From Local to Contextually-Enriched Local Representations: A Mechanism for Holistic Processing in DINOv2 ViTs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Self-supervised Vision Transformers (ViTs) such as DINOv2 achieve robust holistic shape processing, but the transformations that support this ability remain unclear. Probing with visual anagrams, we find that DINOv2's intermediate layers constitute a necessary stage for holistic vision. Our analyses reveal a structured sequence of computations. First, local content representations remain spatially anchored deeper into the network than in supervised ViTs. Second, attention heads progressively extend their range, producing a systematic local-to-global transition, facilitating local representations that are contextually enriched. Third, positional signals are not merely lost with depth but become more sharply aligned with the model's learned positional embeddings in mid-level layers. Models without these properties, such as supervised ViTs, rapidly lose spatially specific content and fail on holistic tasks. Finally, when register tokens are present, high-norm global activations are redirected into these tokens rather than overwriting low-information patch embeddings, allowing patches to maintain their positional identity, also leading to improvements on holistic tasks. Together, these findings show that holistic vision in ViTs emerges from a structured progression of representational transformations that preserve both content and spatial information while enabling global integration.

1 Introduction

Vision Transformers (ViTs) [1] have demonstrated a range of impressive capabilities that distinguish them from their convolutional counterparts. Studies have shown they have a stronger shape bias, greater resilience to severe occlusions and a more flexible attention mechanism [2, 3, 4, 5]. However, while these properties are broadly true, a significant performance gap emerges on tasks requiring holistic processing [6, 7, 8, 9]. Recent work [7] using visual anagrams [10], images that depict different object categories but are built from the same set of local patches just merely rearranged (Fig. 1), revealed that strong holistic understanding is not a universal feature of ViTs but is particularly emergent in self-supervised models like DINOv2 [11]. This capability dissociated from object recognition and the classic shape-vs-texture bias, with the critical computations found in the intermediate transformer blocks of these ViTs.

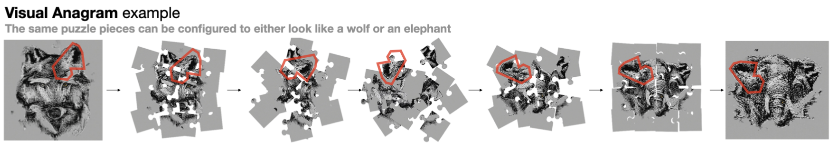
Visual Anagram example
The same puzzle pieces can be configured to either look like a wolf or an elephant

Figure 1: An example of a visual anagram pair. The same set of local image patches can be rearranged to form images of a wolf and an elephant.

This finding presents a puzzle when viewed through the lens of prior mechanistic work [12, 13] which concluded that intermediate layers of supervised ViTs were surprisingly redundant: their

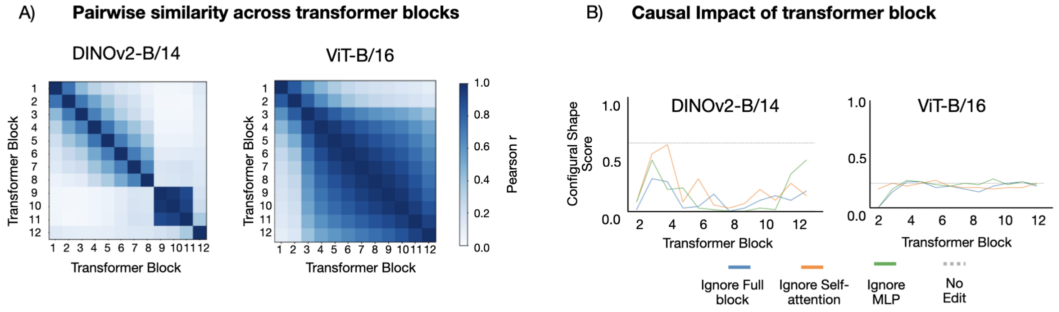

Figure 2: (A) Comparing representations of anagram images between transformer blocks. (B) Ablating intermediate transformer blocks and measuring impact on Configural Shape Scores.

representations remained highly correlated through the blocks, and removing intermediate blocks had minimal effect on recognition performance. Nonetheless, other work has shown that, ViTs in principle, can still learn structured algorithms: when fine-tuned on specific relational tasks, they learned distinct "perceptual" and "relational" processing stages [5, 14].

