I Walk the Line: Examining the Role of Gestalt Continuity in Object Binding for Vision Transformers

Alexa R. Tartaglini
Stanford University
alexart@stanford.edu

Michael A. Lepori
Brown University
michael_lepori@brown.edu

Abstract

Object binding is a foundational process in visual cognition, during which low-level perceptual features are joined into object representations. Binding has been considered a fundamental challenge for neural networks, and a major milestone on the way to artificial models with flexible visual intelligence. Recently, several investigations have demonstrated evidence that binding mechanisms emerge in pretrained vision models, enabling them to associate portions of an image that contain an object. The question remains: ***how*** *are these models binding objects together?*

In this work, we investigate whether vision models rely on the principle of Gestalt continuity to perform object binding, over and above other principles like similarity and proximity. Using synthetic datasets, we demonstrate that binding probes are sensitive to continuity across a wide range of pretrained vision transformers. Next, we uncover particular attention heads that track continuity, and show that these heads generalize across datasets. Finally, we ablate these attention heads, and show that they often contribute to producing representations that encode object binding.

1. Introduction

Natural scenes are replete with low-level perceptual features like colors, shapes, textures, and motion cues. Despite this complexity, humans can instantly parse this information into sets of coherent objects. This ability to join low-level features together into object-level representations is known as object binding [22], and it has long been thought to play a fundamental role in visual cognition [8, 23, 25].

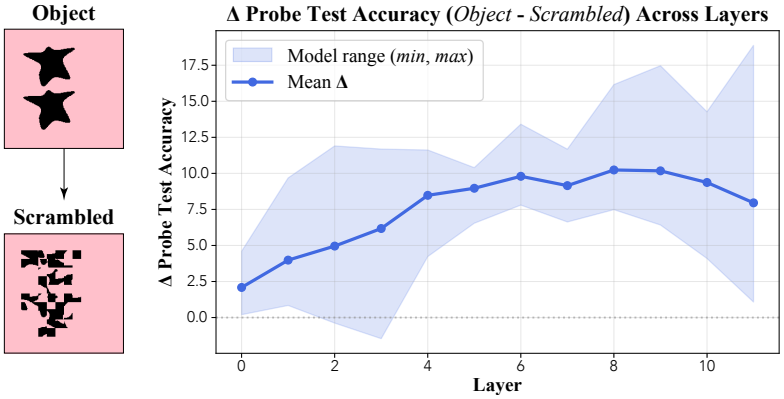
Many performant contemporary vision models are based on the transformer architecture, which processes images as sets of fixed-size square patches [5, 24]. Recent work has suggested that these models have learned mechanisms for binding objects, resulting in representations indicating which image patches contain a particular object [16]. The question remains: *how* are these models implementing object binding (i.e., how does the model know which patches to assign binding representations to)? For insight, we look to the study of human perception. 20th century Gestalt Psychology has long articulated principles that humans rely on when parsing scenes from low-level perceptual cues [14, 27]. These principles (e.g. similarity, proximity, closure, etc.) describe common low-level visual motifs that suggest structure within the visual world. In this work, we focus on the principle of *continuity*: the intuitive heuristic that visual elements that form a continuous curve are likely to belong to the same object or contour.

We create controlled synthetic datasets in order to isolate the effect of continuity from two other Gestalt principles: proximity and similarity. Across a range of vision transformer models, we find that continuity routinely contributes to forming binding representations. Furthermore, we discover specific attention heads — *Gestalt continuity heads* — that track continuous curves across image patches, and show that these naturally arise across vision models that were trained using diverse learning objectives. Finally, we analyze the impact of ablating Gestalt continuity heads on object binding, and observe that ablating these heads often selectively impacts binding representations for images that can leverage continuity for object binding.

2. Related Work

The "binding problem" is thought to pose a particular challenge to neural network models [8, 25], leading some to conclude that models may require strong architectural inductive biases to learn bound object representations [18]. However, recent work has found widespread evidence of binding mechanisms across contemporary vision [15, 16], language [6], and multimodal [1] transformer models.

