Shape-based Figure-Ground Organization in Vision Transformers

Matthias Tangemann[1,2] **Benjamin Lo**[3,4] **Zygmunt Pizlo**[5] **Kaleem Siddiqi**[3,4]
Dirk B. Walther[1] **Sven Dickinson**[1,2]

[1]University of Toronto [2]Vector Institute [3]McGill University [4]MILA [5]UC Irvine

mtangemann@cs.toronto.edu

Abstract

Rich perceptual structure emerges in modern vision encoders. Several studies have shown that the internal representation in many vision transformers separates foreground objects from the background, however, little is known about the visual cues that have been learned by these models. In this work, we study whether figure-ground organization in a broad range of 25 vision transformers relies on generalizable shape cues, rather than on superficial texture statistics. To do so we train linear probes to classify patches as figure or ground, and evaluate them zero-shot on synthetic stimuli where individual cues are systematically manipulated. For a broad range of vision encoders, we find that probes are robust to the random replacement of textures in the input image, suggesting that the emerging figure-ground organization relies on shape cues. Probes also generalize in a zero-shot fashion to synthetic images where individual shape cues, recognizability and spatial priors are controlled for. Our experiments reveal a strong influence of surroundedness and convexity cues, while results for parallelism and symmetry are more nuanced. More detailed analyses reveal stronger evidence for shape-based cues in larger models, that models trained for object-recognition tend to be less robust to texture changes and that figure-ground information emerges most prominently in middle layers. Taken together, our study demonstrates that emerging figure-ground organization in recent vision encoders relies strongly on generalizable shape cues, advancing our understanding of the internal processing in vision transformers, and opening an exciting avenue for computational research on human visual perception.

Code and data is available at https://github.com/mtangemann/mlvbench.

1 Introduction

Figure-ground organization is an important process in mid-level vision that identifies a foreground object (the *figure*) from its surroundings (the *ground*). The resulting structured representation forms the basis for higher level tasks such as recognition or spatial reasoning, and enables generalization to novel situations. For example, adequately responding to an unknown obstacle when driving a car requires correctly separating the object from its surroundings, while recognizing the object is not necessary. Better understanding figure-ground organization in mid-level representations of vision models therefore greatly contributes to building more robust visual intelligence.

Perceptual structure, including figure-ground organization, has been shown to emerge from end-to-end training in recent vision transformers [14, 52, 11]. Which image cues these models use to separate figure and ground, however, remains largely unknown. Research on human visual perception has highlighted the strong influence of shape cues, including many of the classical Gestalt principles

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

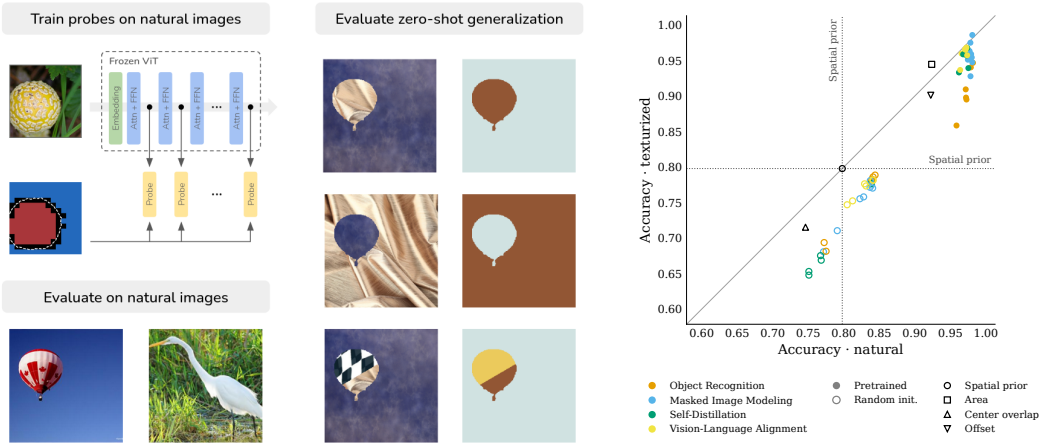

Figure 1: *(left)* We extract patch representations from frozen, pre-trained ViTs and fit linear probes to predict figure-ground assignment on natural images, excluding patches that span both foreground and background (shown in black). We evaluate probes on both natural images and cue-controlled stimuli where textures are randomly replaced. *(right)* Probes for most models are largely unaffected by disrupting texture information, indicating the use of shape cues. Baselines rule out architectural biases and trivial strategies based on region area and location.

[59]. Several of these cues have been shown to be rooted in the statistics of natural scenes [28, 23], so it is plausible that vision encoders might acquire similar cues during training. On the other hand, DNNs have been classically shown to rely primarily on texture cues [25, 7, 12]. This raises the question of whether the emerging figure-ground organization is mainly driven by texture cues as well, or whether recent methods have succeeded in learning more Gestalt-like cues.

In this work, we systematically study the role of shape cues for emergent figure-ground organization by evaluating the zero-shot generalization of linear probes (Figure 1). We extract spatial features from pretrained vision transformers and fit linear probes to classify patches as figure or ground on natural images. Testing the zero-shot generalization of natural image probes to cue-controlled experimental stimuli then allows us to draw conclusions about the figure-ground cues used by the network. If the model relies primarily on texture cues, probes trained on natural images will not generalize to stimuli in which all textures are randomly replaced. Conversely, if the network relies on shape cues, we would expect the probe performance to be unaffected. Following the same logic, we further test the use of individual shape cues discussed in human vision science: surroundedness, convexity, symmetry and parallelism [59].

Our results show that shape-based cues for figure-ground organization emerge in a broad range of 25 vision encoders. All tested models encode figure-ground information for natural scenes and show some level of robustness to texture disruptions, clearly outperforming random features and trivial baselines based on region area and location. A subset of models achieves the same performance on randomly textured images as on the original test set, and shows strong biases towards surroundedness and convexity. More detailed analyses further reveal that larger models tend to show stronger use of shape cues while we do not observe a clear influence of the pretraining task. These results demonstrate that the emerging figure-ground organization in many vision transformers relies on generalizable shape cues that have been learned during pretraining.

In summary, our paper makes the following contributions:

- We show that figure-ground information emerges across a broad range of 25 Vision Transformers pretrained on different tasks.
- We demonstrate that figure-ground organization strongly relies on shape cues for most models, and in particular surroundedness and convexity cues.
- We find that figure-ground information typically peaks in middle layers and that model size and the pretraining task influence the emergence of shape cues.

Our findings represent an important contribution towards understanding the emerging representations in vision transformers. To our knowledge, we are the first to demonstrate the role of shape cues for emergent figure-ground organization. This qualitative similarity to classical Gestalt cues provides an exciting direction for computational research on perceptual organization in human vision.

2 Related Work

Figure-ground organization. Figure-ground organization is a central capability of the human visual system that contributes to structuring the sensory input into coherent surfaces and objects [59, 41, 66]. Research in psychophysics has identified a range of cues that are used by the human visual system for figure-ground assignment ([47, 5, 31, 54, 42, 43] see [59] for a review). Among these, surroundedness, convexity, and symmetry play central roles in figure-ground assignment and shape interpretation. Regions that are spatially surrounded are more likely to be perceived as figure rather than background, while convex regions tend to dominate over concave counterparts in determining object boundaries. Symmetry further biases perception by promoting the grouping of elements into unified forms, reflecting the visual system's sensitivity to regularity and structural simplicity.

Emerging perceptual organization in vision transformers. Caron et al. [14] observed that object-centric structure emerges in Vision Transformers trained using the self-supervised DINO method. Following models like iBOT [72] and improved versions of the DINO model [40, 18, 52] further confirmed this observation. This inspired a range of segmentation methods based on clustering patch features from pretrained ViTs ([51, 29, 38, 63, 3, 49, 45, 69, 67, 46], see [50] for a review). Using the emergent segmentation as pseudo-ground truth has further enabled unsupervised segmentation methods to approach the performance of their supervised counterparts [61, 62, 68]. While DINO models remain the most popular models for studying emergent perceptual organization, other works have demonstrated the emergence of object segmentation in a broader range of vision transformers [17, 36].

Shape-based processing in vision encoders. A long line of work has studied whether DNNs learn shape-based visual processing similar to the human visual system (see [64] for a recent review). One line of work has revealed that object recognition CNNs are strongly driven by texture cues, while humans dominantly rely on global shape [7, 26]. Moreover, early DNNs could be approximated quite well by networks that only consider local texture information [12]. This trend however seems to be changing, as more recent studies increasingly find evidence for more human aligned, shape-based processing in vision models [27, 19, 13]. Recently, [1] have further shown that the emerging segmentation in self-supervised vision encoders aligns with human object perception.

Several studies have investigated whether DNNs follow classical Gestalt principles of perceptual grouping, including both failure and success cases depending on the model and experimental paradigm [6, 2, 22, 32, 10, 55, 70, 35, 71, 37, 35, 48, 56, 34]. Far fewer works have studied whether figure-ground organization in vision encoders follows similar cues as the human visual system, suggesting a role of convexity in masked image models [35, 4].

3 Methods

We examine whether pretrained ViTs use shape cues for figure-ground organization by studying the generalization of linear probes to cue-controlled stimuli. All probes are trained on natural images to extract figure-ground information from frozen ViT features. Evaluating the zero-shot generalization of these probes to cue-controlled stimuli then allows us to draw conclusions about the cues that the model relies on: Probes for models that primarily rely on texture rather than shape cues will not generalize from natural images to stimuli where the textures are randomly replaced. Below, we describe the protocol for probe fitting and evaluation (Section 3.1) and the vision encoders considered in our study (Section 3.2). The cue-controlled test conditions are described with the experiments in Section 4.

3.1 Probe Fitting and Evaluation

We fit linear figure-ground probes on images and ground truth segmentation masks from the MSRA-10K dataset [15], which provides photographs featuring a single, dominant foreground object. The dataset is randomly split into 80% training, 10% validation, and 10% test images. Each image is resized so that the shorter side matches the model's native input resolution and is then center-cropped.

The probed ViT is applied to the input images $x \in \mathbb{R}^{H \times W \times 3}$, and intermediate features $f^{(l)} \in \mathbb{R}^{h \times w \times C}$ are extracted for each layer l, where $(h, w) = (H/p, W/p)$ and p is the patch size of the model. We downscale the ground truth segmentation masks $s \in \{0, 1\}^{H \times W}$ to the internal grid, yielding $\tilde{s} \in \{0, 1\}^{h \times w}$. Patches that overlap both foreground and background regions receive ambiguous labels and are therefore masked in the training loss and evaluation metrics.

Probes. For each model layer, we fit a logistic regression model that predicts the probability of a patch belonging to the foreground:

$$\text{probe}(f_{i,j}) = \sigma(w^T f_{i,j} + b)$$

Probes are trained with binary cross-entropy (BCE) loss using the Adam optimizer [33] with a batch size of 256 for 2000 steps, a learning rate of 0.001 and no weight decay. Inspection of validation loss curves confirmed that probes converge within this limit. We evaluate probes every 200 steps, restore the best step and select the best layer based on the validation loss. All results are reported on held-out test sets. In total, we fit 3108 probes on combined 1.6B samples within 120 GPU hours (NVIDIA L40S).

Evaluation. We evaluate all probes on the held-out i.i.d. test set, and apply them zero-shot to several cue-controlled conditions. We use accuracy as the primary metric, and additionally report BCE, information gain over the spatial prior, and intersection-over-union (IoU) in the appendix.

Baselines. We consider several baselines to rule out trivial explanations for the observed results:

- *Spatial prior*: We compute the prior foreground probability for each location (i, j) on the training set and evaluate this constant prediction on all test sets. This corresponds to a probe that purely relies on the positional information in ViT features.
- *Random features*: We fit probes for randomly initialized variants for all models using the same protocol as described above in order to rule out any architectural biases.
- *Region properties*: We use ground truth texture segmentation information and select the foreground region based on heuristic properties including the region area, overlap with the image center, and the offset of the region centroid from the image center. This simulates models that apply texture segmentation in combination with a trivial heuristic.

