What Positional Embeddings Really Do In Vision Transformers

Anonymous Authors[1]

Abstract

Positional embeddings (PEs) in Vision Transformers (ViTs) are typically viewed as mechanisms for injecting absolute spatial information. We show that ViTs trained without PEs can nonetheless recover non-trivial spatial structure using patch content alone, questioning the fundamental functional role of PEs. We demonstrate that PEs induce a sharp increase in early-layer representational diversity, characterized by higher effective rank and reduced token homogenization. However, our analysis reveals that this diversity alone is insufficient for robustness. Using a new metric, Spatial Similarity Distance Correlation (SSDC), we show that PEs facilitate a qualitative shift from content-based to absolute-position-based spatial organization, yielding representations that remain stable under distributional shifts where content-only models fail.

1. Introduction

Vision Transformers (ViTs) have emerged as a powerful alternative to convolutional architectures for visual recognition by modeling images as sequences of patch tokens processed through self-attention mechanisms (Dosovitskiy et al., 2021). Unlike Convolutional Neural Networks (CNNs), ViTs lack strong built-in biases toward locality and translation equivariance. To compensate, most ViT architectures rely on Positional Embeddings (PEs) to inject explicit spatial information, allowing the model to distinguish patches originating from different locations in the image.

Existing studies suggest that ViTs can retain substantial performance even when positional information is removed or degraded (Chu et al., 2023). This potentially indicates that transformers partially reconstruct spatial relationships from patch content alone, analogous to how CNNs learn implicit positional information from zero-padding (Islam* et al., 2020). These findings challenge the conventional view that PEs are strictly necessary and raise fundamental questions about what functional advantages they provide beyond basic spatial identifiability.

Prior work has largely explored positional embeddings through downstream performance or architectural variations. While informative, such approaches reveal little about how PEs shape the internal representations of the model. In particular, the effects of positional embeddings on the geometry, dimensionality, and stability of token representations across the transformer stack remain poorly understood.

In this work, we adopt a mechanistic perspective to study the role of positional embeddings in ViTs. We analyze the evolution of token representations in the residual stream using tools from representational geometry (Raghu et al., 2021), introducing the Spatial Similarity Distance Correlation (SSDC), a metric designed to quantify how spatial relationships are reflected in token similarity patterns. Using this framework, we systematically compare ViTs trained with and without positional embeddings, examining their impact on representational dimensionality, the spatial reasoning strategies employed by the model, and robustness to distributional shifts such as stylization and noise. Using this framework, we demonstrate that PEs perform three critical functions:

- **Promoting Representational Diversity:** PEs causally increase representational diversity (characterized by high dimensionality and lower token similarities), thereby enriching internal geometric representations.
- **Shifting Spatial Organization Strategy:** We demonstrate that PEs push ViTs from a dominantly content-based strategy of forming internal spatial structure to an absolute-position-based strategy, where the model can always reason "where" a patch is without a great dependence on what it "looks" like.
- **Enhancing Robustness to Distributional Shifts:** We show that this absolute-position-based shift is precisely what allows ViTs with PEs to remain robust to distributional shifts, such as stylization, where local patch content becomes unreliable.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Together, our results provide a detailed and novel view of how explicit positional signals shape the internal organization of vision transformers, offering new insights into why positional embeddings play a critical role in stable and robust visual representation learning.

2. Background and Setup

2.1. Vision Transformer Architecture

All models used in our experiments are vanilla Vision Transformers trained from scratch, with approximately 12M parameters. Images are divided into fixed-size patches, which are linearly projected into token embeddings and processed by a stack of self-attention and feedforward layers. When present, positional embeddings are learned parameters optimized jointly with the rest of the model. No architectural modifications or auxiliary inductive biases are introduced beyond standard ViT components.

2.2. Positional Embedding Ablation

To isolate the functional role of positional embeddings, we train a parallel set of models in which positional embeddings are entirely removed. Throughout the paper, we refer to models trained without positional embeddings as *ablated models*, and to models trained with positional embeddings as *intact models*. Apart from the presence or absence of positional embeddings, all architectural choices, optimization settings, and training procedures are held constant.

2.3. Datasets

We evaluate models on CIFAR-10 and a stylized variant derived from CIFAR-10 to assess robustness under distributional shift. The stylized dataset is generated using Adaptive Instance Normalization (AdaIN) with a mixing coefficient $\alpha = 0.1$, which significantly alters texture statistics while preserving coarse spatial structure. Models are trained on standard CIFAR-10 images and evaluated on both the original and stylized datasets.

3. Methods

3.1. Residual Stream Geometry

To analyze the evolution of internal representations across depth, we extract the residual stream at selected layers of the model (layers 0, 2, 4, and 9, where layer 9 corresponds to the final layer). At each layer, we represent the residual stream as a matrix $R \in \mathbb{R}^{T\times C}$, where T denotes the number of tokens and C the embedding dimension. Each row of R corresponds to the residual stream representation of a single token.

