE}(I)$. □

C Order of magnitude of rotations

Let us parameterize the continuous rotation by $t \in [0, \theta]$ as a curve $R(t) \in SO(d)$ such that $R(0) = I$ and $R(\theta) = R_\theta$:

$$R_\theta - I = \int_0^\theta \frac{d}{dt} R(t)\, dt$$

For any rotation path, the velocity vector is given by $\frac{d}{dt}R(t) = WR(t)$, where $W = \frac{dR}{dt}\big|_{t=0}$ is an anti-symmetric matrix (Hall, 2015). Taking the spectral norm on both sides and applying the integral triangle inequality we get:

$$\|R_\theta - I\|_2 = \left\| \int_0^\theta WR(t)\, dt \right\|_2 \leq \int_0^{|\theta|} \|WR(t)\|_2\, dt$$

Because $R(t)$ is an orthogonal matrix, its spectral norm is $\|R(t)\|_2 = 1$. Therefore, $\|WR(t)\|_2 \leq \|W\|_2 \|R(t)\|_2 = \|W\|_2$:

$$\|R_\theta - I\|_2 \leq \int_0^{|\theta|} \|W\|_2\, dt = \|W\|_2 |\theta|$$

Since $\|W\|_2$ is a constant independent of θ, it follows immediately that:

$$\|R_\theta - I\|_2 = O(|\theta|)$$