3.2 Models

We study a diverse set of 25 Vision Transformer models. For all models, we use the implementations and checkpoints as provided by the `timm` library [65] and test the ViT-B and ViT-L variants. We consider ViTs trained for supervised and unsupervised training objectives:

- **Object recognition** models are trained with supervision on ImageNet [20]. We include an improved version of the original ViT (AugReg, [53]), DeIT III [57] and FlexiViT [9].
- **Vision-Language Alignment** models have been pioneered by CLIP [44]. We additionally consider the more recent SigLIP 2 [58] and Perception Encoder Spatial models [11].
- **Masked Image Modeling** is a self-supervised pretraining task where the model is trained to reconstruct missing patches in the input image. We include the original Masked Auteoncoder [30] and BEiT [8]. Further, we consider EVA-02 [24], which is trained to reconstruct CLIP features for masked patches, and BEiT-3 [60], which is jointly trained on masked images and text and finetuned on ImageNet.
- **Self-Distillation** is a self-supervised pretext task used by the DINO models. We include the original DINO model, DINOv2 with registers, and DINOv3 [14, 40, 18, 52].

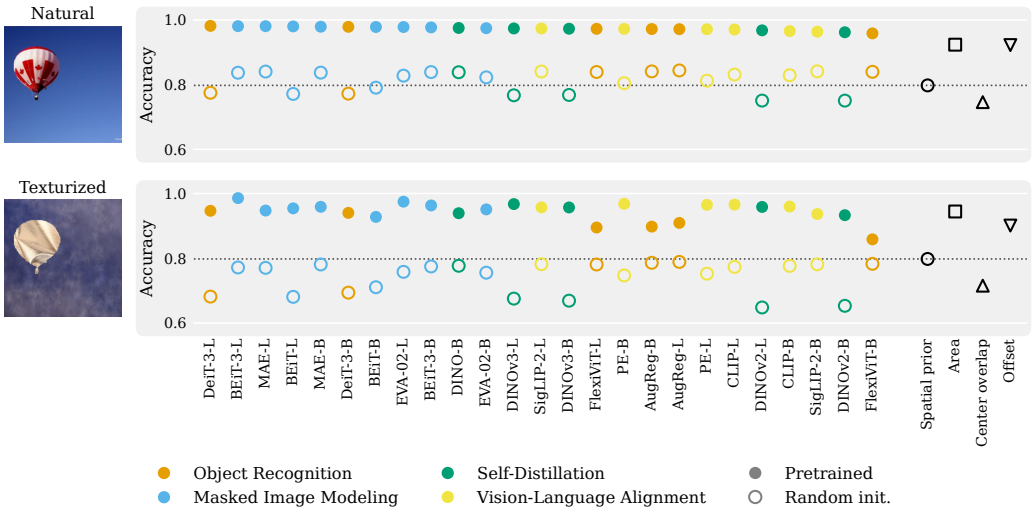

Figure 2: Probe performance for the i.i.d. natural test set, and applied zero-shot to the same images but with the textures randomly replaced. Many models reach high accuracies for both the natural and texturized images, demonstrating the use of shape cues in these models.

Links to the precise model checkpoints are provided in Table 9 in the appendix.

4 Experiments

All experiments are based on linear probes that are trained to classify image patches as figure or ground in natural scenes. The performance of the models on the i.i.d. test set in Figure 2 shows that patch tokens from all 25 tested models provide strong figure-ground information. The results further reveal a substantial spatial bias, which is likely caused by the tendency of photographs to contain the foreground object near the image center. While the probes for pretrained models outperform this spatial prior by a large margin, probes for random features perform much worse. Figure-ground organization is therefore not an architectural bias but requires learning.

So far, our results corroborate previous studies showing the emergence of perceptual structure in Vision Transformers and further extend the range of tested models. In order to gain insight about the visual cues that underlie this phenomenon, we assess whether figure-ground probes trained on natural images generalize in a zero-shot fashion to stimuli in which shape and texture cues are systematically controlled. In Section 4.1 we test whether models are robust to the disruption of texture cues, while in Section 4.2 we take a more detailed look at individual shape cues. Finally, in Section 4.3 we evaluate the influence of model size and pretraining task, and in Section 4.4 we test which layers provide the strongest figure-ground information.

4.1 Generalization to texturized stimuli

If figure-ground organization in Vision Transformers relies on shape cues, we would expect the probe predictions to be invariant to changes in texture. Conversely, if figure-ground organization relies purely on texture cues, we would expect the performance to drop to chance level. To test these possibilities, we generate a variant of the natural test set with the same segmentation mask, but where each segment is filled with a random texture from the DTD dataset [16] or a random color. In total, we create 6 variants for each image (see Figure 1 for examples). We fill figure and ground with random textures or colors, and as a control include variants where the assignment is reversed. We observed that the performance of baselines using texture segmentation with a trivial foreground heuristic was quite high for these variants. We therefore include additional stimulus variations where the figure is split by a straight line and filled with two different textures or colors. All results reported in the main paper average over the different variants, and more detailed results are included in the appendix.

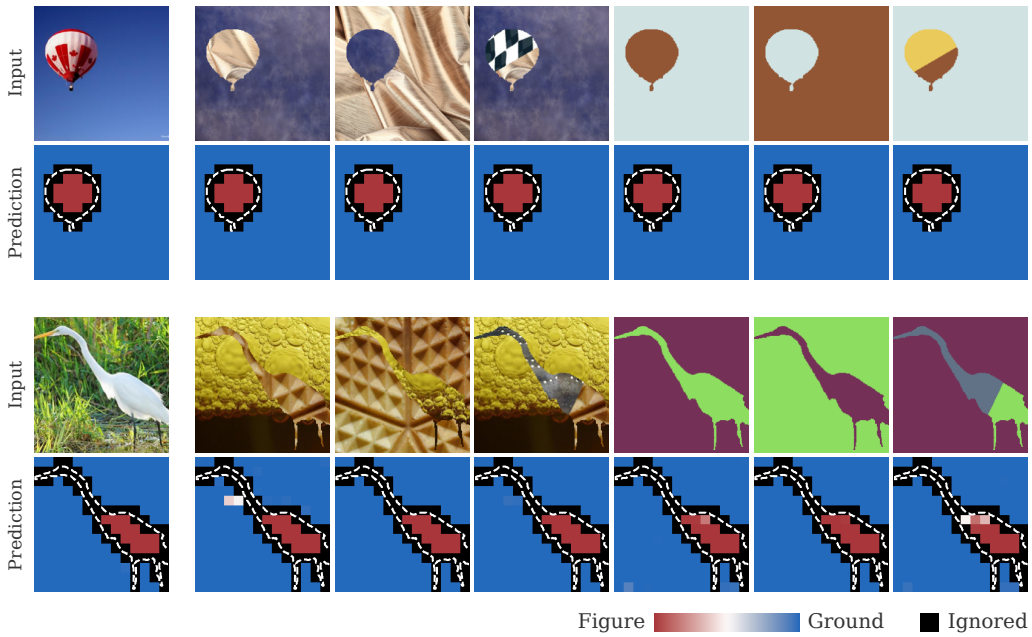

Figure 3: Figure-ground predictions based on layer 12 from BeiT-3-L [60]. The probe generalizes from natural images to randomized textures with hardly any loss in performance, providing support for shape-based figure-ground cues used by the model.

The results in Figure 2 provide strong evidence for shape-based figure-ground organization in many Vision Transformers, as probes from most models generalize well to synthetic stimulus variants where texture information is destroyed. The direct comparison of the i.i.d. and zero-shot performances in Figure 1 further shows that for several models, probes generalize to changes in texture without any loss in performance. For other models, and in particular object recognition models except DeiT-III [57], we observe a substantial drop in performance, yet there is still a clear improvement over the spatial prior baseline. The results for the individual variants (Figure 11 in the appendix) show that these models struggle on the colorized stimuli, while they are less affected for the textured variants.

4.2 Isolation of shape cues

The previous results demonstrate that figure-ground organization in many ViTs relies on shape cues. However, we do not know *which* shape cues have been learned by the models. Moreover, due to the strong spatial bias in natural images, we cannot fully rule out a strong influence of trivial positional heuristics. In this section, we therefore test the figure-ground probes zero-shot on stimuli designed to isolate the role of individual shape cues. We focus on four cues that have been discussed in the human vision literature: surroundedness, convexity, parallelism and symmetry (see examples in Figure 4).

- *Surroundedness*: Stimuli contain a random shape from the idsprites dataset ([21]). The shapes include both convex and concave regions, and are surrounded by the background on all sides. The relative area covered by the shape is uniformly sampled from $[0.08, 0.25]$ and $[0.5, 0.6]$, respectively, where both intervals are chosen with equal probability. When selecting the first interval, we additionally ensure that the image center is not covered by the shapes.
- *Convexity*: Foreground and background are separated by a random parabolic arc, $v = ku^2 + c$ in a randomly oriented coordinate frame, with $k \sim \text{Uniform}(1.0, 3.0)$ and c chosen to ensure a specific region area. We randomly sample the region area in $[0.2, 0.3]$ and $[0.7, 0.8]$, respectively, and both intervals are selected with equal probability.
- *Symmetry*: The image is divided into four equally sized vertical columns by three Bézier-curve edges. Two consecutive edges are made mirror-symmetric by copying and inverting

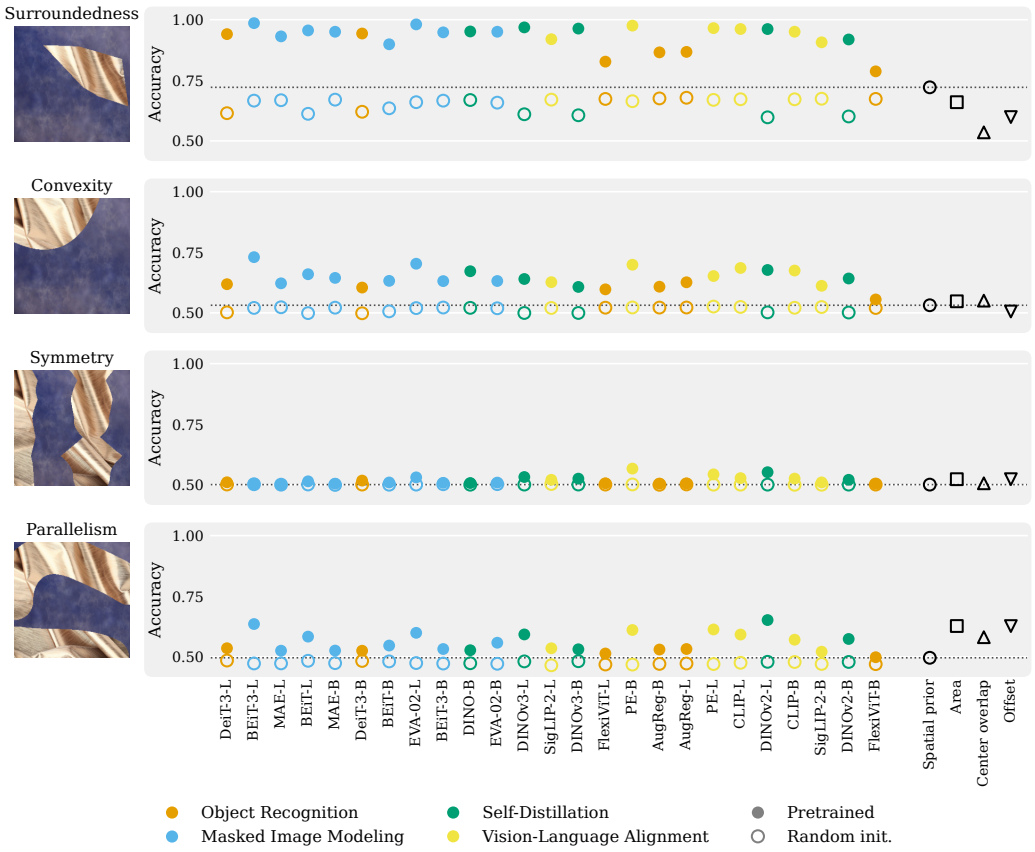

Figure 4: Zero-shot performance of the figure-ground probes on cue-controlled test sets. The results shows strong biases towards surrounded and convex regions, a weaker role of parallelism and no clear evidence for the use of symmetry.

their control points, producing two symmetric columns (foreground) flanked by two asymmetric columns (background). This design follows the classic Bahnsen column paradigm [5].

- *Parallelism*: Stimuli contain random stripes with a constant thickness of one third of the stimulus width. We generate the stripes by sampling curvature values from a Gaussian process prior with a squared exponential kernel and integrating the sampled function, starting from a random point on the image boundary.