Given the singular values $\{\sigma_i\}$ of R, we compute the effective rank using the participation ratio:

$$\mathrm{ER} = \frac{\left(\sum_i \sigma_i^2\right)^2}{\sum_i \sigma_i^4}.$$

Effective rank quantifies the number of dimensions that meaningfully contribute to the representation, with higher values indicating more distributed and heterogeneous representations.

In addition, we compute pairwise cosine similarities between all token representations in R to form a token cosine similarity matrix, where the (i, j)-th entry corresponds to the cosine similarity between tokens i and j. This matrix is symmetric by construction. We average the token cosine similarity matrix across the batch dimension to obtain a layer-wise summary of inter-token relationships. This matrix serves both as a proxy for inter-token heterogeneity and as the basis for computing the Spatial Similarity Distance Correlation (SSDC).

3.2. Spatial Similarity Distance Correlation

To quantify the emergence of spatial structure, we introduce the Spatial Similarity Distance Correlation (SSDC). For a given layer, we compute the pairwise cosine similarity matrix between token representations and the corresponding matrix of pairwise spatial distances between token positions, where Spatial distance is defined as the Manhattan distance between patch coordinates on the image grid. SSDC is defined as the Spearman rank correlation between cosine similarity and the negative spatial distance, such that higher values indicate that tokens which are spatially closer tend to have more similar representations. Because SSDC measures the monotonic alignment between spatial proximity and representational similarity, it serves as a proxy for the presence of relative positional structure in the residual stream. We use Spearman rank correlation to remain agnostic to the precise functional form relating spatial distance and representational similarity.

3.3. Fragility Score

To quantify a model's sensitivity to distributional shifts, we define a simple *Fragility Score* (FS), which measures the relative drop in top-1 accuracy under distribution shift. It is defined as

$$\mathrm{FS} = 1 - \frac{A_{\text{shift}}}{A_{\text{normal}}},$$

where A_{normal} and A_{shift} denote top-1 accuracy on the normal and shifted datasets, respectively. Higher values indicate greater performance degradation.

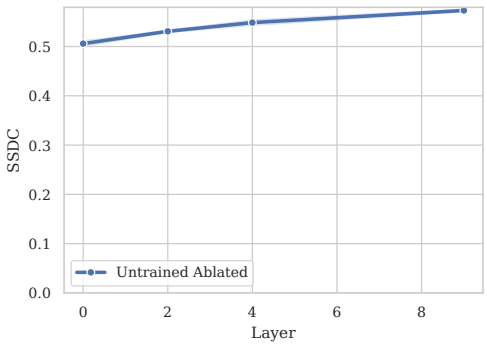


Figure 1. **SSDC across depth for an untrained ablated model.** SSDC remains approximately constant and at a relatively high value across layers, indicating static spatial correlations induced by architectural and data priors rather than learning. Shaded regions indicate variability across runs (±1 standard deviation).

4. Results

4.1. Architectural Priors Induce Static Spatial Correlations at Initialization

Experimental Setup: We evaluate SSDC at layers 0, 2, 4, and 9 on the CIFAR-10 dataset using untrained ablated models. We also extract from each one of these layers the Token Cosine Similarity Matrix and plot them using the same color scale.

Results: The untrained ablated model exhibits a non-zero SSDC (0.5) that remains approximately constant across depth (Fig. 1). This behavior is consistent across runs and indicates the presence of static spatial correlations induced by architectural and data priors rather than learning. Importantly, SSDC magnitude alone is insufficient to characterize learned spatial organization; instead, changes in SSDC across depth are the relevant signal.

This static spatial structure can be visually demonstrated by token cosine similarity heatmaps that are shown in Appendix Fig. A1.

This provides a static baseline against which we can measure the emergence of learned spatial structure in trained models.

4.2. Validating Emergent Spatial Structure via Extreme Counterfactuals

Experimental setup: We compute Token Cosine Similarity Matrices from layers 0, 2, 4, and 9 and evaluate the SSDC on each of these layers. We compare three settings: (i) an untrained ablated model with tokens randomly permuted at inference time, serving as an extreme baseline with no spatial structure; (ii) a trained model without positional embeddings (trained ablated); and (iii) a fully trained intact model.

Results: The untrained ablated model with random permutation consistently exhibits a near-zero and depth-invariant SSDC value, reflecting the absence of meaningful spatial organization in token representations. In contrast, the trained ablated model shows a clear and consistent increase in SSDC from layer 0 to layer 2, after which SSDC remains roughly constant across deeper layers (Fig. 3). This depth-dependent increase indicates the emergence of non-trivial spatial structure during training, despite the absence of explicit positional embeddings.

As expected, the trained intact model exhibits higher SSDC values across all layers, consistent with the direct contribution of explicit positional embeddings. Importantly, however, the presence of a strong SSDC increase in the trained ablated model demonstrates that spatial structure is not solely inherited from positional embeddings, but can emerge implicitly through training dynamics and architectural biases.