Much of this work posits the existence of a "binding-ID" subspace in a model's intermediate activations, which can be populated with "binding vectors" that are applied to bind activations to one another. However, it is unclear how these binding vectors are created. The nature of this mechanism is further obscured by another line of work arguing that purely feedforward networks (including transformers) may struggle with perceptual grouping according to Gestalt principles



(a) (b)

Model	max(**Probe Test Acc.**) Blobs	Curves
DINO ViT-B/16	91.2% (+4.3)	86.9% (+10.6)
DINOv2 ViT-B/14	98.7% (+1.4)	97.3% (+16.6)
MAE ViT-B/16	99.0% (+3.7)	95.3% (+16.1)
ImageNet ViT-B/16	95.5% (+6.1)	89.4% (+11.7)
ImageNet ViT-B/32	95.5% (+4.6)	90.9% (+15.6)
CLIP ViT-B/16	91.3% (+1.4)	89.9% (+9.4)
CLIP ViT-B/32	90.4% (+3.2)	87.2% (+11.2)

(c)

Figure 1. **Object binding probe stimuli and results. (a)** An example stimulus pair from the dataset used to train and test the binding probes (Object and Scrambled$_{location}$). More example images from the Object and Scrambled datasets can be found in Appendix A. **(b)** Difference between object binding probe test accuracy on Blobs Object stimuli vs. Scrambled stimuli (taking the maximum probe test accuracy over {Scrambled$_{orientation}$, Scrambled$_{location}$}), averaged across models. See Appendix B for results on the Curves dataset. **(c)** Maximum object binding probe test accuracy for Object stimuli from the Blobs and Curves datasets across model layers. The quantities in the parentheses show the delta with maximum probe test accuracy on the Scrambled datasets; for example, (+6.1) indicates that maximum probe test accuracy on the Object dataset is 6.1 points higher than the probe test accuracy on either of the Scrambled datasets. Scrambled dataset accuracies are reported explicitly in Appendix B. Overall, we find that binding probes achieve higher performance when the stimuli obey continuity, even when controlling for proximity and surface similarity of the patches. This indicates that continuity plays a role in generating binding representations across vision transformer models. See Appendix B for full results.

[12, 17, 21]. In contrast, we demonstrate the emergence of attention heads that track continuity across patch boundaries in a variety of transformers. We then demonstrate that these heads are implicated in producing binding vectors.

3. Methods

Models We study a range of vision transformers, which are pretrained on different tasks but share most of their architectural details. Specifically, we study several 12-layer self-supervised models: DINO ViT-B/16 [2], DINOv2 ViT-B/14 [19], and a Masked Autoencoder model (MAE ViT-B/16; He et al. [11]); two versions of the CLIP model [20]: CLIP ViT-B/16 and B/32 (which have patch sizes of 16×16 and 32×32 pixels, respectively); and two supervised models: ImageNet ViT-B/16 and B/32 [5].

Datasets For our main set of experiments, we generate a synthetic dataset of filled blobs in which each image consists of a randomly colored blob on a randomly colored background. The boundary between blob and background forms a continuous closed curve whose shape we can disrupt, allowing us to probe for representations of Gestalt continuity in models. Additionally, we replicate our main results using a dataset of open curves. See Appendix A for dataset generation details and examples.

Binding Probes We adopt the highest performing probing strategy from Li et al. [16], the quadratic probe. This probe takes two patch activations, x and y, as input and computes $\sigma(xWy^T + b)$, where W is constrained to be symmetric. This probe is trained with weight decay in order to encourage a low effective rank.

4. Object Binding is Sensitive to Visual Continuity Cues

First, we establish that the binding representations discovered by the quadratic binding probes are sensitive to object continuity, over and above their sensitivity to object similarity and proximity. We train binding probes on a dataset comprised of images containing two copies of identical random blob shapes (see Figure 1a, top row). The probe is trained on pairs of activations corresponding to patches belonging to these objects, and must discriminate whether the two patches belong to the same or different objects. These objects are identical in shape, color, and, texture, reducing the probe's reliance on object similarity to establish binding. We train probes for each layer of all models described in Section 3 for 20 epochs using the Adam optimizer [13] with a learning rate of 0.001 weight decay factor of 0.01 and a stepwise learning rate scheduler with $\gamma = 0.2$.