The parameters for all stimulus classes, including the size ranges, have been selected in order to minimize the accuracy of the simple center-related baselines. We use random textures and colors for the regions, including the reverse variants. For the surroundedness and convexity stimuli, we additionally consider variants where the foreground region is split into two subregions.

The results in Figure 4 show clear evidence for surroundedness and convexity cues in most models. Probes for all ViTs outperform the baselines in the surroundedness condition, with near perfect predictions in many cases. The results in the convexity condition are weaker, but we observe a clear bias towards the convex region in most models that cannot be explained by any of the baselines. In the parallelism condition, we see a bias towards the stripes with parallel outlines, but the results require more careful interpretation. Since some of the baselines achieve similar accuracies, we cannot fully exclude the possibility that models rely on trivial heuristics. However, the layer-wise analysis in Section 4.4 shows that for certain models, later layers than those selected on the natural test set are substantially more biased towards parallelism, which provides evidence against the baselines in these cases. Finally, for the symmetry condition, we do not observe a clear bias towards perceiving the symmetric region as foreground.

Overall, these results contribute further evidence for generic, shape-based figure-ground organization in several Vision Transformers. The generalization to the surroundedness and convexity conditions demonstrates that figure-ground organization does not depend on texture cues, recognizable shapes, or the strong center bias of natural images.

4.3 Factors influencing the emergence of shape cues

In Figure 5, we analyze which model attributes influence the emergence of shape-based figure-ground organization in ViTs. We find effects of both model size and pretraining task. In all cases, except for convexity and symmetry, probes for ViT-L models perform significantly better than probes for the smaller ViT-B models. Furthermore, we see significant differences between pretraining tasks in many cases, suggesting a stronger overall performance of Masked Image Models, and a weaker overall performance of object recognition models. However, there are exceptions, as the DeiT-III object recognition model performs substantially better than other object recognition models.

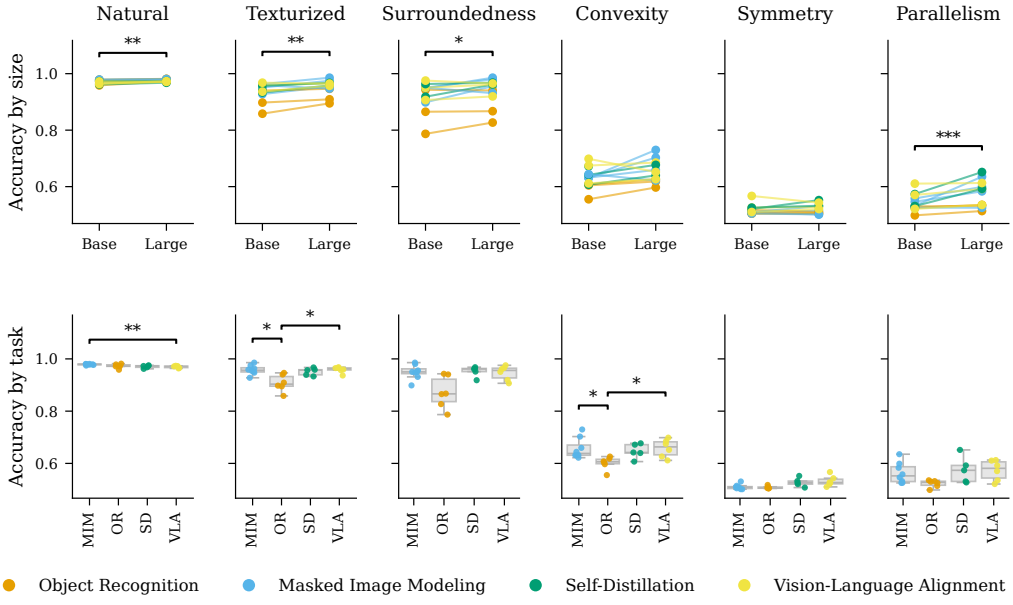

Figure 5: Probe performance by model size (top row) and pretraining task (bottom row: OR = Object Recognition, VLA = Vision-Language Alignment, MIM = Masked Image Modeling, SD = Self-Distillation). Brackets indicate statistical significance: (*/**/***: $p < .05/.01/.001$, non-significant pairs are omitted for clarity).

4.4 Layer-wise analysis

So far, we have analysed the performance of the best layer for each model, selected based on the validation loss. In Figure 6, we take a more detailed look at the performance of the individual layers. On natural images, we see a rise in figure-ground information during the first few layers, with a peak around the middle. For some models, we observe a drop in performance in later layers.

While early layers enable non-trivial predictions for natural images, the probes for early layers perform around chance-level for the generalization conditions. This suggests that the information extracted by the probes corresponds to features that are superficially correlated but do not provide robust figure-ground information. In some other cases, and in particular for the convexity and parallelism conditions, it seems that later layers are more strongly biased towards these cues than the layers selected on the validation set. Taking into account the relatively flat curves in the natural conditions, this is compatible with the hypothesis that figure-ground information plateaus early for the easier natural images but converges later for the more difficult synthetic conditions.

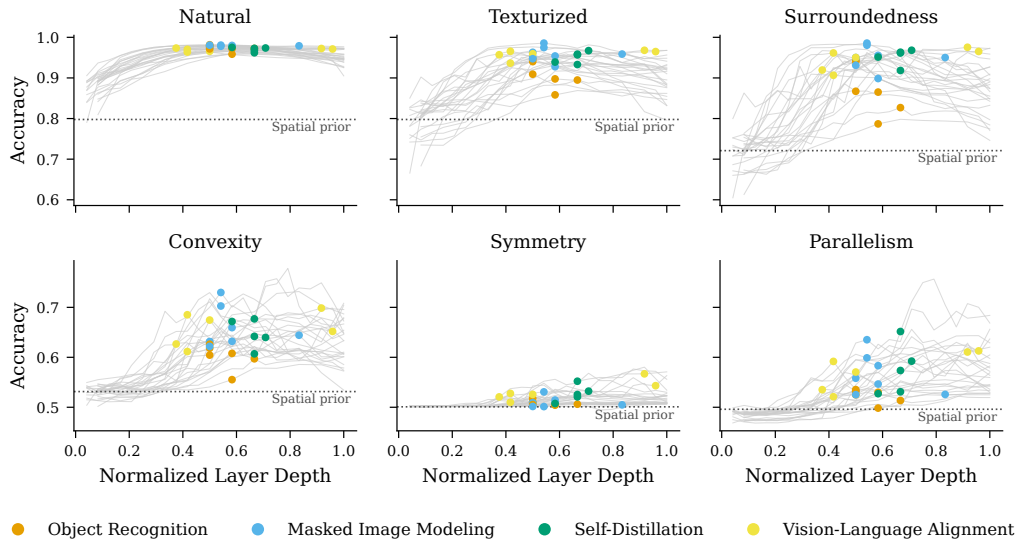

Figure 6: Per-layer performance for each test condition. The best layer selected on the validation set is marked with a colored dot. For most models, figure-ground information is best decoded from intermediate layers. More detailed information is provided in Figure 30 in the appendix.

4.5 Comparing generalization and decodability

We train probes directly on training sets matching the cue-controlled stimuli in order to estimate how well the cues can be decoded from the network representations *in principle*. The results in Figure 7 indicate that several cues, and in particular convexity, symmetry and parallelism, can be decoded much better from internal features than the zero-shot performance of the figure-ground probes. This is the case even for the best performing vision encoders where training directly on texturized images or the surroundedness condition does not improve results. Overall, this suggests that convexity, symmetry and parallelism are less strongly linked to figure-ground organization in these models than surroundedness.

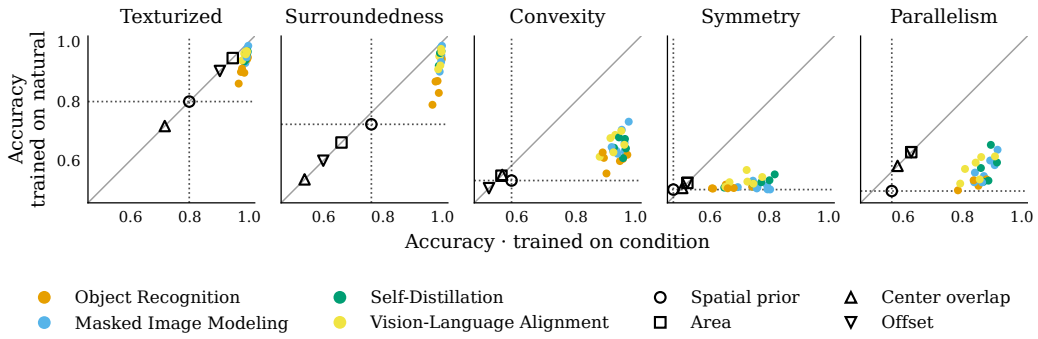

Figure 7: Comparison of probes trained on natural images with probes trained directly on the cue controlled conditions. For the strongest models, direct training on the texturized or surroundedness condition hardly improves over the zero-shot results. The performances improve substantially for convexity, symmetry and parallelism which suggests a weaker link to figure-ground organization.

5 Limitations

We rely on patch-wise, linear probes to decode figure-ground information from intermediate ViT features. This ensures that any positive result reflects readily accessible information that can be used

in principle by the key, query and value projections in subsequent layers. In cases where we did not find strong shape-based figure-ground information, it is nevertheless possible that a more powerful probe could successfully decode figure-ground assignment. Furthermore, while this paradigm allows us to study the cues that influence figure-ground information, our paradigm does not support any conclusions about how figure-ground information influences any higher-level capabilities of the respective network.

Finally, we do not directly compare DNNs with human behavior. While our study provides qualitative evidence for cues that are believed to underlie figure-ground information in humans as well, we cannot make strong claims about the similarity to human vision. Since our work focuses on establishing the use of shape cues for perceptual organization in DNNs, this question is out of scope for the present paper. However, given the qualitative similarity of some of the cues to well known Gestalt principles, this provides a highly promising direction for future work.

6 Discussion

We performed a systematic study evaluating whether emergent figure-ground organization in recent Vision Transformers relies on shape cues. Our experimental approach is based on testing zero-shot generalization of linear probes: only if the emergent structure in ViTs relies on shape cues would we expect the zeros-shot generalization of probes to stimuli where textures are randomly replaced. The results demonstrate strong generalization for the majority of models, and therefore strongly indicate shape-based processing in ViTs.

This finding challenges the common characterization of deep neural networks as primarily texture-driven. Classical work has shown strong texture biases of early CNNs [25, 7, 12], marking a striking difference to human visual perception. Our work strongly corroborates more recent studies that suggest a stronger shape bias in more recent models [27, 19] since probe performance was not impacted by randomizing textures in several models. Furthermore, the successful generalization to stimuli with randomly generated shapes rules out a central role of recognizable outlines and spatial biases, while the random feature control rules out architectural biases. While we cannot conclude that texture cues are never used by the models, our results show that shape-based processing is sufficient for figure-ground organization. The emergence of these cues across supervised, self-supervised, and vision-language models suggests that learning shape-based figure-ground cues is a robust phenomenon rather than an artifact of any particular training paradigm.

Finally, our results open an exciting avenue for research on human visual perception. Our results show that several ViTs learned to use human-like cues, including surroundedness and convexity and, to a lesser degree, parallelism. While we do not find evidence for the use of symmetry, the effectiveness of this cue has been debated for humans as well [31, 39]. Qualitatively, these results therefore seem compatible with human vision, but direct comparisons will be necessary to make stronger claims. If these similarities hold, modern vision encoders would be highly valuable model systems for studying open questions about the mechanisms and learning of Gestalt principles in human vision.

Acknowledgments and Disclosure of Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Samsung. The authors thank the Digital Research Alliance of Canada (alliancecan.ca) for providing computing resources.

References

[1] H. Adeli, S. Ahn, A. Luo, M. Zhang, N. Kriegeskorte, and G. Zelinsky. Human-Like Object Grouping in Self-supervised Vision Transformers. In: *Computer Vision – ECCV 2026*. Ed. by P. Favaro, Z. Kukelova, A. Maki, A. Rohrbach, K. Schindler, and F. Tombari. Cham: Springer Nature Switzerland, 2026, pp. 113–130.

[2] A. Amanatiadis, V. G. Kaburlasos, and E. B. Kosmatopoulos. Understanding Deep Convolutional Networks through Gestalt Theory. In: *2018 IEEE International Conference on Imaging Systems and Techniques (IST)*. 2018, pp. 1–6.