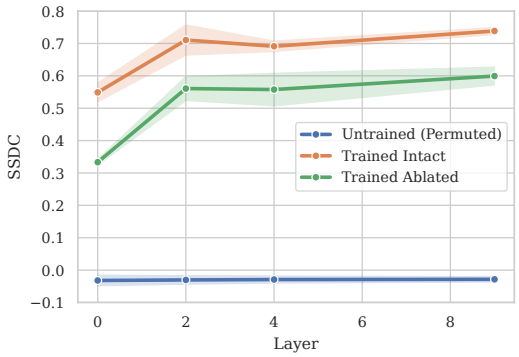
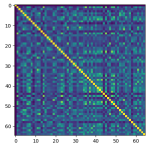
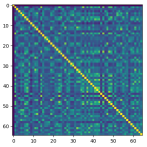


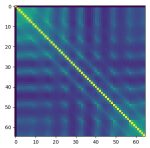
Figure 2. **SSDC across depth for an untrained ablated model with random permutation, a trained ablated model, and a trained intact model.** Shaded regions indicate variability across runs (±1 standard deviation).



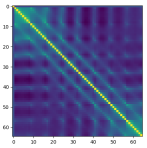
(a) Untrained Ablated Model with Random Permutation Layer 0



(b) Untrained Ablated Model with Random Permutation Layer 2

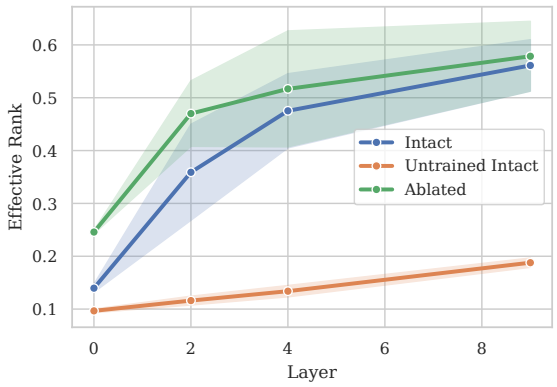


(c) Trained Ablated Model Layer 0

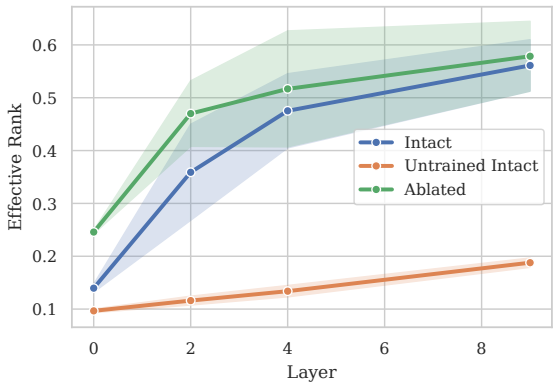


(d) Trained Ablated Model Layer 2

Figure 3. **Representative token cosine similarity matrices.** All matrices share the same color scale.



(a) Effective Rank across depth (Intact vs Post-hoc Ablated vs Untrained Intact).



(b) Mean Token Cosine Similarity across depth (Intact vs Post-hoc Ablated vs Untrained Intact).

Figure 4. Early-layer representational diversity under intact and post-hoc ablated positional embeddings. Left: Effective Rank of the residual stream across layers. Right: Mean token-wise cosine similarity across layers. Results are shown for a fully intact ViT, the same trained model with positional embeddings removed post-hoc at inference time, and a completely untrained intact model. The intact model exhibits substantially higher Effective Rank at layer 0 followed by an early collapse, while the post-hoc ablated model starts from a low-rank regime and remains consistently lower. Mean token cosine similarity is higher for the post-hoc ablated model across all layers.

4.3. Positional Embeddings Inject Early-Layer Representational Diversity

Experimental setup: We evaluate representational diversity in Vision Transformers using two complementary metrics: the Effective Rank of the residual stream and the mean token-wise cosine similarity. Experiments are conducted on CIFAR-10 using two models: (1) a standard ViT with intact positional embeddings, (2) the same trained model with positional embeddings removed post-hoc at inference time, and (3) an untrained model with untrained PEs initialized at random.

Effective Rank is computed at layers 0, 2, 4, and 9, following standard practice as a proxy for the dimensional diversity of representations. In parallel, we compute the token-to-token cosine similarity matrix at the same layers and report its mean value as a measure of representational collapse across spatial tokens.

Crucially, because positional embeddings are removed only at inference time, any observed differences reflect the functional role of positional information during forward computation rather than differences in learned parameters. Additionally, by observing the behavior of the untrained intact model, we can distinguish functionally meaningful representational diversity from representational diversity that is a result of random noise injections.

Results: Figure 4 reveals a pronounced difference in representational diversity between intact and post-hoc ablated models that emerges immediately at the input layer. At layer 0, the intact model exhibits an Effective Rank approximately four times higher than that of the post-hoc ablated model, indicating that positional embeddings inject substantial high-dimensional diversity directly into the residual stream prior to any attention-based mixing.