This leads to our central question: what computations in DINOv2 give rise to holistic vision? We address this through mechanistic analyses of self-supervised ViTs, with three main findings:

- **Intermediate layers are indispensable.** Unlike supervised ViTs, whose mid-layers are functionally redundant, ablating DINOv2's intermediate blocks sharply reduces holistic performance, indicating that these stages implement non-trivial transformations.
- **Attention shifts from local to global.** DINOv2 maintains short-range attention deeper into the network than supervised ViTs, before gradually extending to long-range links. This staged transition preserves local content while enabling global integration.
- **Positional signals become more distinct mid-network.** Rather than fading with depth, positional information becomes more separable (quasi-orthogonal) from content in the intermediate layers. With register tokens present, global activations are absorbed by registers instead of overwriting patch embeddings, allowing patches to maintain their positional identity.

Together, these results indicate that DINOv2 forms holistic percepts through a structured progression: early layers preserve local content, intermediate layers enrich that content with contextual information and make positional identity more distinct, and later layers assemble global representations.

2 Background: Probing Holistic Vision with Visual Anagrams

To quantitatively measure a model's capacity for holistic processing, recent work introduced the Configural Shape Score (CSS) using the "visual anagram" stimulus set [7]. Visual anagrams are pairs of images that depict two different object categories (e.g., a wolf vs. an elephant) but are constructed from the exact same set of local image patches, merely rearranged [10]. Because local textural cues are matched between the two images in the anagram pair, to correctly classify both images in a pair, a model can only succeed by processing the configural arrangement of the patches. While standard supervised ViTs performed poorly, self-supervised models like DINOv2 showed high CSS scores. Further representational similarity analysis (RSA) using carefully controlled image pairs revealed a distinct transition within the intermediate layers. Early-layer representations were dominated by local puzzle piece similarity, whereas later layer representations were dominated by the global object category, suggesting that a critical transition from part-based to category-based representations occurs in the intermediate transformer blocks.

3 Intermediate Layers in DINOv2 are Causally Critical for Holistic Processing

While prior work has established that the intermediate layers of supervised ViTs are functionally redundant [12, 13], we first investigated if the same holds for the DINOv2 which is one of the highest performing model class on CSS. We ran two analyses: first, we traced the flow of representations through the network, and second, we performed a causal ablation study to measure each layer's direct contribution to the CSS.

Following the approach by [13], we extracted the activations of all anagram images in each transformer block and then measured similarity between all block pairs (Fig. 2A). For a standard supervised ViT-

B/16, the off-diagonal similarities remained high after the initial layers, keeping the representations largely similar. This confirmed the redundancy hypothesis for standard ViTs. In DINOv2-B/14, however, the similarity matrix showed a sharp representational change in the intermediate layers (approx. blocks 8-9). To further test if this drastic change in representation was causally critical, we performed an ablation study where we bypassed each transformer block individually during inference [12] and measured the downstream impact on CSS (Fig. 2B). For the supervised ViT-B/16, ablating any intermediate block had negligible effect on CSS but for DINOv2-B/14, the same intervention, especially in blocks 4-10, collapsed the CSS performance to near chance. These results provide evidence that the intermediate layers in DINOv2-B/14 are causally critical for holistic processing.

4 Dissecting the Intermediate Computations

"What" and "Where" Does a Head Attend? Having established that DINOv2's intermediate layers are necessary, we ask what computations they implement. We separate the analysis into "where" and "what". **Where.** For each head in layer (i.e. the transformer block) ℓ, we compute the Mean Attention Head Distance (MAHD) following [13]: the attention-weighted Euclidean distance (in pixel coordinates) between a query patch and all key patches, averaged over N=3925 Imagenette images (CLS excluded). Within each layer, heads are sorted by MAHD and visualized as a heatmap to reveal the depth-wise transition from short- to long-range links. **What.** A short-range MAHD is only meaningfully "local" if the attended patch still carries local content. To investigate this, we separate content from global and positional structure using a principled decomposition. Let $\boldsymbol{x} \in \mathcal{D}$ denote an input image with patch set $\Omega = \{1, \ldots, W\} \times \{1, \ldots, H\}$, $\boldsymbol{h}_p^{(\ell)}(\boldsymbol{x}) \in \mathbb{R}^d$ the layer-ℓ embedding of patch p and $z_q^{(0)}(x)$ be the input (layer-0) patch embedding at position q with $p, q \in \Omega$. We now formally introduce our decomposition method:

Definition 1 (Vision Transformer Spatial-Content Decomposition)**.** *For any vision transformer layer ℓ and dataset $\mathcal{D}$, every patch embedding admits the unique additive decomposition:*

$$\boldsymbol{h}_p^{(\ell)}(\boldsymbol{x}) = \boldsymbol{\mu}^{(\ell)} + \boldsymbol{\mu}_p^{(\ell)} + \boldsymbol{c}_p^{(\ell)}(\boldsymbol{x})$$

where:

$$\begin{cases} \boldsymbol{\mu}^{(\ell)} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},p\sim\Omega}(\boldsymbol{h}_p^{(\ell)}(\boldsymbol{x})) & \textit{(global mean)} \\ \boldsymbol{\mu}_p^{(\ell)} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}(\boldsymbol{h}_p^{(\ell)}(\boldsymbol{x})) & \textit{(positional effect)} \\ \boldsymbol{c}_p^{(\ell)}(\boldsymbol{x}) = \boldsymbol{h}_p^{(\ell)}(\boldsymbol{x}) - \boldsymbol{\mu}^{(\ell)} - \boldsymbol{\mu}_p^{(\ell)} & \textit{(content residual)} \end{cases}$$

The content residual $\boldsymbol{c}_p^{(\ell)}(\boldsymbol{x})$ isolates input-specific semantic information and is statistically orthogonal[1] *to both global and positional components by construction.*

To measure spatial localization of content, we define the localization score between patch p at layer ℓ and initial position q:

$$s_{p,q}^{(\ell)}(\boldsymbol{x}) = \frac{\langle \boldsymbol{c}_p^{(\ell)}(\boldsymbol{x}),\, \boldsymbol{z}_q^{(0)}(\boldsymbol{x}) \rangle}{\|\boldsymbol{c}_p^{(\ell)}(\boldsymbol{x})\|\, \|\boldsymbol{z}_q^{(0)}(\boldsymbol{x})\|}, \quad q \in \Omega \tag{1}$$

The field $q \mapsto s_{p,q}^{(\ell)}(\boldsymbol{x})$ measures spatial localization: higher concentration near q=p indicates that content at patch p remains tied to its original spatial coordinate through layer ℓ.

Applying our decomposition to both supervised ViT-B/16 and DINOv2-B/14 reveals striking differences in their computational strategies (Fig. 3). In the supervised ViT, MAHD increases rapidly in early layers: most attention heads transition to long-range interactions (MAHD > 100 pixels) by layer 3. In contrast, DINOv2 maintains a heterogeneous mixture of short-range (MAHD < 50 pixels) and long-range heads throughout the first half of the network, with the transition occurring gradually between layers 4-8 (see Appendix A.1 for MAHD in DINOv2 S/L/G variants). Content localization exhibits a corresponding pattern. In the supervised model, the content residual $\boldsymbol{c}_p^{(\ell)}(\boldsymbol{x})$ loses its spatial specificity early: by layer 3, the localization score $s_{p,q}^{(\ell)}(\boldsymbol{x})$ becomes uniformly distributed across the spatial grid, meaning patches no longer carry information specific to their original location. DINOv2 preserves this spatial specificity significantly longer, with concentrated localization scores maintained through layer 6-7. Moreover, our spatial-content decomposition reveals another distinction. While

[1] It has zero expectation covariance with them, in expectation over data and patches.

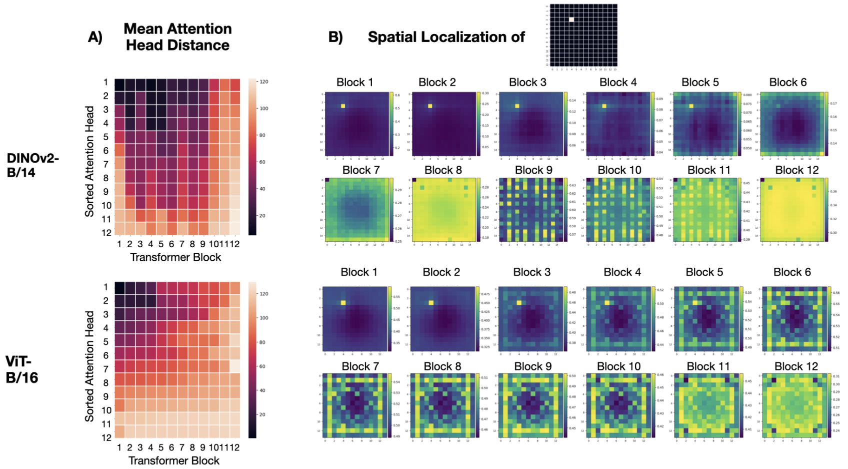

Figure 3: (A) Mean distance in each attention head for all layers of the ViT. (B) Cosine similarity between the content residual representation of the selected patch and the input patch embedding.