[3] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel. On the Effectiveness of ViT Features as Local Semantic Descriptors. In: *Computer Vision – ECCV 2022 Workshops*. Ed. by L. Karlinsky, T. Michaeli, and K. Nishino. Cham: Springer Nature Switzerland, 2022, pp. 39–55.

[4] S. Arnold and R. Gröbner. *Locating and Editing Figure-Ground Organization in Vision Transformers*. 2026. arXiv: 2603.06407.

[5] P. Bahnsen. Eine Untersuchung über Symmetrie und Asymmetrie bei visuellen Wahrnehmungen. In: *Zeitschrift für Psychologie* 108 (1928), pp. 129–154.

[6] N. Baker, G. Erlikhman, P. Kellman, and H. Lu. Deep Convolutional Networks do not Perceive Illusory Contours. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 40. 2018, pp. 1310–1315.

[7] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman. Deep convolutional networks do not classify based on global object shape. In: *PLOS Computational Biology* 14.12 (2018), e1006613.

[8] H. Bao, L. Dong, S. Piao, and F. Wei. BEiT: BERT Pre-Training of Image Transformers. In: *The Tenth International Conference on Learning Representations (ICLR)*. 2022.

[9] L. Beyer et al. FlexiViT: One Model for All Patch Sizes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 14496–14506.

[10] V. Biscione and J. S. Bowers. Mixed Evidence for Gestalt Grouping in Deep Neural Networks. In: *Computational Brain & Behavior* 6.3 (2023), pp. 438–456.

[11] D. Bolya et al. Perception Encoder: The best visual embeddings are not at the output of the network. In: *Advances in neural information processing systems*. Ed. by D. Belgrave et al. Vol. 38, Main Conference. Curran Associates, Inc., 2025, pp. 60884–60937.

[12] W. Brendel and M. Bethge. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In: *7th International Conference on Learning Representations (ICLR)*. 2019.

[13] T. Burgert, O. Stoll, P. Rota, and B. Demir. ImageNet-trained CNNs are not biased towards texture: Revisiting feature reliance through controlled suppression. In: *Advances in neural information processing systems*. Ed. by D. Belgrave et al. Vol. 38, Main Conference. Curran Associates, Inc., 2025, pp. 60809–60830.

[14] M. Caron et al. Emerging Properties in Self-Supervised Vision Transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 9650–9660.

[15] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu. Global Contrast Based Salient Region Detection. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.3 (2015), pp. 569–582.

[16] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing Textures in the Wild. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 3606–3613.

[17] P. Couairon, M. Shukor, J.-E. Haugeard, M. Cord, and N. Thome. DiffCut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. In: *Advances in neural information processing systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., 2024, pp. 13548–13578.

[18] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski. Vision transformers need registers. In: *The Twelfth International Conference on Learning Representations (ICLR)*. Ed. by B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun. 2024, pp. 2632–2652.

[19] M. Dehghani et al. Scaling Vision Transformers to 22 Billion Parameters. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 7480–7512.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[21] S. Dziadzio, Ç. Yıldız, G. M. van de Ven, T. Trzciński, T. Tuytelaars, and M. Bethge. Infinite dSprites for Disentangled Continual Learning: Separating Memory Edits from Generalization. In: *3rd Conference on Lifelong Learning Agents (CoLLAs)*. 2024.

[22] G. Ehrensperger, S. Stabinger, and A. R. Sánchez. Evaluating CNNs on the Gestalt Principle of Closure. In: *Artificial Neural Networks and Machine Learning – ICANN 2019: Theoretical Neural Computation*. Ed. by I. V. Tetko, V. Kůrková, P. Karpov, and F. Theis. Cham: Springer International Publishing, 2019, pp. 296–301.

[23] J. H. Elder and R. M. Goldberg. Ecological statistics of Gestalt laws for the perceptual organization of contours. In: *Journal of Vision* 2.4 (2002), p. 5.

[24] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao. EVA-02: A visual representation for neon genesis. In: *Image and Vision Computing* 149 (2024), p. 105171.

[25] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: *7th International Conference on Learning Representations (ICLR) 2019*. 2019.

[26] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

[27] R. Geirhos et al. Partial success in closing the gap between human and machine vision. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 23885–23899.

[28] W. S. Geisler, J. S. Perry, B. J. Super, and D. P. Gallogly. Edge co-occurrence in natural images predicts contour grouping performance. In: *Vision Research* 41.6 (2001), pp. 711–724.

[29] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman. *Unsupervised Semantic Segmentation by Distilling Feature Correspondences*. 2022. arXiv: 2203.08414.

[30] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked Autoencoders Are Scalable Vision Learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 16000–16009.

[31] G. Kanisza and W. Gerbino. "Convexity and symmetry in figure-ground organization". In: M. Henle. *Vision and artifact*. New York: Springer, 1976, pp. 25–32.

[32] B. Kim, E. Reif, M. Wattenberg, S. Bengio, and M. C. Mozer. Neural Networks Trained on Natural Scenes Exhibit Gestalt Closure. In: *Computational Brain & Behavior* 4.3 (2021), pp. 251–263.

[33] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In: *3rd International Conference on Learning Representations (ICLR) 2015*. arXiv, 2015.

[34] T. Li, I. Lin, L. Song, and T. S. Lee. *Learning Perceptual Rules in a Fragmented Visual World*. 2026. arXiv: 2610.13988.

[35] T. Li, Z. Wen, L. Song, J. Liu, Z. Jing, and T. S. Lee. *From Local Cues to Global Percepts: Emergent Gestalt Organization in Self-Supervised Vision Models*. 2025. arXiv: 2506.00718.

[36] Y. Li, S. Salehi, L. Ungar, and K. Kording. Does object binding naturally emerge in large pretrained vision transformers? In: *Advances in neural information processing systems*. Ed. by D. Belgrave et al. Vol. 38, Main Conference. Curran Associates, Inc., 2025, pp. 3394–3423.

[37] B. Lonnqvist, E. Scialom, A. Gokce, Z. Merchant, M. Herzog, and M. Schrimpf. Contour Integration Underlies Human-Like Vision. In: *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025, pp. 40290–40311.

[38] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi. Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 8364–8375.

[39] A. J. Mojica and M. A. Peterson. Display-wide influences on figure–ground perception: The case of symmetry. In: *Attention, Perception, & Psychophysics* 76.4 (2014), pp. 1069–1084.

[40] M. Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. In: *Transactions on Machine Learning Research* (2024).

[41] S. E. Palmer. *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press, 1999.

[42] M. A. Peterson. Object Recognition Processes Can and Do Operate Before Figure–Ground Organization. In: *Current Directions in Psychological Science* 3.4 (1994), pp. 105–111.

[43] M. A. Peterson and E. Salvagio. Inhibitory competition in figure-ground perception: Context and convexity. In: *Journal of Vision* 8.16 (2008), p. 4.

[44] A. Radford et al. Learning Transferable Visual Models From Natural Language Supervision. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763.

[45] S. S. Rambhatla, I. Misra, R. Chellappa, and A. Shrivastava. MOST: Multiple Object Localization with Self-Supervised Transformers for Object Discovery. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 15823–15834.

[46] S. Rawlekar, A. Swain, Y. Cai, Y. Wang, M.-H. Yang, and N. Ahuja. Finding Distributed Object-Centric Properties in Self-Supervised Transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2026, pp. 31326–31335.

[47] E. Rubin. *Visuell wahrgenommene Figuren: Studien in psychologischer Analyse*. Original Title: Synsoplevede figurer: Studier i psykologisk analyse. Copenhagen, Denmark: Gyldendalske Boghandel, 1921.

[48] J. Sha, K. Kersting, and D. Singh Dhami. Gestalt Vision: Evaluating Perceptual Grouping in Neural Visual Reasoning. In: *Neurosymbolic Artificial Intelligence* 2 (2026), p. 29498732261469488.

[49] O. Siméoni, C. Sekkat, G. Puy, A. Vobecký, É. Zablocki, and P. Pérez. Unsupervised Object Localization: Observing the Background To Discover Objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 3176–3186.

[50] O. Siméoni, É. Zablocki, S. Gidaris, G. Puy, and P. Pérez. Unsupervised Object Localization in the Era of Self-Supervised ViTs: A Survey. In: *International Journal of Computer Vision* 133.2 (2025), pp. 781–808.

[51] O. Siméoni et al. Localizing Objects with Self-Supervised Transformers and no Labels. In: *BMVC*. 2021.

[52] O. Siméoni et al. DINOv3. In: *Transactions on Machine Learning Research* (2026).

[53] A. P. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. In: *Transactions on Machine Learning Research* (2022).

[54] K. A. Stevens and A. Brookes. The Concave Cusp as a Determiner of Figure—Ground. In: *Perception* 17.1 (1988), pp. 35–42.

[55] L. Tang and D. Ley. Degraded Polygons Raise Fundamental Questions of Neural Network Perception. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Datasets and Benchmarks Track. Curran Associates, Inc., 2023, pp. 9695–9706.

[56] A. R. Tartaglini and M. A. Lepori. *I Walk the Line: Examining the Role of Gestalt Continuity in Object Binding for Vision Transformers*. 2026. arXiv: 2604.09942.

[57] H. Touvron, M. Cord, and H. Jégou. DeiT III: Revenge of the ViT. In: *Computer Vision – ECCV 2022*. Ed. by S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner. Cham: Springer Nature Switzerland, 2022, pp. 516–533.

[58] M. Tschannen et al. *SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features*. 2025. arXiv: 2502.14786.

[59] J. Wagemans et al. A Century of Gestalt Psychology in Visual Perception: I. Perceptual Grouping and Figure–Ground Organization. In: *Psychological Bulletin* 138.6 (2012), pp. 1172–1217.

[60] W. Wang et al. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 19175–19186.

[61] X. Wang, R. Girdhar, S. X. Yu, and I. Misra. Cut and Learn for Unsupervised Object Detection and Instance Segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 3124–3134.

[62] X. Wang, J. Yang, and T. Darrell. Segment Anything without Supervision. In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., 2024, pp. 138731–138755.

[63] Y. Wang, X. Shen, S. X. Hu, Y. Yuan, J. L. Crowley, and D. Vaufreydaz. Self-Supervised Transformers for Unsupervised Object Discovery Using Normalized Cut. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 14543–14553.

[64] F. A. Wichmann and R. Geirhos. Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception? In: *Annual Review of Vision Science* 9 (2023), pp. 501–524.

[65] R. Wightman. *PyTorch Image Models*. 2019.

[66] J. M. Wolfe et al. *Sensation & perception*. Sixth edition. New York, NY, USA: Oxford University Press, 2021. 522 pp.

[67] Z. Wu, A. Moemeni, and P. Caleb-Solly. Ensemble Foreground Management for Unsupervised Object Discovery. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025, pp. 20268–20279.

[68] J. Yu, T. Darrell, and X. Wang. UnSAMv2: Self-Supervised Learning Enables Segment Anything at Any Granularity. In: *Transactions on Machine Learning Research* (2026).

[69] X. Zhang, J. Xie, Y. Yuan, M. B. Mi, and R. T. Tan. HEAP: Unsupervised Object Discovery and Localization with Contrastive Grouping. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 7. 2024, pp. 7323–7331.

[70] Y. Zhang, D. Soydaner, F. Behrad, L. Koßmann, and J. Wagemans. Investigating the Gestalt Principle of Closure in Deep Convolutional Neural Networks. In: *32nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning Bruges, Belgium October 09 - 11*. i6docs.com, 2024, pp. 679–684.

[71] Y. Zhang, D. Soydaner, L. Koßmann, F. Behrad, and J. Wagemans. Finding Closure: A Closer Look at the Gestalt Law of Closure in Convolutional Neural Networks. In: *Computational Brain & Behavior* (2025).

[72] J. Zhou et al. Image BERT Pre-training with Online Tokenizer. In: *The Tenth International Conference on Learning Representations (ICLR)*. 2022.