We then evaluate the probe on three datasets: a held out within-distribution test set of whole objects (**Object**) and two control datasets designed to disrupt continuity cues (**Scrambled**; see Figure 1a). The **Scrambled**$_{orientation}$ dataset randomly rotates each image patch in the Object images by $d \in \{90°, 180°, 270°\}$. This preserves the content and position of each patch while systematically breaking

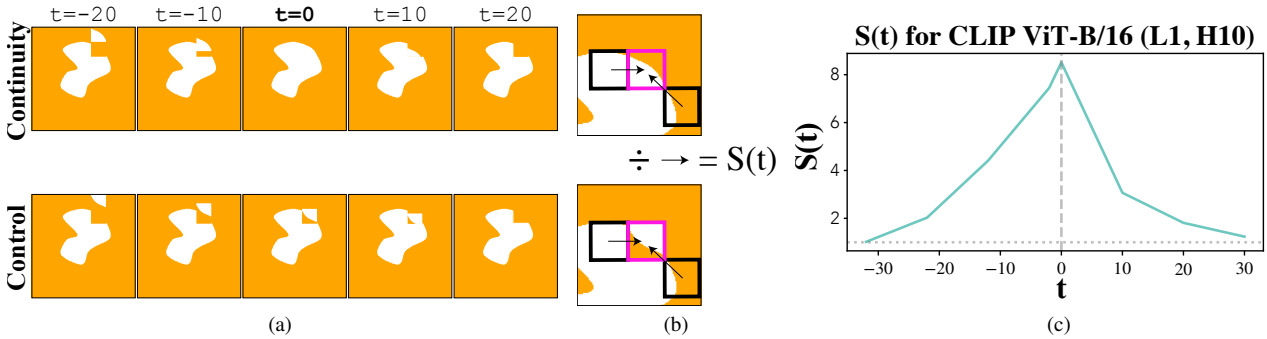


Figure 2. **Procedure for identifying Gestalt continuity heads. (a)** Top row: We parametrically vary continuity by translating a target perimeter patch. At $t = 0$, the patch is maximally aligned with its neighbors, creating a continuous curve. Bottom row: We create a control condition where the target patch at $t = 0$ contains the same pixel information but randomly rotated as to disrupt continuity between neighboring patches. **b** At each t and for each attention head, we measure attention from neighboring patches containing the curve (black squares) to the target patch (blue square). We then divide the maximum neighbor attention for the continuity condition by the same quantity for the control condition to get $S(t)$, the continuity sensitivity index. **c** $S(t)$ for an attention head from CLIP ViT-B/16 with a high continuity sensitivity index (Layer 1, Head 10). Attention peaks at the maximally-aligned time step $t = 0$. See Appendix C for all models.

the continuity of the object border. The **Scrambled**$_{location}$ dataset randomly permutes the locations of patches that contain a particular object. This preserves the overall content and position of the *set* of patches containing the object. The extent to which a model's Object probe performance exceeds its Scrambled performance is indicative of its reliance on continuity cues when performing object binding (as opposed to simple patch similarity or proximity).

From Figure 1, we see that probe performance on the Object dataset regularly outperforms performance on the Scrambled control datasets, with performance increasing from early to middle layers. However, we note that performance on the control sets far exceeds the majority-class baseline accuracy of 50.1%(see Table 1c). This indicates that models are relying on continuity *in addition to* other Gestalt principles, such as proximity.

5. Specific Attention Heads Detect Continuity

Having shown that object binding is sensitive to continuity cues in the image, we now ask whether this sensitivity can be localized to specific attention heads. We hypothesize that certain heads may act as dedicated continuity detectors, preferentially attending between neighboring patches when the visual content of each patch forms a continuous curve across patch boundaries.

To identify these hypothesized *Gestalt continuity heads*, we design an experiment that parametrically varies continuity in stimuli from our Blobs dataset. For this analysis, we use images that contain just one blob (in contrast to the datasets used to assess object binding). For each blob stimulus, we select a single perimeter patch and translate it along a trajectory such that at $t = 0$ it is maximally aligned with the curve formed by the object boundary, and at $t = \pm n$ it is progressively displaced (see Figure 2a, top row). For a given attention head, we measure the maximum attention directed from neighboring curve-completing perimeter patches to the target perimeter patch at each time step t. We repeat this for all possible perimeter patches in a given image and average to arrive at a scalar-valued score of that head's attention to continuity.