analyses of the full patch representations $\boldsymbol{h}_p^{(\ell)}(\boldsymbol{x})$ show apparent spatial structure in both models (Appendix A.2), our decomposition shows that in supervised ViTs, this structure comes primarily from the positional component $\boldsymbol{\mu}_p^{(\ell)}$ rather than content. In other words, patches "remember" where they are, but not what local visual information they originally contained. Only DINOv2 maintains both positional and content-specific spatial information $\boldsymbol{c}_p^{(\ell)}(\boldsymbol{x})$ deep into the network. This dual preservation likely enables DINOv2's superior holistic processing: patches retain knowledge of both their spatial location and their local visual content, allowing the model to reason about how local parts relate to global structure.

Alignment with Learned Positional Embeddings. Beyond keeping content local, DINOv2 also showed another interesting signature in the encoded positional signal ($\boldsymbol{\mu}_p^{(\ell)}$). We estimated each patch's positional signal in a specific layer (Def. (1)), computed its cosine similarity with that patch's learned positional embedding at the start of the network, and averaged over patches to obtain a layer-wise positional similarity curve (Fig. 4). In the supervised ViT, positional similarity drops steadily from the first block. In DINOv2 models, it rises through the early blocks, peaks mid-network (blocks 4–10), and then drops in the last blocks. In other words, DINOv2 sharpens a patch's location information up until the middle of the network, keeping the "where" information crisp while it gathers context, whereas the supervised model steadily forgets this location information. This increase in alignment also relates to "attention-sinks" and high-norm "outlier" tokens reported in large ViTs [15, 16, 17, 18, 19] which occurs when low-information patches are repurposed to carry global signals, subsequently degrading the local positional information that they encode. Models trained with registers [15], that likely absorb these high-norm activations, showed higher CSS and maintained the positional alignment even further into the network (AppendixA.3).

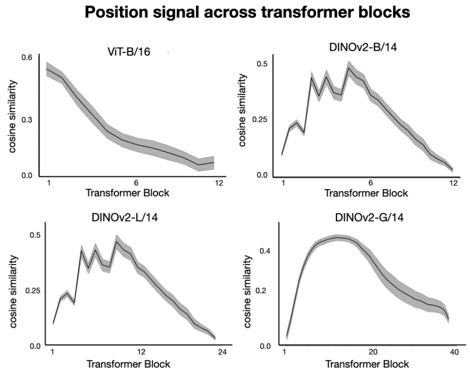

Figure 4: Mean cosine similarity of estimated positional signal with learned positional embeddings.

5 Conclusion

Overall, we asked how self-supervised DINOv2 ViTs arrive at an object's global shape and characterized the computations underlying holistic processing. We found that mid-layers matter: ablating these layers sharply lowers CSS in DINOv2, whereas the same manipulation in a supervised ViT has little impact; representation structure also changes drastically in these layers. We observed three key signatures in these intermediate stages of processing: (i) patch content stays local much deeper than in the supervised ViT, (ii) a large mixture of heads that maintain short-range attention while also drawing from distant patches, enabling a contextually-enriched local representation, and (iii) an increase in alignment to the learned positional embeddings in the intermediate layers.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[4] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. *Advances in Neural Information Processing Systems*, 35:16276–16289, 2022.

[5] Michael Lepori, Alexa Tartaglini, Wai Keen Vong, Thomas Serre, Brenden M Lake, and Ellie Pavlick. Beyond the doors of perception: Vision transformers represent relations between objects. *Advances in Neural Information Processing Systems*, 37:131503–131544, 2024.

[6] Nicholas Baker and James H Elder. Deep learning models fail to capture the configural nature of human shape perception. *Iscience*, 25(9), 2022.

[7] Fenil R Doshi, Thomas Fel, Talia Konkle, and George Alvarez. Visual anagrams reveal hidden differences in holistic shape processing across vision models. *arXiv preprint arXiv:2507.00493*, 2025.

[8] Fenil R Doshi, Talia Konkle, and George A Alvarez. Quantifying the quality of shape and texture representations in deep neural network models. *Journal of Vision*, 24(10):1263–1263, 2024.

[9] Valerio Biscione and Jeffrey S Bowers. Mixed evidence for gestalt grouping in deep neural networks. *Computational Brain & Behavior*, 6(3):438–456, 2023.