A Detailed results

We report the following evaluation metrics in the extended results:

- **Accuracy (ACC)** is computed by thresholding the predicted foreground probabilities for each patch at 0.5 and comparing to the binary ground truth.
- **Binary cross entropy (BCE)** between the predicted foreground probabilities and the binary ground truth.
- **Information gain (IG)** is defined as the reduction of the binary cross entropy relative to the spatial prior:
$$\text{IG} = \text{BCE}_{prior} - \text{BCE}_{probe}$$
- **Information gain explained (IGE)** is the fraction of the maximum possible information gain achieved by the respective probe.
$$\text{IGE} = \frac{\text{IG}}{\text{BCE}_{prior}} = 1 - \frac{\text{BCE}_{probe}}{\text{BCE}_{prior}}$$
An IGE of 0% indicates that the probe performs no better than the spatial prior, while an IGE of 100% indicates perfect prediction.
- **Mean intersection over union (mIoU)** is computed for the predictions thresholded at 0.5. For each sample, we compute the IoU between the predicted and target foreground and background regions, respectively, and average their scores.

The region property baselines predict hard binary masks so that the binary cross entropy is not finite for incorrect predictions. We therefore use accuracy as the main metric throughout the paper.

A.1 Natural stimuli



Figure 8: Example images from the natural test set (MSRA-10K, [15])

Table 1: Probe performances on the natural test set.

Model	ACC ↓	BCE	IG	IGE	mIoU
DeiT-3-L	0.982	0.051	0.360	0.876	0.951
BEiT-3-L	0.981	0.054	0.358	0.870	0.950
MAE-L	0.981	0.054	0.357	0.869	0.951
BEiT-L	0.980	0.056	0.356	0.865	0.948
MAE-B	0.979	0.058	0.354	0.860	0.947
DeiT-3-B	0.979	0.059	0.352	0.856	0.946
BEiT-B	0.978	0.060	0.351	0.853	0.945
EVA-02-L	0.978	0.061	0.359	0.855	0.945
BEiT-3-B	0.977	0.062	0.349	0.848	0.941
DINO-B	0.975	0.068	0.344	0.835	0.939
EVA-02-B	0.974	0.071	0.349	0.831	0.937
DINOv3-L	0.974	0.072	0.348	0.829	0.935
SigLIP-2-L	0.974	0.075	0.345	0.821	0.934
DINOv3-B	0.973	0.074	0.345	0.823	0.935
FlexiViT-L	0.973	0.073	0.343	0.824	0.931
PE-B	0.972	0.078	0.376	0.829	0.937
AugReg-B	0.972	0.080	0.331	0.806	0.930
AugReg-L	0.971	0.082	0.330	0.801	0.930
PE-L	0.971	0.081	0.372	0.821	0.934
CLIP-L	0.970	0.080	0.341	0.810	0.926
DINOv2-L	0.968	0.088	0.370	0.809	0.926
CLIP-B	0.966	0.095	0.317	0.770	0.915
SigLIP-2-B	0.964	0.104	0.307	0.746	0.908
DINOv2-B	0.962	0.102	0.356	0.777	0.913
FlexiViT-B	0.958	0.112	0.304	0.731	0.898
Spatial prior	0.798	0.433	0.000	0.000	0.571
Area	0.924	–	–	–	–
Center overlap	0.746	–	–	–	–
Offset	0.922	–	–	–	–

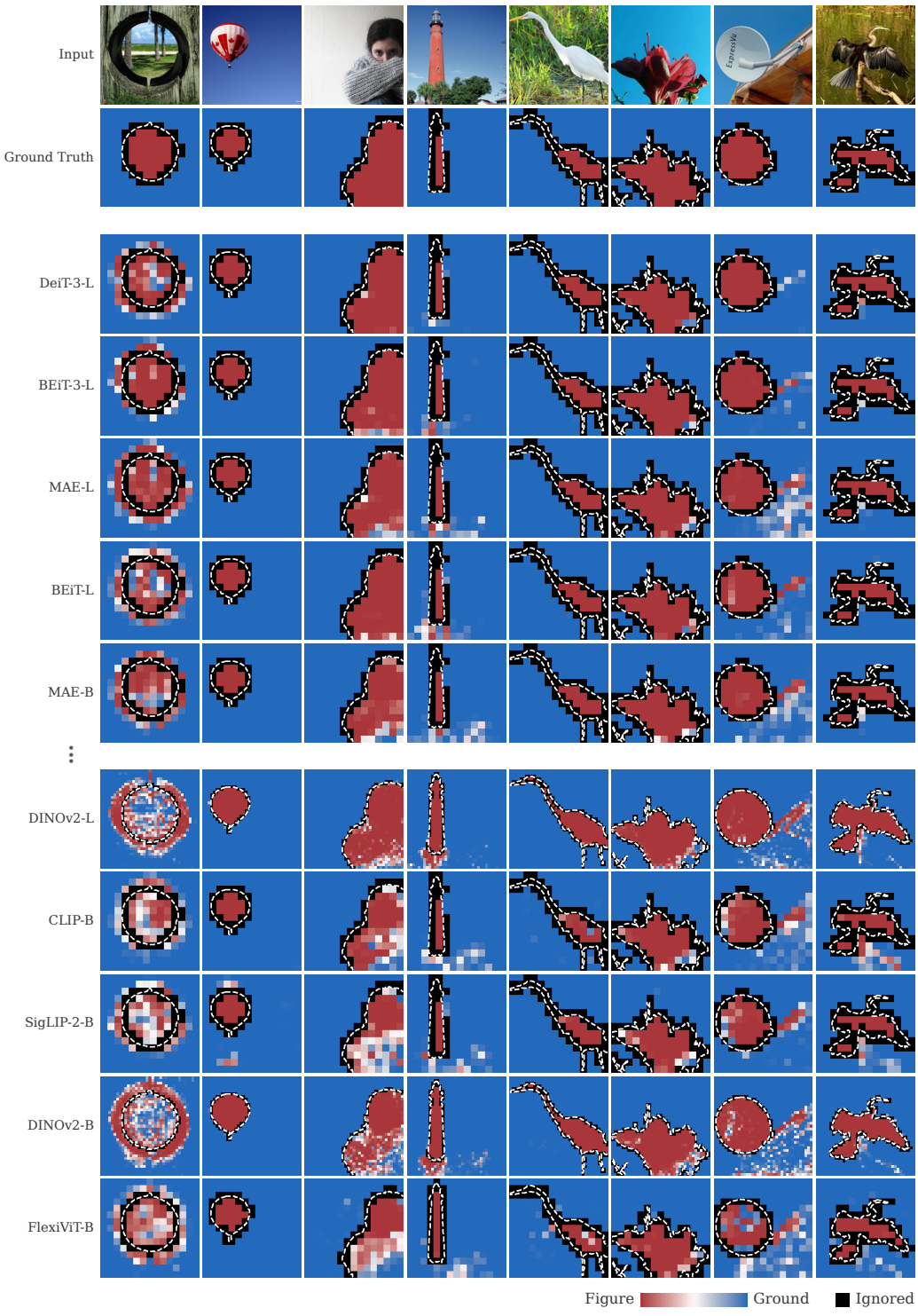

Figure 9: Probe predictions for the best and worst models in the natural condition.

A.2 Texturized stimuli



Figure 10: Example images from the texturized test set. The same segmentation masks are used as in the natural condition, but filled with random textures or colors.

Table 2: Probe performances on the texturized test set, averaged over all six variants (texture and color variants of the original, reversed, and split stimuli).

	Trained on natural					Trained on texturized				
Model	ACC ↓	BCE	IG	IGE	mIoU	ACC	BCE	IG	IGE	mIoU
BEiT-3-L	0.986	0.044	0.367	0.893	0.963	0.997	0.011	0.400	0.973	0.990
EVA-02-L	0.975	0.070	0.350	0.833	0.941	0.992	0.023	0.397	0.945	0.979
PE-B	0.968	0.093	0.361	0.796	0.933	0.991	0.025	0.429	0.945	0.980
DINOv3-L	0.967	0.092	0.328	0.782	0.925	0.995	0.016	0.404	0.962	0.985
CLIP-L	0.966	0.093	0.327	0.779	0.923	0.987	0.039	0.381	0.906	0.969
PE-L	0.965	0.103	0.351	0.773	0.925	0.993	0.022	0.432	0.952	0.984
BEiT-3-B	0.963	0.097	0.314	0.764	0.914	0.995	0.017	0.395	0.960	0.984
CLIP-B	0.960	0.118	0.293	0.712	0.906	0.983	0.054	0.357	0.868	0.958
MAE-B	0.959	0.113	0.298	0.725	0.914	0.987	0.037	0.375	0.911	0.968
DINOv2-L	0.958	0.116	0.342	0.746	0.916	0.990	0.030	0.428	0.935	0.976
SigLIP-2-L	0.957	0.115	0.304	0.725	0.913	0.988	0.039	0.381	0.908	0.970
DINOv3-B	0.957	0.121	0.298	0.710	0.907	0.991	0.025	0.395	0.941	0.978
BEiT-L	0.954	0.119	0.292	0.711	0.897	0.992	0.023	0.389	0.944	0.980
EVA-02-B	0.951	0.122	0.298	0.709	0.894	0.987	0.037	0.384	0.913	0.968
MAE-L	0.947	0.131	0.281	0.683	0.890	0.992	0.025	0.387	0.940	0.980
DeiT-3-L	0.946	0.152	0.260	0.632	0.887	0.995	0.015	0.396	0.963	0.987
DeiT-3-B	0.940	0.162	0.249	0.606	0.869	0.993	0.023	0.389	0.945	0.981
DINO-B	0.939	0.149	0.263	0.639	0.879	0.988	0.034	0.377	0.917	0.970
SigLIP-2-B	0.936	0.182	0.229	0.557	0.863	0.976	0.073	0.338	0.822	0.936
DINOv2-B	0.933	0.164	0.294	0.642	0.867	0.983	0.048	0.410	0.895	0.960
BEiT-B	0.928	0.176	0.236	0.573	0.840	0.987	0.039	0.372	0.905	0.966
AugReg-L	0.909	0.244	0.167	0.406	0.806	0.976	0.071	0.340	0.828	0.939
AugReg-B	0.898	0.277	0.134	0.326	0.776	0.972	0.078	0.334	0.811	0.928
FlexiViT-L	0.895	0.256	0.159	0.383	0.771	0.983	0.053	0.362	0.872	0.958
FlexiViT-B	0.858	0.350	0.066	0.159	0.683	0.964	0.104	0.312	0.751	0.913
Spatial prior	0.798	0.433	0.000	0.000	0.571	0.798	0.433	0.000	0.000	0.571
Area	0.944	–	–	–	–	0.944	–	–	–	–
Center overlap	0.715	–	–	–	–	0.715	–	–	–	–
Offset	0.901	–	–	–	–	0.901	–	–	–	–