Strikingly, this initial advantage does not persist unaltered. The intact model undergoes a sharp collapse in Effective Rank within the first two layers, after which its rank stabilizes and remains only slightly higher than that of the post-hoc ablated model for the remainder of the network. In contrast, the post-hoc ablated model begins in a low-rank regime and does not exhibit a comparable early collapse, instead maintaining a relatively stable but consistently lower Effective Rank across depth.

Mean token cosine similarity provides a complementary perspective. Across all layers and all runs, the post-hoc ablated model exhibits significantly higher token-wise similarity than the intact model, indicating persistent spatial homogenization when positional embeddings are removed. This elevated similarity is present from the earliest layer and does not disappear with depth, suggesting that the lack of positional information induces a sustained form of representational collapse across tokens.

Comparing the evolution of representational diversity across depth, we observe that untrained intact models do not exhibit the sharp post-early-layer collapse in effective rank seen in trained intact models. Consistent with this, the mean token-wise cosine similarity in untrained intact models remains approximately constant across depth, in contrast to the progressive increase observed in trained models.

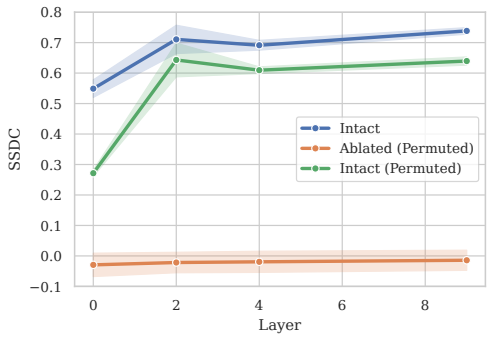
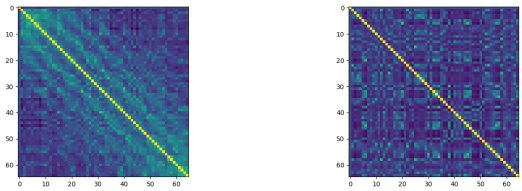


Figure 5. **SSDC across depth for intact models, and intact/ablated models that have had their tokens permuted randomly at inference.** The SSDC collapses to near-zero values upon permutation of the tokens of the ablated model, whereas the intact model's SSDC only takes a slight hit after permutation. Shaded regions indicate variability across runs (±1 standard deviation).



(a) Token Cosine Similarity Matrix for an intact model (permuted) at layer 2

(b) Token Cosine Similarity Matrix for an ablated model (permuted) at layer 2

Figure 6. **Representative Token Cosine Similarity Matrices.** Once its tokens are permuted, the ablated model's Token Cosine Similarity Matrix becomes chaotic and loses all spatial structure, whereas the diagonal band persists (though fuzzier) in the intact model's.

4.4. Patch-Relative and Absolute-Position-Based Modes of Spatial Organization

Experimental Setup: We extract Token Cosine Similarity Matrices from layers 0, 2, 4, and 9 from intact models, permuted intact models, and permuted ablated models. We plot these matrices and evaluate the SSDC across depth.

Notably, this permutation allows us to probe whether spatial organization is anchored to absolute position or emerges purely from patch content. Since token cosine similarity matrices are constructed according to token indices, this permutation disrupts the correspondence between matrix proximity and spatial proximity. Consequently, any diagonal structure observed after permutation cannot be attributed to local patch relationships and instead reflects the model's reliance on absolute positional information.

Results: We find that ablated models collapse to near-zero SSDC values after permutation (Fig. 5). Combined with the fact that these models' performance remains unchanged under permutation (Appendix Fig. A2), this indicates that ablated models form spatial structure primarily through patch content and relative relationships, without reliance on absolute token indices. As a result, permuting token order has no effect on which tokens become more similar as representations propagate through the network.

In contrast, intact models exhibit only a modest reduction in SSDC under permutation, suggesting that they rely more strongly on absolute token indices to organize spatial structure.

Interestingly, the SSDC of intact models under permutation drops sharply at layer 0 before gradually recovering across subsequent layers, remaining slightly below the unpermuted intact baseline. This pattern suggests that absolute positional information introduced by positional embeddings is progressively integrated within the encoder blocks, rather than being fully expressed at the input layer.

These trends are visually corroborated by the Token Cosine Similarity Matrices (Fig. 6). In the ablated models, permutation leads to a complete collapse of spatial structure, yielding matrices that appear random and unstructured. In contrast, intact models retain a fuzzy diagonal pattern after permutation, potentially suggesting that even intact models may partially rely on patch-content for spatial structure.

4.5. Robustness Is Tightly Linked to Spatial Encoding Strategy

Experimental Setup: We evaluate Fragility Scores as defined in section 3.3 across three distinct training and inference regimes designed to disentangle patch-content-based spatial organization from absolute-position-based strategies. The first regime is an intact model, trained and evaluated under standard conditions with positional embeddings (PEs) enabled. The second is an ablated model, in which positional embeddings are removed during and after training, serving as a reference point for models that lack explicit access to absolute position information. The third regime is an intact model trained with Random Permutation Training and evaluated with Random Permutation at Inference, hereafter referred to as RPT–RPI Intact.