To control for low-level content of the patches, we construct a control condition in which the target patch at each time step is randomly rotated by $d \in \{90°, 270°\}$ (see Figure 2a). This preserves the pixel content of the patch at $t = 0$ while disrupting the continuity of the curve between the target and its neighbors. We then compute a *continuity sensitivity index* $S(t)$ at each time step by dividing the maximum neighbor-to-target attention in the base condition by the corresponding quantity in the rotated condition, applying a floor of 1 (see Figure 2b). A sensitivity index greater than 1 indicates that the head allocates more attention when the curve is completed between neighboring patches than when it is not, despite identical pixel content at the target location; for example, a sensitivity index of 8 means the head is 8× more sensitive to continuity than mere pixel content or patch proximity. Plotting $S(t)$ across time steps produces a "tuning curve" [3] — heads that are sensitive to continuity exhibit a characteristic Gaussian-like profile peaking at $t = 0$ where curve alignment is maximal (see Figure 2c).

Across models, we find a set of heads exhibiting high continuity sensitivity indices at $t = 0$, mostly clustered within the first four layers (Appendix D). To validate these findings, we repeat the procedure on two additional stimulus sets: the Curves dataset described in Section 4, and a dataset of naturalistic objects drawn from ImageNet [4] (see Appendix A for details). We compute Pearson's r of

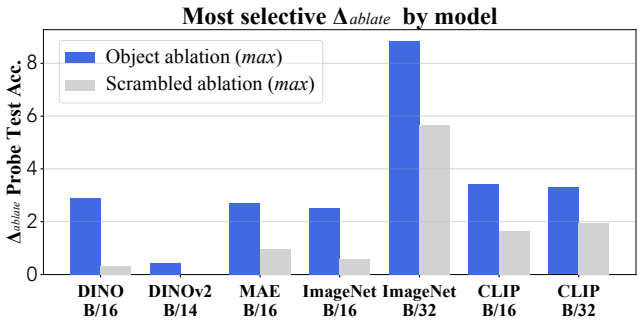


Figure 3. **Attention ablation results.** We plot the impact of ablating Gestalt continuity heads (Δ_{ablate}) on either the Object dataset or the Scrambled datasets. We plot the most *selective* ablations across layers for each model on the Blobs dataset, where selectivity is defined as $\Delta^{\text{Object}}_{ablate} - \Delta^{\text{Scrambled}}_{ablate}$. We find that ablating Gestalt continuity heads impacts probe performance more on the Object dataset than the Scrambled datasets. See Appendix F for full results (including Curves results) and additional controls.

the continuity sensitivity index at $t = 0$ over all heads between each pair of datasets (Blobs, Curves, and ImageNet). Continuity heads are highly consistent across datasets, with an average Pearson's r of 0.664 across models and dataset pairs (see Appendix E for a breakdown by model and p-values). Furthermore, when examining the full tuning curves across all time steps, identified continuity heads display the characteristic Gaussian-like tuning curves expected of a continuity detector: attention between neighboring patches peaks at $t = 0$, where curve alignment is maximal, and decays monotonically as the target patch becomes increasingly misaligned with its neighbors (Appendix C). Overall, this indicates that particular attention heads are sensitive to Gestalt continuity cues, providing at least one concrete mechanism for representing continuity across patch boundaries. We note that the concentration of continuity heads across layers roughly mirrors probe performance across layers (see Appendix D).

6. Gestalt Heads are Causally Implicated in Binding

Having found that binding representations are sensitive to continuity in Section 4, and having identified a particular set of attention heads that are sensitive to continuity in Section 5, we now seek to understand whether Gestalt continuity heads play a role in the construction of binding representations. To do so, we perform a mean ablation [26] of the outputs of a set of Gestalt continuity heads and observe the impact on binding probe performance on the Object and Scrambled datasets. Specifically, for each probe, we mean-ablate the top 5 attention heads with the highest continuity sensitivity index scores (Section 5) across all layers preceding the probed representations.