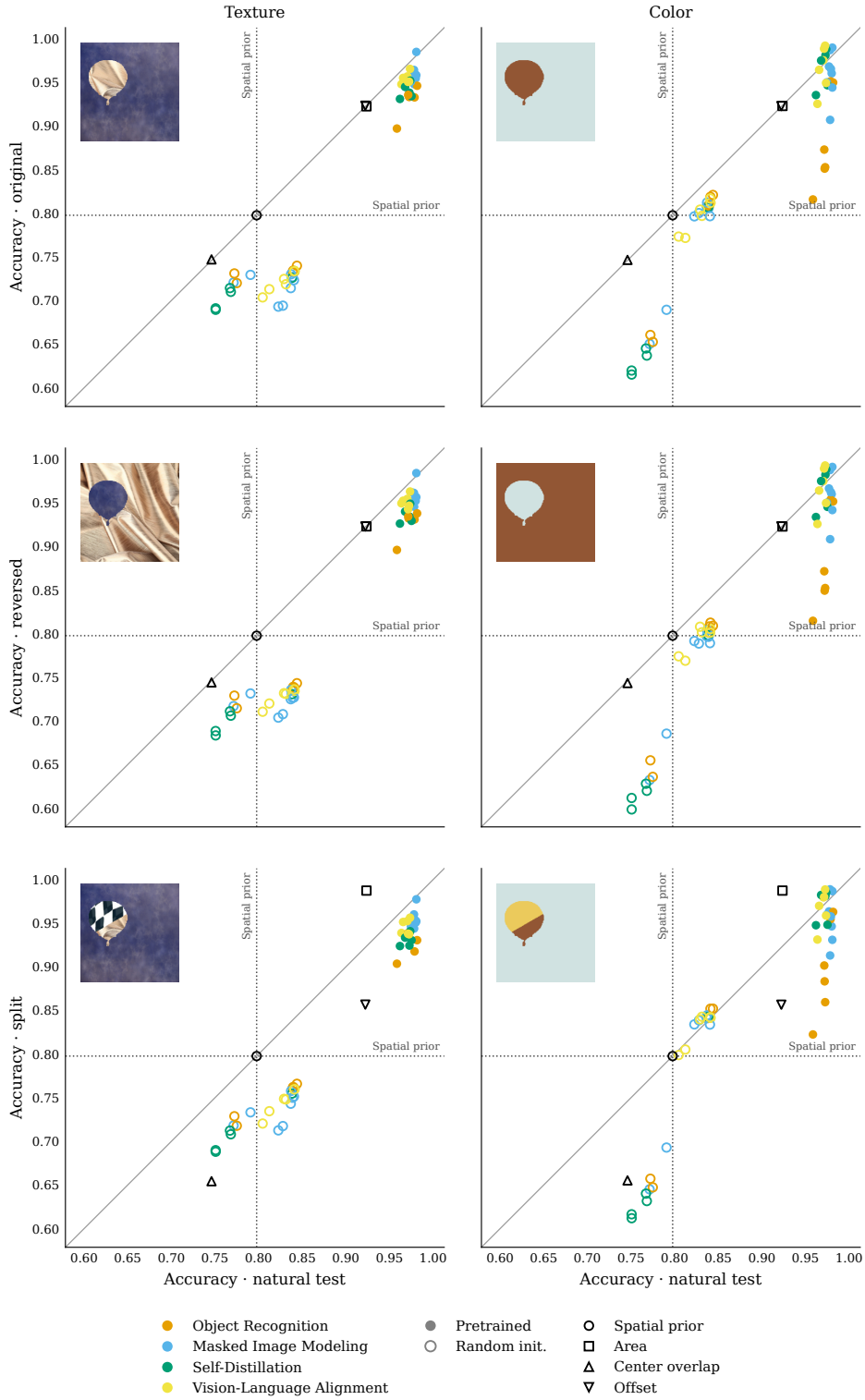

Figure 11: Comparison of the probe performances on the i.i.d. test set of natural images (x-axis) and applied zero-shot to texturized stimulus variants (y-axis).

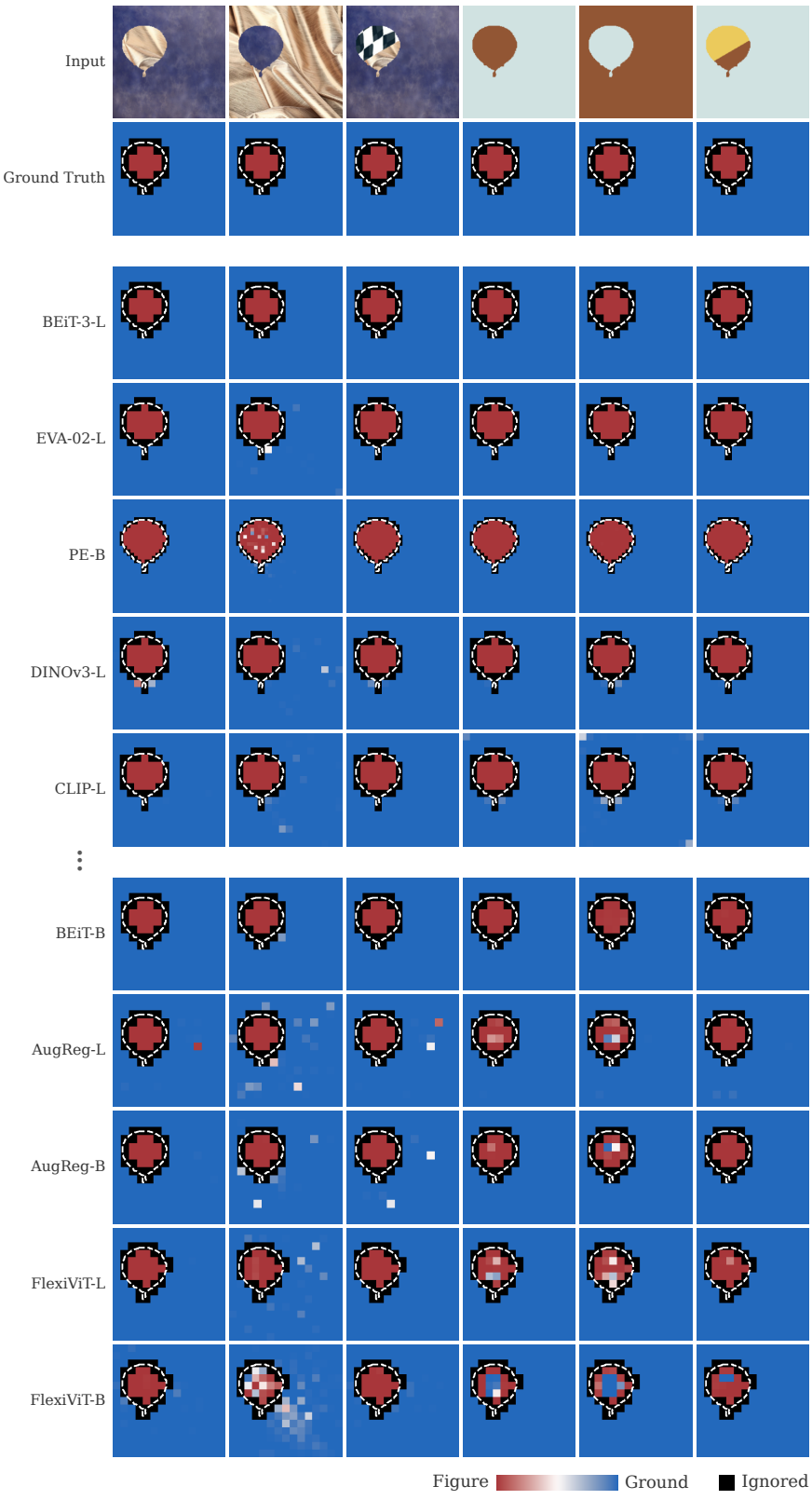

Figure 12: Probe predictions for the best and worst models in the texturized condition.

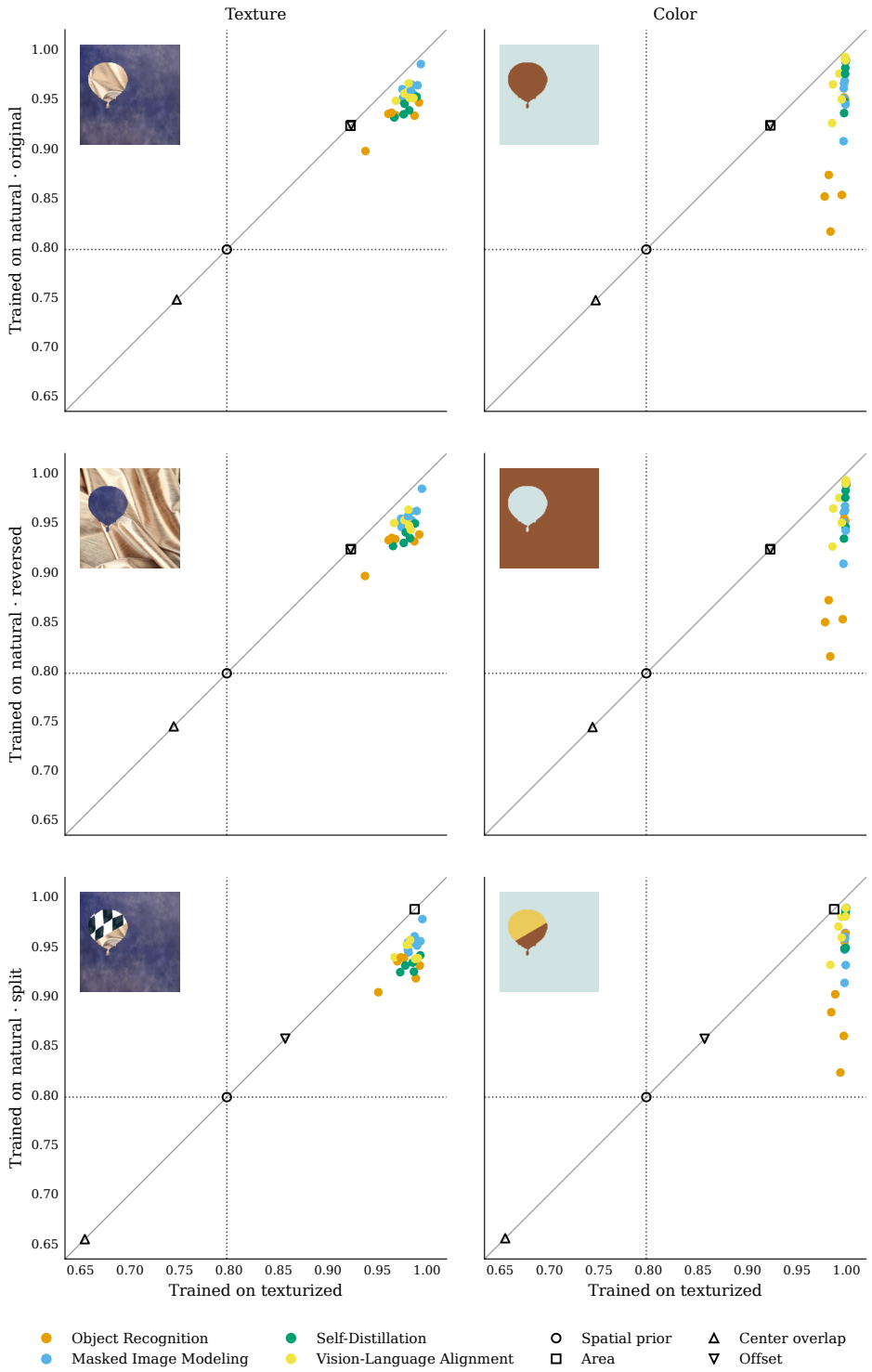

Figure 13: Performance comparison of probes directly trained on the texturized images (x-axis) with zero-shot probes trained on natural images (y-axis).

A.3 Surroundedness



Figure 14: Example images from the surroundedness test set. Each image contains a random shape from the idSprites dataset [21], surroundend by the background on all sides.

Table 3: Probe performances on the surroundedness test set, averaged over all six variants (texture and color variants of the original, reversed, and split stimuli).

	Trained on natural					Trained on surroundedness				
Model	ACC ↓	BCE	IG	IGE	mIoU	ACC	BCE	IG	IGE	mIoU
BEiT-3-L	0.986	0.051	0.439	0.897	0.970	0.999	0.004	0.419	0.990	0.996
EVA-02-L	0.981	0.066	0.430	0.867	0.958	0.995	0.014	0.418	0.967	0.988
PE-B	0.975	0.073	0.447	0.859	0.952	0.996	0.014	0.453	0.970	0.990
DINOv3-L	0.968	0.096	0.400	0.806	0.935	0.997	0.009	0.424	0.979	0.991
PE-L	0.965	0.095	0.425	0.817	0.931	0.997	0.009	0.458	0.981	0.993
DINOv3-B	0.963	0.110	0.386	0.779	0.928	0.995	0.014	0.419	0.967	0.988
CLIP-L	0.961	0.101	0.396	0.797	0.922	0.992	0.032	0.401	0.926	0.980
DINOv2-L	0.961	0.118	0.405	0.774	0.924	0.992	0.024	0.447	0.949	0.980
BEiT-L	0.956	0.110	0.380	0.775	0.916	0.996	0.011	0.412	0.973	0.989
DINO-B	0.951	0.126	0.364	0.743	0.912	0.992	0.024	0.400	0.944	0.979
CLIP-B	0.950	0.151	0.338	0.691	0.900	0.989	0.051	0.372	0.880	0.973
EVA-02-B	0.950	0.118	0.379	0.762	0.904	0.991	0.027	0.405	0.937	0.977
MAE-B	0.950	0.122	0.368	0.752	0.911	0.992	0.023	0.400	0.945	0.981
BEiT-3-B	0.948	0.120	0.370	0.755	0.903	0.997	0.010	0.413	0.976	0.989
DeiT-3-B	0.943	0.141	0.349	0.712	0.887	0.996	0.013	0.410	0.969	0.988
DeiT-3-L	0.940	0.136	0.354	0.723	0.891	0.998	0.007	0.416	0.983	0.994
MAE-L	0.931	0.150	0.340	0.694	0.881	0.996	0.014	0.409	0.967	0.989
SigLIP-2-L	0.919	0.166	0.330	0.665	0.865	0.993	0.026	0.407	0.941	0.982
DINOv2-B	0.918	0.179	0.345	0.658	0.852	0.988	0.036	0.435	0.923	0.970
SigLIP-2-B	0.907	0.263	0.227	0.463	0.833	0.986	0.051	0.373	0.880	0.961
BEiT-B	0.899	0.211	0.279	0.569	0.822	0.991	0.027	0.396	0.935	0.977
AugReg-L	0.867	0.384	0.105	0.215	0.767	0.984	0.052	0.371	0.876	0.958
AugReg-B	0.865	0.394	0.095	0.195	0.764	0.977	0.069	0.355	0.837	0.941
FlexiViT-L	0.827	0.414	0.080	0.162	0.701	0.988	0.044	0.384	0.896	0.969
FlexiViT-B	0.787	0.504	-0.010	-0.020	0.616	0.967	0.109	0.319	0.745	0.919
Spatial prior	0.721	0.506	0.000	0.000	0.479	0.760	0.446	0.000	0.000	0.582
Area	0.660	–	–	–	–	0.660	–	–	–	–
Center overlap	0.535	–	–	–	–	0.535	–	–	–	–
Offset	0.598	–	–	–	–	0.598	–	–	–	–