[10] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24154–24163, 2024.

[11] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[12] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021.

[13] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

[14] Hojin Jang, Pawan Sinha, and Xavier Boix. Configural processing as an optimized strategy for robust object recognition in neural networks. *Communications Biology*, 8(1):386, 2025.

[15] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

[16] Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. *arXiv preprint arXiv:2506.08010*, 2025.

[17] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[18] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024.

[19] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

A Appendix

A.1 Mean Attention Head Distance for DINOv2 ViTs

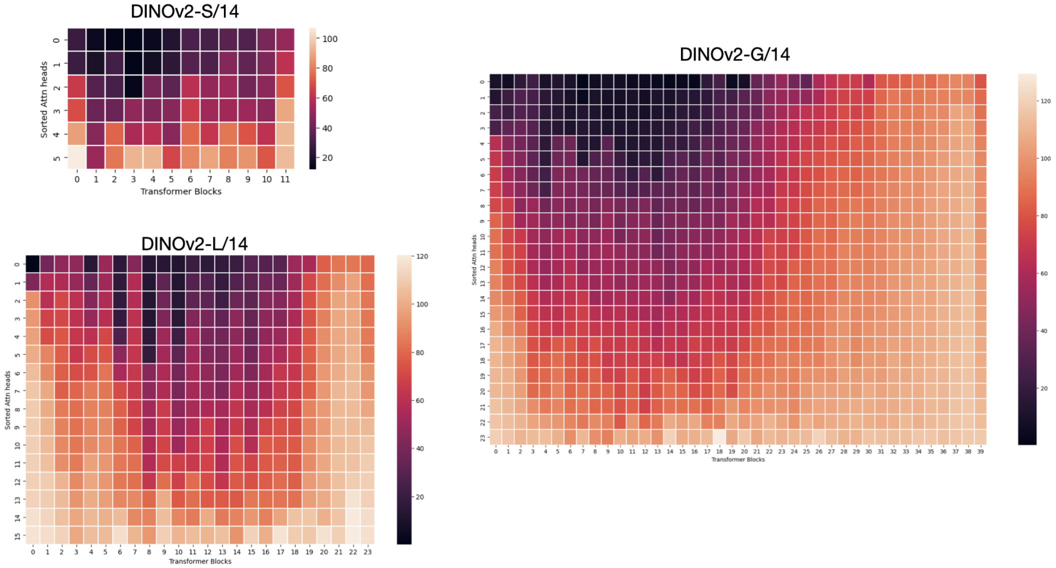

Figure 5: Mean distance in each attention head for all layers of Dinov2-S/14, Dinov2-L/14, and Dinov2-G/14..

A.2 Spatial Localization of content residual vs. full patch representations

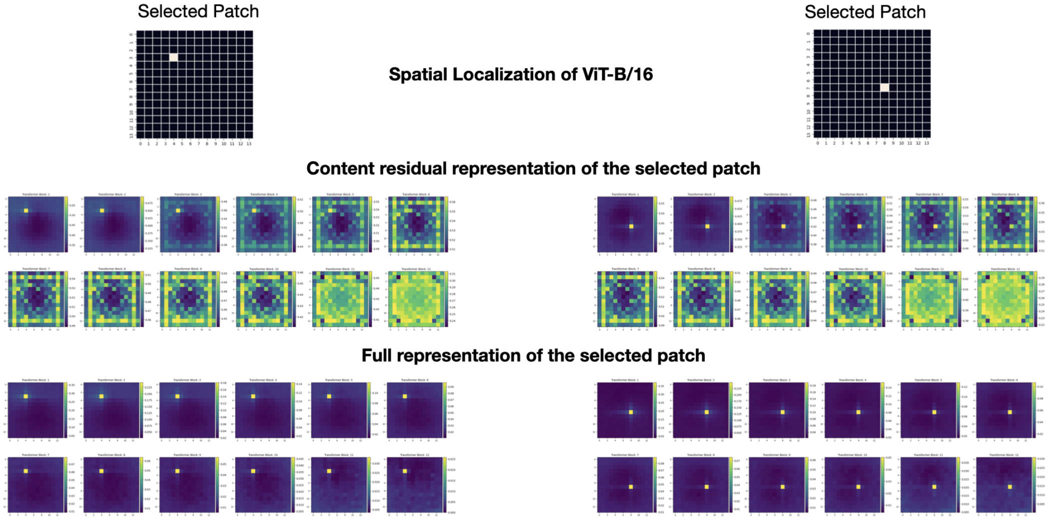

Figure 6: (Top) 2 selected patches in ViT-B/16. (Middle) Cosine similarity between the content residual representation of the selected patches and the input patch embedding.(Bottom) Cosine similarity between the full patch representation and the input patch embedding.