The RPT–RPI Intact model preserves positional embeddings throughout training and inference, but is exposed to a different random permutation of patch tokens at every forward pass. As a result, any fixed mapping between token index and spatial location is systematically destroyed. This prevents the model from exploiting absolute positional cues, even though PEs are present in the architecture. Importantly, this regime does not collapse representational diversity: prior analyses show that RPT–RPI Intact models retain a large fraction of the effective rank and avoid the degeneracies observed in fully ablated models, although diversity

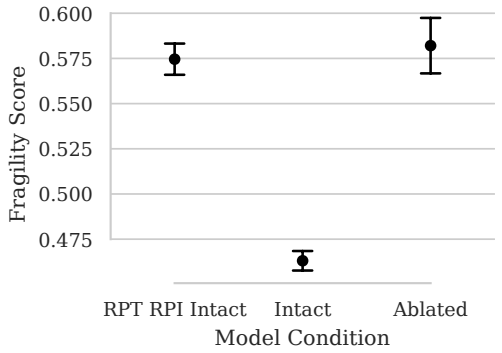


Figure 7. **Fragility scores for intact, ablated, and permutation-trained intact models under the stylized dataset.** Black markers show the mean Fragility Scores with ±1 standard deviation. Intact models are substantially more robust, while permutation-trained intact models exhibit fragility comparable to ablated models despite retaining positional embeddings.

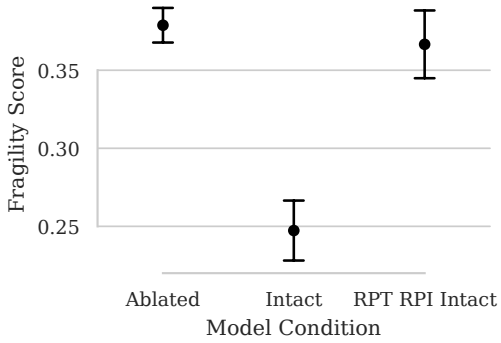


Figure 8. **Fragility scores for intact, ablated, and permutation-trained intact models exposed to Gaussian Blur.** Black markers show the mean Fragility Scores with ±1 standard deviation. Intact models are substantially more robust, while permutation-trained intact models exhibit fragility comparable to ablated models despite retaining positional embeddings.

is not perfectly preserved. This makes RPT–RPI Intact a controlled intervention that selectively disables absolute-position-based strategies without inducing the broader representational pathologies associated with PE removal.

Fragility Scores are computed identically across all regimes to ensure comparability. By contrasting the intact and ablated models with the RPT–RPI Intact model, we isolate whether robustness to spatial perturbations arises from spatial organization strategy or from the Representational Diversity introduced by the PEs. In particular, if a model trained under RPT–RPI conditions exhibits high fragility despite retaining Representational Diversity, this provides direct evidence that robustness can be explained by the differences in spatial organization strategies employed by these models. A dedicated sanity check is included to verify that RPT–RPI Intact models indeed rely on patch-content-driven cues for spatial structure, rather than implicitly recovering absolute position through degenerate shortcuts in the Appendix.

Results: We consistently find that the models that rely dominantly on patch content for spatial structure (ablated and RPT–RPI Intact models) are significantly more fragile to distributional shifts than the models that rely on an Absolute Position mode of spatial organization (Fig. 7). In particular, the RPT–RPI Intact models exhibit Fragility Scores that are substantially higher than standard intact models and nearly comparable to fully ablated models, despite retaining representational diversity. This indicates that robustness is not merely a consequence of diverse token representations, but critically depends on the presence of absolute-position-based strategies during training and inference.

Comparing intact and ablated models further clarifies the role of positional embeddings: intact models leverage PEs to construct stable spatial representations that mitigate sensitivity to random perturbations, while ablated models, lacking any absolute positional cues, show the lowest robustness. The RPT–RPI Intact regime provides a clean dissociation, showing that even when representational diversity is preserved, disruption of absolute positional mapping alone is sufficient to induce fragility.

Taken together, these results demonstrate that absolute-position-based spatial strategies, enabled by positional embeddings, are a key contributor to the stability of ViT representations under distributional shifts. Robustness cannot be explained solely by representational richness; the mode of spatial organization (absolute versus patch-relative) is a decisive factor. These findings complement our earlier analyses of representational geometry and spatial structure, linking internal organizational principles directly to functional resilience.