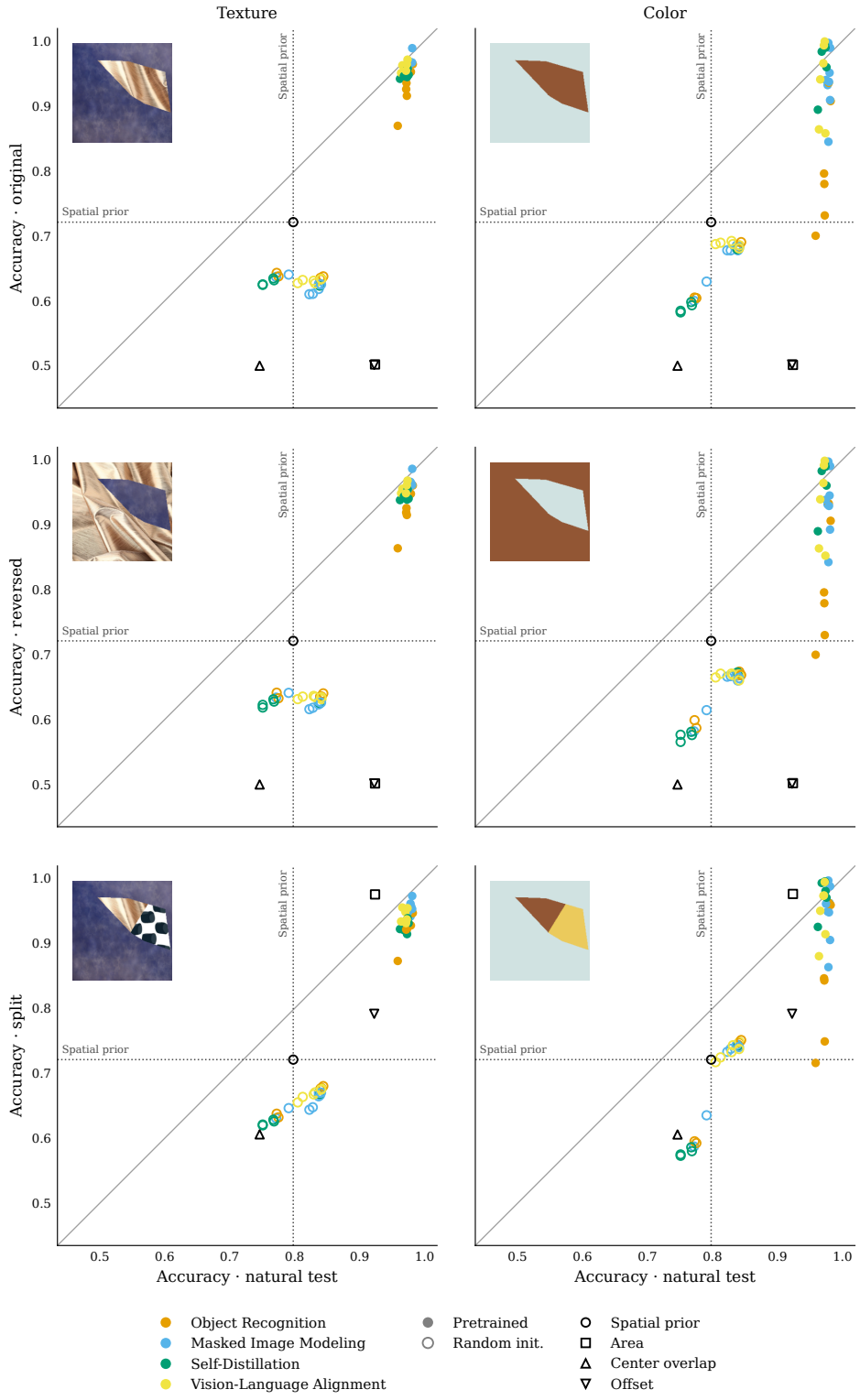

Figure 15: Comparison of the probe performances on the i.i.d. test set of natural images (x-axis) and applied zero-shot to the stimuli isolating surroundedness (y-axis).

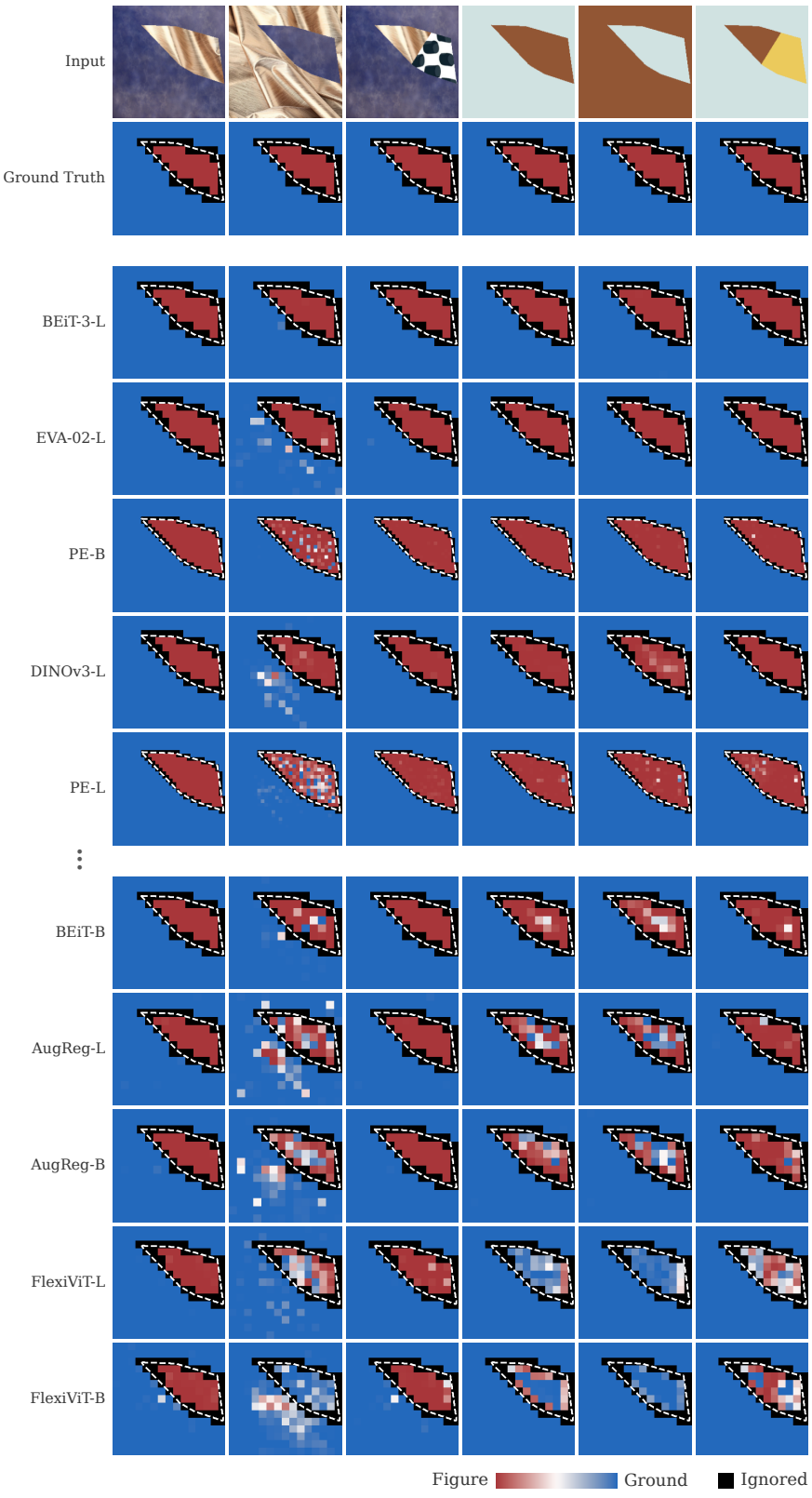

Figure 16: Probe predictions for the best and worst models in the surroundedness condition.

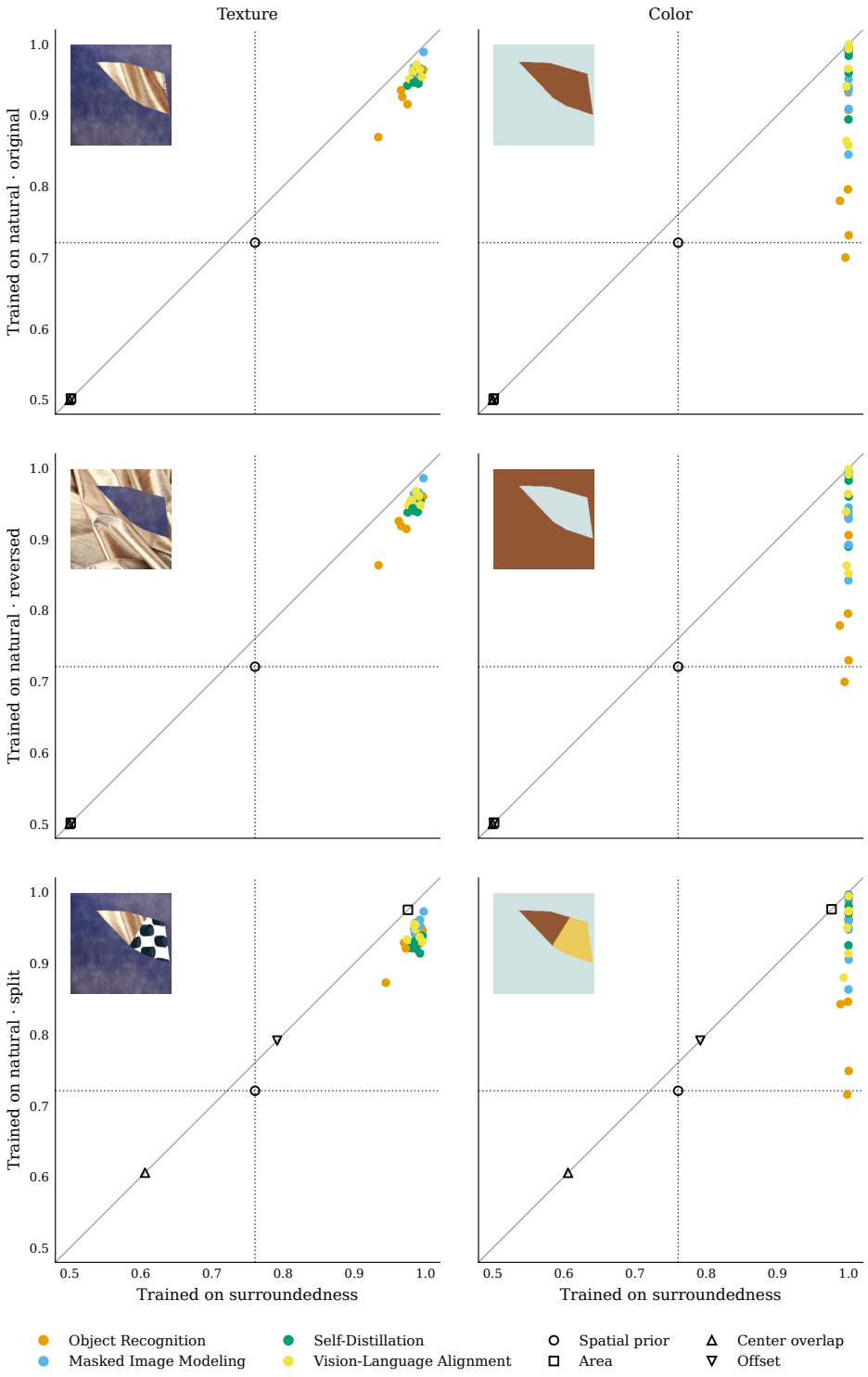

Figure 17: Performance comparison of probes directly trained on the surroundedness stimuli (x-axis) with zero-shot probes trained on natural images (y-axis).