In Figure 3, we see the results of this ablation analysis across models. We seek to demonstrate that these Gestalt continuity heads are *selective*, or that ablating them impacts performance on the Object dataset more than the control datasets. Across models, we find that there exists a set of Gestalt continuity heads that are selectively implicated in producing binding vectors for the Object dataset. In Appendix F we compare the selectivity of the identified Gestalt continuity heads to other sets of randomly selected heads. We find that ablating the Gestalt continuity heads often impacts binding probe performance more selectively than ablating random sets of heads.

Our earlier experiments explicitly demonstrated that binding representations are not uniquely the result of perceptual continuity cues, so we expect several heads beyond Gestalt continuity heads to be implicated in creating binding representations. However, our results suggest that Gestalt continuity heads play a role in the formation of object binding representations for many of the models.

7. Discussion

Limitations and Future Work This work can be improved and extended in several ways. First, one can explore different ways to select Gestalt continuity heads for ablation. We rely on a simple heuristic for assessing selecting heads: taking the top-k heads according to continuity score, with $k = 5$. More dynamic or layer-specific selection procedures might reveal even stronger contribution to binding.

Additionally, while our work reveals that continuity plays a role in producing binding representations, the degree to which it contributes when other Gestalt principles are available (as is the case with naturalistic stimuli) remains an open question. Future work should try to discover attention heads that are sensitive to other Gestalt principles, such as closure or surface form similarity. Finally, existing work has sought to understand the mechanisms behind "detokenization" in language models [7, 9], or the process by which tokens are bound together to form meaningful units. Our work seeks to understand detokenization mechanisms in vision models by using principles from the vision sciences, but other work might apply techniques from language model mechanistic interpretability to this problem.

Conclusion This work employs a synthetic experimental paradigm and classic ideas from the vision sciences to investigate the mechanisms that give rise to object binding in vision transformers. We find that 1) continuity contributes to the formation of object binding representations (as discovered by binding probes), 2) specific heads in vision transformers track continuity across image patch boundaries, and 3) these heads oftentimes specifically contribute to the formation of the binding representation found by binding probes. In summary, this work advances our understanding of how contemporary vision models form object-level representations from low-level perceptual cues.

References

[1] Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. *arXiv preprint arXiv:2506.15871*, 2025. 1

[2] 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. 2

[3] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[5] 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. 1, 2

[6] Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2023. 1

[7] Sheridan Feucht, David Atkinson, Byron C Wallace, and David Bau. Token erasure as a footprint of implicit vocabulary items in llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9727–9739, 2024. 4

[8] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020. 1

[9] Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023. 4

[10] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 6

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[12] Junkyung Kim, Drew Linsley, Kalpit Thakkar, and Thomas Serre. Disentangling neural mechanisms for perceptual grouping. In *International Conference on Learning Representations*, 2019. 2

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[14] Kurt Koffka. *Principles of Gestalt psychology*. routledge, 1935. 1

[15] Michael A Lepori, Alexa R Tartaglini, Wai K 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. 1

[16] Yihao Li, Saeed Salehi, Lyle Ungar, and Konrad P Kording. Does object binding naturally emerge in large pretrained vision transformers? *arXiv preprint arXiv:2510.24709*, 2025. 1, 2

[17] Drew Linsley, Junkyung Kim, Vijay Veerabadran, Charles Windolf, and Thomas Serre. Learning long-range spatial dependencies with horizontal gated recurrent units. *Advances in neural information processing systems*, 31, 2018. 2

[18] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020. 1

[19] 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. *Transactions on Machine Learning Research Journal*, 2024. 2

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[21] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2020. 2

[22] Anne Treisman. The binding problem. *Current opinion in neurobiology*, 6(2):171–178, 1996. 1

[23] Anne Treisman and Hilary Schmidt. Illusory conjunctions in the perception of objects. *Cognitive psychology*, 14(1): 107–141, 1982. 1

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[25] Christoph Von der Malsburg. The what and why of binding: the modeler's perspective. *Neuron*, 24(1):95–104, 1999. 1

[26] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *The Eleventh International Conference on Learning Representations*, 2022. 4

[27] Max Wertheimer. *Untersuchungen zur Lehre von der Gestalt*. Springer, 1923. 1