5. Discussion

Our results indicate that positional embeddings play a role that extends beyond injecting absolute positional information. Across models with intact positional embeddings, we consistently observe increased representational diversity in early layers, as reflected by elevated effective rank and reduced mean token-wise cosine similarity. This diversity persists weakly through depth and appears to act as a structural capacity rather than a direct determinant of performance. Notably, this capacity is necessary but not sufficient: RPT-RPI intact models retain high effective rank and low cosine similarity while exhibiting weak performance and low robustness (Appendix Fig. LALALALA). The causal role of positional embeddings is further supported by post-hoc ablation experiments, where removing positional embeddings at inference time induces an immediate collapse in effective

rank and increased token homogeneity, despite identical learned parameters.

A natural alternative explanation for these observations is that the increased effective rank and reduced cosine similarity associated with positional embeddings merely reflect unstructured decorrelation or noise. Several findings argue against this interpretation. Although untrained intact models also exhibit high effective rank, this diversity neither collapses across depth nor is accompanied by an increase in token-wise cosine similarity. In contrast, trained intact models display a coordinated pattern in which early representational diversity is progressively consolidated, with effective rank collapsing and token similarity increasing across layers. This behavior is inconsistent with stochastic noise, which would be expected to persist diffusely rather than undergo selective compression. Instead, these dynamics suggest that positional embeddings introduce a high-dimensional representational capacity that training actively organizes into structured, task-aligned subspaces.

Beyond representational geometry, positional embeddings also induce a qualitative shift in how spatial structure is formed. In the absence of positional embeddings, ViTs rely predominantly on patch content to infer relative spatial relationships, resulting in a content-based mode of spatial organization. When positional embeddings are present and combined with consistent patch ordering, models increasingly adopt an absolute-position-based strategy that leverages absolute positional information. While this aligns with common intuitions about the role of positional embeddings, our results indicate that this shift is not absolute: even intact models continue to subtly exploit patch content when forming spatial structure, suggesting that these strategies coexist rather than replace one another.

Finally, this shift in spatial organization has direct implications for robustness. Models that rely more heavily on absolute-position-based strategies are consistently more robust to distributional shifts. In contrast, content-based spatial organization is particularly vulnerable to perturbations that alter local patch statistics, which can disrupt the model's ability to infer coherent spatial relationships. This supports the interpretation that positional embeddings improve robustness not simply by increasing performance, but by biasing ViTs toward a mode of spatial reasoning that is less sensitive to changes in patch content.

Taken together, these findings suggest that positional embeddings act as a structural bias that reshapes the geometry of the representation space and alters the dominant strategy by which spatial information is encoded and used. Rather than serving as a passive positional signal, positional embeddings influence how diversity is introduced, organized, and ultimately integrated into task-relevant computation, with measurable consequences for robustness and generalization.

Limitations and Future Directions

Despite providing a detailed mechanistic analysis, our study has several limitations. First, our experiments focus on a specific class of ViT architectures and positional embedding schemes. While we expect the qualitative distinction between absolute-position-based and patch-relative strategies to generalize, the precise dynamics may vary across architectures, embedding types, or training regimes. Extending this analysis to alternative positional encoding mechanisms, such as relative or rotary embeddings, remains an important direction for future work.

Second, our robustness evaluation centers on a particular family of distributional shifts. Although these shifts are well-motivated and commonly used, they do not exhaust the space of possible perturbations. It is possible that patch-relative strategies may confer advantages under other forms of shift not considered here.

Finally, our study is limited in its experimental scope: all analyses were conducted using relatively small Vision Transformers (12M parameters) trained on CIFAR-10. While this controlled setting enabled detailed mechanistic analysis, it raises natural questions about the generality of our findings. The magnitude of the effects we observe, such as the scale of the early-layer effectve rank spike or the precise SSDC values may vary with model scale and dataset complexity. However, we hypothesize that the core directional conclusions are rooted in a causal, architectural mechanism. But we explicitly encourage future works to validate and quantify these effects in larger-scale, state-of-the-art ViT models.

6. Related Work

Positional Information in Vision Transformers

The standard Vision Transformer (ViT) breaks the permutation invariance of self-attention by adding learnable absolute positional embeddings (PEs) to patch tokens (Dosovitskiy et al., 2021), establishing the dominant paradigm for spatial encoding in vision transformers. However, it has been found that ViTs retain substantial performance even when PEs are degraded or removed (Dosovitskiy et al., 2021; Chu et al., 2023), suggesting ViTs may be able to partially recover spatial structure and implicit positional information without PEs. In fact, similar observations have been reported outside the vision domain. For example, recent work on decoder-only Transformers shows that models trained without PEs can implicitly recover positional information and that they tend to use relative positions in practice (Kazemnejad et al., 2023). While this analysis is specific to generative language models, it reinforces the broader notion that explicit PEs are not strictly required for structured positional information to emerge. This parallels earlier findings in CNNs, where

it was demonstrated that convolutional networks learn substantial positional information implicitly, for instance from architectural features like zero-padding (Islam* et al., 2020). These observations create the fundamental puzzle our work addresses: if spatial structure can emerge without explicit guidance, what functional role do PEs actually play? Prior studies on PEs in ViTs have primarily focused on architectural variants (d'Ascoli et al., 2022; Liu et al., 2021) or downstream performance comparisons [2, 3], leaving the mechanistic impact of PEs on internal representations largely unexplored.

