What Positional Embeddings Really Do In Vision Transformers

Mahmoud Mannes [1]

Abstract

Positional embeddings (PEs) in Vision Transformers (ViTs) are often viewed as simple spatial injectors that enable the model to encode absolute position. In this work, we show that even ViTs trained without PEs can recover meaningful spatial structure and distinguish relative position using patch content alone. This raises a central question: if spatial structure can be reconstructed without PEs, what functional role do PEs actually play? We find that PEs causally induce a dramatic increase in early-layer residual stream dimensionality and token heterogeneity, leading to richer image representations. PEs also introduce spatial structure prior to the encoder blocks and encourage representations that jointly rely on positional information and patch content. In contrast, models without PEs rely exclusively on content-based heuristics to infer spatial structure, a strategy that is significantly more fragile to distributional shifts.

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., 2020). Unlike Convolutional Neural Networks (CNNs), however, ViTs lack strong built-in inductive biases toward locality and translation equivariance. As a result, most ViT architectures rely on Positional Embeddings (PEs) to inject explicit spatial information, enabling the model to distinguish between patches originating from different locations in the image.

Despite their widespread use, the precise role of positional embeddings in vision transformers remains incompletely understood. While PEs are generally assumed to be essential for encoding spatial structure, several empirical studies have shown that ViTs can retain non-trivial performance even when explicit positional information is removed or degraded (Chu et al., 2021). These findings suggest that transformers may partially reconstruct spatial relationships from patch content alone, much like how CNNS learn implicit positional information from zero-padding (Islam et al., 2020), raising questions about what functional advantages positional embeddings actually provide beyond basic spatial identifiability.

Most prior work has investigated the role of positional embeddings primarily through downstream performance metrics or architectural variations. While informative, such analyses offer limited insight into how positional information influences the internal representations learned by the model. In particular, the impact of positional embeddings on the geometry, dimensionality, and stability of token representations throughout the transformer stack has received comparatively little attention.

In this work, we adopt a mechanistic perspective to study the role of positional embeddings in Vision Transformers. We analyze the evolution of token representations in the residual stream using tools from representational geometry (Raghu et al., 2021), with the goal of characterizing how spatial structure emerges and is maintained across layers. To support this analysis, we introduce the Spatial Structure–Diagonal Coefficient (SSDC), a metric designed to quantify the degree to which spatial relationships are reflected in token similarity patterns.

Using this framework, we conduct a comparative study of ViT models trained with and without positional embeddings. We examine how positional embeddings affect representational dimensionality, the nature of spatial reasoning employed by the model, and robustness to distributional shifts such as stylization and noise. Together, our results provide a more detailed picture of how explicit positional signals shape the internal organization of vision transformers, offering new insight into why positional embeddings play a critical role in stable and robust visual representation learning.

---
[1] Independent Researcher. Correspondence to: Mahmoud Mannes <mannesmahmoud@gmail.com>.

2. Background and Setup

2.1. Vision Transformer Architecture

All models used in our experiments are vanilla Vision Transformers trained from scratch, with approximately 1.2M 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

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

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

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.
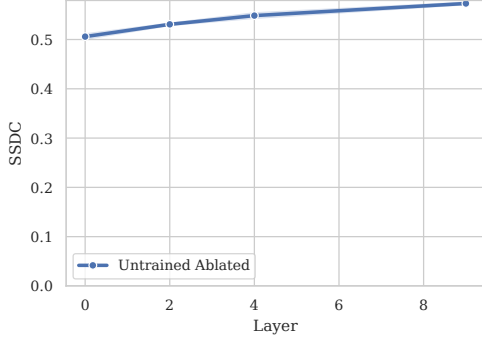
Figure 1. **SSDC across depth for an untrained ablated model.** SSDC remains approximately constant across layers, indicating static spatial correlations induced by architectural and data priors rather than learning. Shaded regions indicate variability across runs.

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 proposed SSDC metric. We compare three settings: (i) an untrained 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 model with random permutation exhibits a consistently negative 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 encodings, but can emerge implicitly through training dynamics and architectural biases.