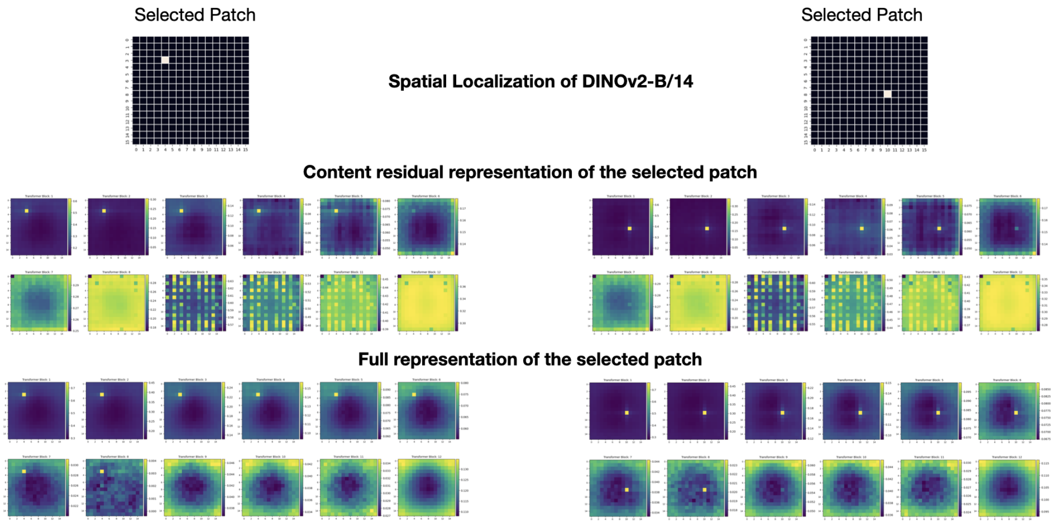

Figure 7: (Top) 2 selected patches in DINOv2-B/14. (Middle) Cosine similarity between the content residual representation of the selected patches and the input patch embedding.(Bottom) Cosine similarity between the full patch representation and the input patch embedding.

A.3 Alignment with Learned Positional Embeddings in model with vs. without trained registers

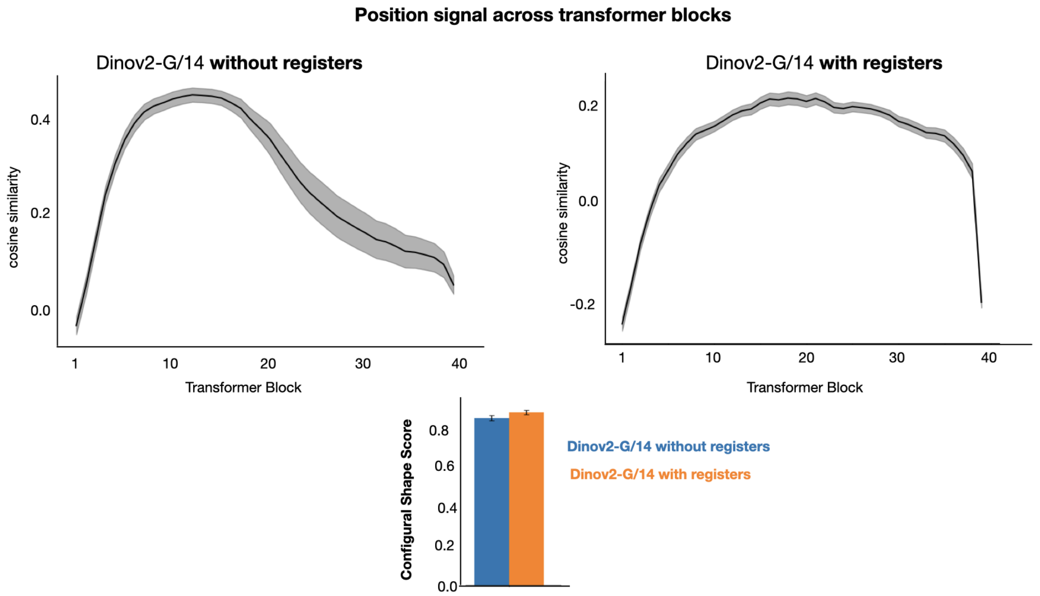

Figure 8