Representational Analysis of Transformers

A separate line of work analyzes the geometry and dynamics of transformer representations using tools from representational analysis. Earlier work provided foundational comparisons between ViT and CNN representations (Raghu et al., 2021), revealing distinct spatial organization patterns. Subsequent work has examined how attention mechanisms transform representations(Kobayashi et al., 2021), and the evolution of representational rank through depth (Dong et al., 2021), and the tendency for token representations to homogenize in deep layers (Bhojanapalli et al., 2021). The residual stream framework provides the conceptual foundation for our analysis of token representations across layers (Elhage et al., 2021). However, despite these insights, these representational analyses have not specifically targeted the causal effect of positional embeddings on this geometry, nor have they connected these internal dynamics to external robustness properties.

Robustness of Visual Models

Vision Transformers exhibit distinct robustness profiles compared to convolutional networks. Prior works have systematically compared ViT and CNN robustness, finding transformers exhibit greater resilience to spatial perturbations but increased sensitivity to certain texture changes (Bhojanapalli et al., 2021). Subsequent work further establishes that ViTs demonstrate favorable out-of-distribution generalization properties (Paul & Chen, 2022). These observations connect to the broader literature on shape versus texture bias in visual recognition, where it has been shown that models with stronger shape bias tend to exhibit better generalization (Geirhos et al., 2019). While robustness differences between architectural families have been documented, the link between a model's specific spatial reasoning strategy, such as relying on absolute position versus inferring relations from content, and its robustness to distribution shifts has not been mechanistically established.

Our Contribution

We bridge these disconnected research threads to solve the positional embedding puzzle. Unlike performance-focused ablation studies (Dosovitskiy et al., 2021; Chu et al., 2023), we perform a causal, representational analysis to show how PEs work internally. We demonstrate they (1) induce early-layer representational diversity through a high-rank injection at layer 0, (2) shift the model's spatial organization strategy from content-relative to position-absolute reasoning, and (3) that this strategic shift—rather than representational diversity alone—confers robustness to texture-based distribution shifts. We introduce the Spatial Similarity Distance Correlation (SSDC) metric and employ controlled interventions like Random Permutation Training to isolate these mechanisms, providing the a unified and mechanistic account of why PEs remain crucial beyond their basic function of breaking permutation invariance.

7. Conclusion

In this work, we investigated how positional embeddings shape spatial organization and representation geometry in Vision Transformers. We show that even in the absence of positional embeddings, ViTs retain non-trivial spatial structure through a patch-relative, content-based mode of organization. However, this structure is fragile and corresponds to limited representational consolidation. When positional embeddings are present, they introduce substantial early-layer representational diversity and induce a qualitatively different pattern of representation dynamics, characterized by coordinated rank collapse and increased token similarity across depth.

These findings indicate that positional embeddings act as a structural bias that alters both the geometry of the latent space and the dominant strategy by which spatial information is encoded. By pushing ViTs toward an absolute-position mode of spatial organization, positional embeddings promote more stable and robust representations under distributional shift, while still allowing patch content to play a secondary role. More broadly, our results suggest that spatial reasoning in ViTs emerges from the interaction between inductive bias, training dynamics, and representational geometry, rather than from positional encoding alone. Understanding and controlling these interactions may be critical for designing transformer-based vision models that generalize reliably beyond their training distributions.

References

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., and Veit, A. Understanding robustness of transformers for image classification. pp. 10211–10221, 10 2021. doi: 10.1109/ICCV48922.2021.01007.

Chu, X., Tian, Z., Zhang, B., Wang, X., and Shen, C. Conditional positional encodings for vision transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/

forum?id=3KWnuT-R1bh.

Dong, Y., Cordonnier, J.-B., and Loukas, A. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/dong21a.html.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

d'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., and Sagun, L. Convit: improving vision transformers with soft convolutional inductive biases*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114005, nov 2022. doi: 10.1088/1742-5468/ac9830. URL https://doi.org/10.1088/1742-5468/ac9830.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Islam*, M. A., Jia*, S., and Bruce, N. D. B. How much position information do convolutional neural networks encode? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJeB36NKvB.

Kazemnejad, A., Padhi, I., Natesan, K., Das, P., and Reddy, S. The impact of positional encoding on length generalization in transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Drrl2gcjzl.

Kobayashi, G., Kuribayashi, T., Yokoi, S., and Inui, K. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4547–4568. Association for Computational Linguistics, November 2021. doi: 10.18653/v1/2021.emnlp-main.373. URL https://aclanthology.org/2021.emnlp-main.373/.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002. IEEE Computer Society, October 2021. doi: 10.1109/ICCV48922.2021.00986. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986.

Paul, S. and Chen, P.-Y. Vision transformers are robust learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):2071–2081, Jun. 2022. doi: 10.1609/aaai.v36i2.20103. URL https://ojs.aaai.org/index.php/AAAI/article/view/20103.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do vision transformers see like convolutional neural networks? In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=R-616EWWKF5.