Token Cosine Similarity Matrices (Fig. 4) provide a visual confirmation of these trends. The untrained model shows no coherent spatial pattern, with diffuse similarity and no clear evolution across layers. In contrast, the trained ablated model exhibits a progressively sharpening diagonal structure from layer 0 to layer 2, indicating increasing similarity between nearby tokens and the formation of a locality bias. This diagonal becomes more pronounced with depth, aligning with the observed SSDC growth.
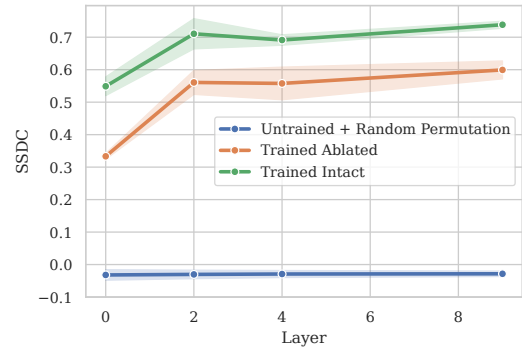

Figure 2. **SSDC across depth for an untrained model with random permutation, a trained ablated model, and a trained intact model.** Shaded regions indicate variability across runs.
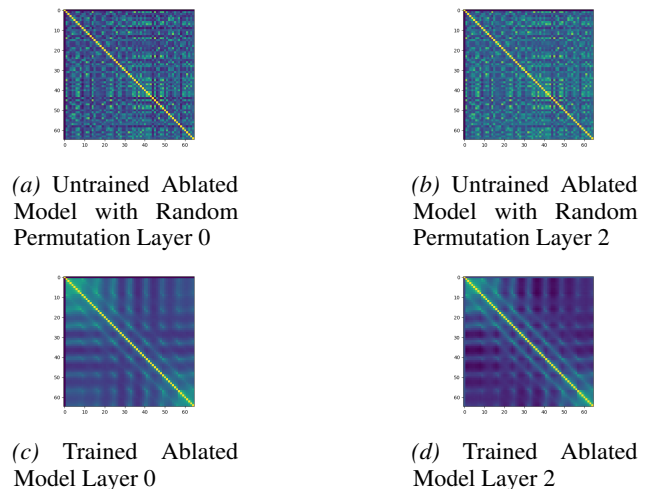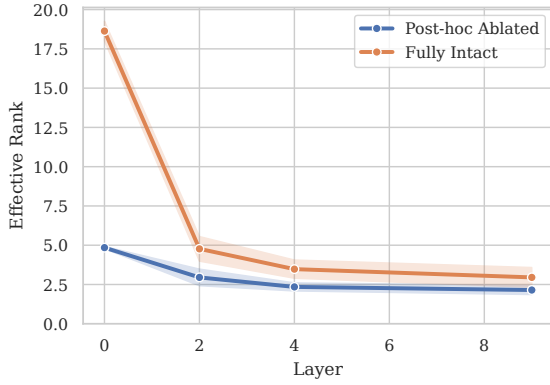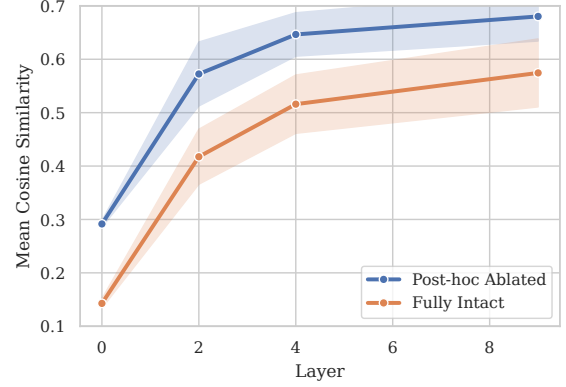

(a) Untrained Ablated Model with Random Permutation Layer 0

(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)

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

Figure 4. Early-layer representational diversity under intact and post-hoc ablated models. Left: Effective Rank. Right: Mean Token Cosine Similarity.

4.3. Positional Embeddings Inject Early-Layer Representational Diversity

Experimental setup: We evaluate the Effective Rank of the Residual Stream Matrix at layers 0, 2, 4, and 9 on the CIFAR-10 dataset using ablated and intact models. We also extract the Token Cosine Similarity Matrix at these same layers from the same models and calculate the mean value of the Token Cosine Similarities. To deduce the causal nature of the effects of PEs on these metrics, we repeat these processes on an intact model that has its PEs removed post-hoc at inference.

Results:

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

References

Chu, Zhi, et al. Conditional positional encodings for vision transformers. 2021.

Dosovitskiy, Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

Islam, Jia, et al. How much positional information do convolutional neural networks encode? 2020.

Raghu, Dosovitskiy, et al. Do vision transformers see like convolutional neural networks? 2021.