A. Experimental Setup and Hyperparameters

Table 1. Model architecture and training hyperparameters used in all experiments.

Parameter	Value
Input & Tokenization	
Input resolution	32×32
Patch size	4×4
Number of patches	64
Input channels (C)	3
ViT Architecture	
Embedding dimension (D)	320
Number of encoder layers	10
Number of attention heads	8
Key/query dimension (d_k)	40
Dropout (embedding)	0.1
Dropout (attention)	0.1
Dropout (MLP)	0.1
Stochastic depth rate	0.1
Training Hyperparameters	
Batch size	128
Optimizer	Adam
Learning rate	1×10^{-3}
Weight decay	5×10^{-4}
Adam β_1	0.9
Adam β_2	0.999
Training epochs	50

B. Metric Definitions and Implementation Details

B.1. Spatial Similarity Distance Correlation (SSDC)

Let T denote the number of patch tokens (excluding the CLS token), arranged on a $\sqrt{T} \times \sqrt{T}$ image grid. For a given layer, let $S \in \mathbb{R}^{T \times T}$ be the pairwise cosine similarity matrix between token representations. We associate each token i with spatial coordinates $\mathbf{p}_i \in \mathbb{Z}^2$ corresponding to its location on the image grid, and define the spatial distance matrix $D \in \mathbb{R}^{T \times T}$ by

$$D_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|_1,$$

where $\|\cdot\|_1$ denotes Manhattan distance.

SSDC is defined as the Spearman rank correlation between representational similarity and negative spatial distance over all unordered token pairs:

$$\text{SSDC} = \rho_{\text{Spearman}}(\{S_{ij}\}_{i<j}, \{-D_{ij}\}_{i<j}) \,.$$

Higher SSDC values indicate that spatially proximal tokens tend to have more similar representations, reflecting stronger relative positional structure in the residual stream. We use Spearman correlation to remain agnostic to the exact functional relationship between spatial distance and representational similarity.

Effective Rank

Given a matrix $X \in \mathbb{R}^{n \times d}$ (e.g., a collection of token representations), let $\sigma_1, \ldots, \sigma_r$ denote its singular values, where $r = \text{rank}(X)$. We define the effective rank of X using the participation ratio as

$$\text{erank}(X) = \frac{\left(\sum_{i=1}^{r} \sigma_i^2\right)^2}{\sum_{i=1}^{r} \sigma_i^4}.$$

This quantity measures how evenly variance is distributed across singular directions: it attains its maximum value of r when all singular values are equal, and decreases as the representation collapses onto fewer dominant directions.

B.2. Mean Cosine Similarity

Let $R \in \mathbb{R}^{B\times T\times D}$ denote the token representations at a given layer, excluding the `[CLS]` token, where B is the batch size, T the number of tokens, and D the hidden dimension. For each sample in the batch, we compute the pairwise cosine similarity between all token representations to obtain a token cosine similarity matrix $S \in \mathbb{R}^{T\times T}$, where

$$S_{ij} = \frac{r_i^\top r_j}{\|r_i\|_2 \, \|r_j\|_2}.$$

The similarity matrices are symmetric by construction and have unit diagonal. We average S across the batch dimension to obtain a layer-wise summary of inter-token relationships. Mean Cosine Similarity (MCS) is defined as the mean of all entries of this averaged similarity matrix. Because the diagonal entries are identically equal to 1 across all models and conditions, including them does not affect comparative analysis. Higher MCS values indicate greater representational homogeneity among tokens.

B.3. Fragility Score

To quantify sensitivity to distributional shifts, we compute the Fragility Score (FS), defined as the relative drop in top-1 accuracy under a given shift:

$$\mathrm{FS} = 1 - \frac{A_{\text{shift}}}{A_{\text{normal}}},$$

where A_{normal} and A_{shift} denote top-1 accuracy on the unshifted and shifted datasets, respectively. Higher values correspond to greater performance degradation. Unless otherwise stated, accuracies are averaged across random seeds prior to computing FS.

B.4. Metric Summary

For clarity, we briefly summarize the metrics used throughout this work and the phenomena they are intended to capture.

- **Mean Cosine Similarity (MCS)** measures average pairwise similarity between token representations and serves as a proxy for inter-token homogeneity or redundancy.
- **Effective Rank** quantifies representational diversity by measuring how evenly variance is distributed across singular directions of the token representation matrix.
- **Spatial Similarity Distance Correlation (SSDC)** captures the degree to which representational similarity aligns with spatial proximity, serving as a proxy for relative positional structure in the residual stream.
- **Fragility Score (FS)** measures sensitivity to distributional shifts by quantifying relative performance degradation under dataset perturbations.

Together, these metrics allow us to disentangle representational diversity, spatial organization strategy, and robustness, and to analyze how positional embeddings causally affect each of these factors.