A.4 Convexity



Figure 18: Example images from the convexity condition. Each image contains two regions separated by a random parabolic arc, filled with random textures or colors.

Table 4: Probe performances on the convexity test sets, averaged over all six variants (texture and color variants of the original, reversed, and split stimuli).

	Trained on natural					Trained on convexity				
Model	ACC ↓	BCE	IG	IGE	mIoU	ACC	BCE	IG	IGE	mIoU
BEiT-3-L	0.730	0.683	0.359	0.344	0.593	0.977	0.072	0.604	0.894	0.952
EVA-02-L	0.703	0.711	0.313	0.305	0.547	0.953	0.131	0.545	0.806	0.909
PE-B	0.699	0.945	0.024	0.025	0.580	0.952	0.127	0.549	0.812	0.914
CLIP-L	0.685	0.810	0.214	0.209	0.533	0.931	0.178	0.498	0.737	0.855
DINOv2-L	0.677	0.732	0.229	0.238	0.516	0.944	0.140	0.536	0.793	0.897
CLIP-B	0.675	0.728	0.314	0.301	0.509	0.915	0.234	0.441	0.653	0.836
DINO-B	0.672	0.881	0.161	0.154	0.532	0.964	0.094	0.581	0.860	0.931
BEiT-L	0.659	0.988	0.053	0.051	0.509	0.963	0.099	0.576	0.853	0.926
PE-L	0.652	1.087	-0.118	-0.122	0.525	0.962	0.105	0.571	0.844	0.930
MAE-B	0.644	0.991	0.051	0.049	0.480	0.919	0.199	0.477	0.706	0.864
DINOv2-B	0.642	0.950	0.010	0.011	0.462	0.927	0.177	0.499	0.738	0.870
DINOv3-L	0.640	1.085	-0.060	-0.059	0.459	0.968	0.088	0.588	0.870	0.935
BEiT-B	0.632	1.137	-0.095	-0.092	0.460	0.935	0.173	0.503	0.744	0.879
EVA-02-B	0.631	1.089	-0.065	-0.064	0.445	0.924	0.199	0.476	0.705	0.856
BEiT-3-B	0.631	1.081	-0.039	-0.038	0.447	0.963	0.111	0.565	0.836	0.925
SigLIP-2-L	0.627	1.034	-0.009	-0.008	0.448	0.925	0.205	0.470	0.696	0.854
AugReg-L	0.626	1.233	-0.191	-0.184	0.453	0.889	0.259	0.416	0.616	0.771
MAE-L	0.622	1.133	-0.091	-0.088	0.447	0.939	0.155	0.521	0.771	0.895
DeiT-3-L	0.618	1.413	-0.372	-0.357	0.418	0.972	0.081	0.594	0.880	0.939
SigLIP-2-B	0.612	1.121	-0.079	-0.076	0.414	0.879	0.283	0.392	0.580	0.761
AugReg-B	0.608	1.302	-0.261	-0.250	0.419	0.893	0.251	0.424	0.628	0.777
DINOv3-B	0.607	1.204	-0.178	-0.174	0.428	0.958	0.110	0.566	0.837	0.921
DeiT-3-B	0.604	1.324	-0.283	-0.271	0.391	0.958	0.111	0.564	0.835	0.909
FlexiViT-L	0.597	1.386	-0.354	-0.342	0.399	0.947	0.167	0.509	0.753	0.890
FlexiViT-B	0.555	1.354	-0.321	-0.311	0.338	0.902	0.273	0.402	0.596	0.811
Spatial prior	0.531	1.003	0.000	0.000	0.308	0.582	0.676	0.000	0.000	0.397
Area	0.548	–	–	–	–	0.548	–	–	–	–
Center overlap	0.550	–	–	–	–	0.550	–	–	–	–
Offset	0.506	–	–	–	–	0.506	–	–	–	–

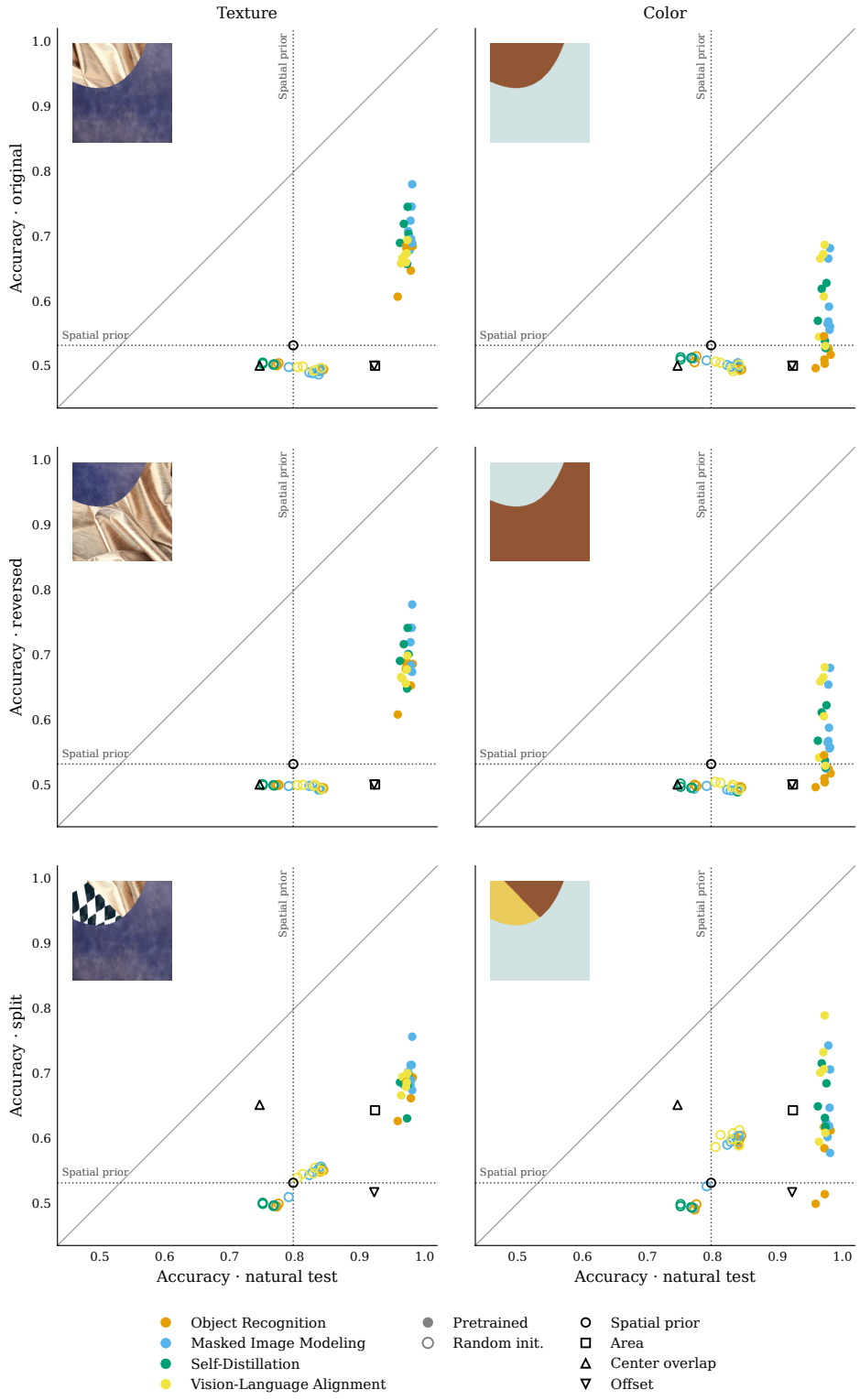

Figure 19: Comparison of the probe performances on the i.i.d. test set of natural images (x-axis) and applied zero-shot to the stimuli isolating convexity (y-axis).

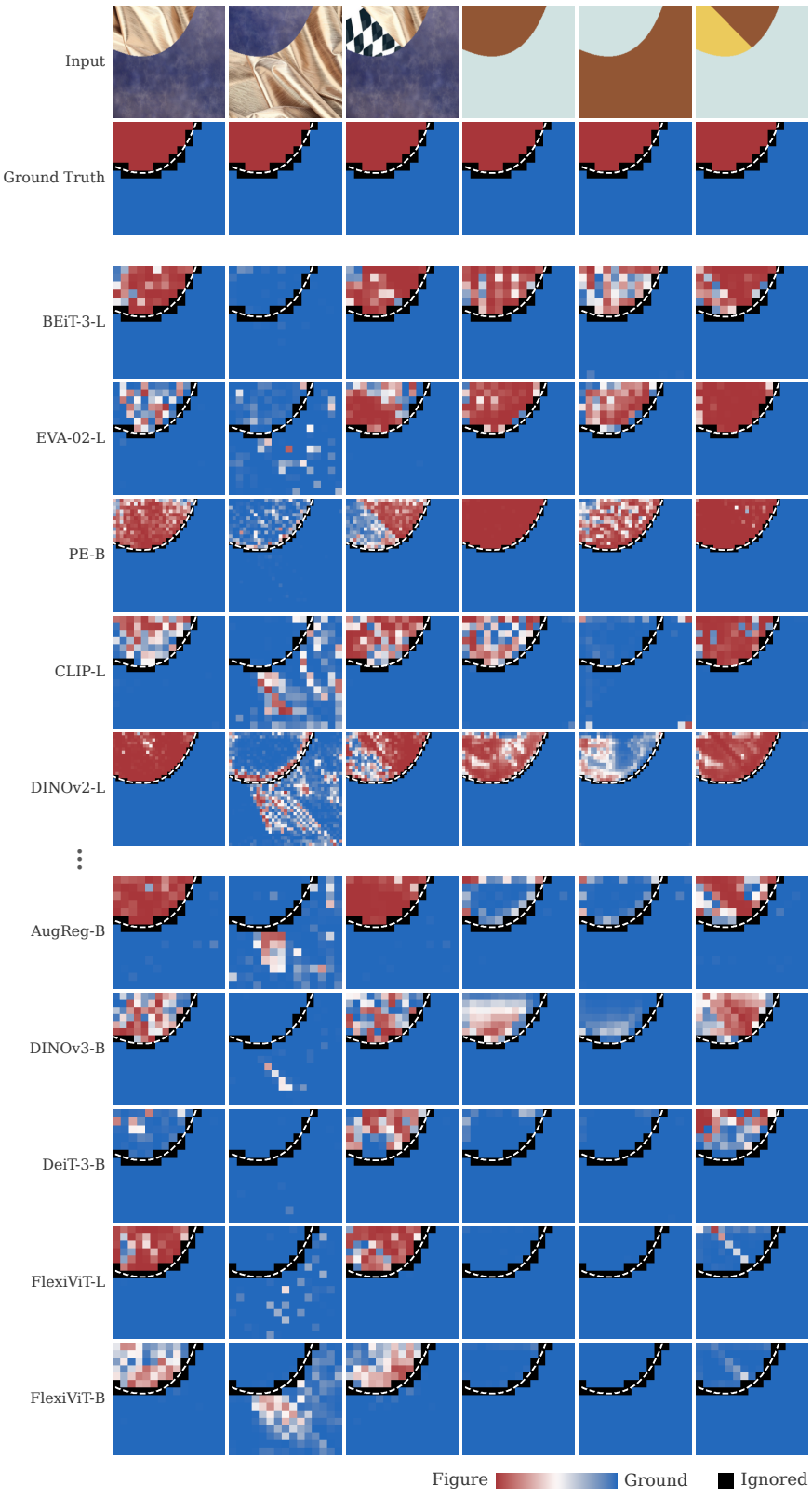

Figure 20: Probe predictions for the best and worst models in the convexity condition.

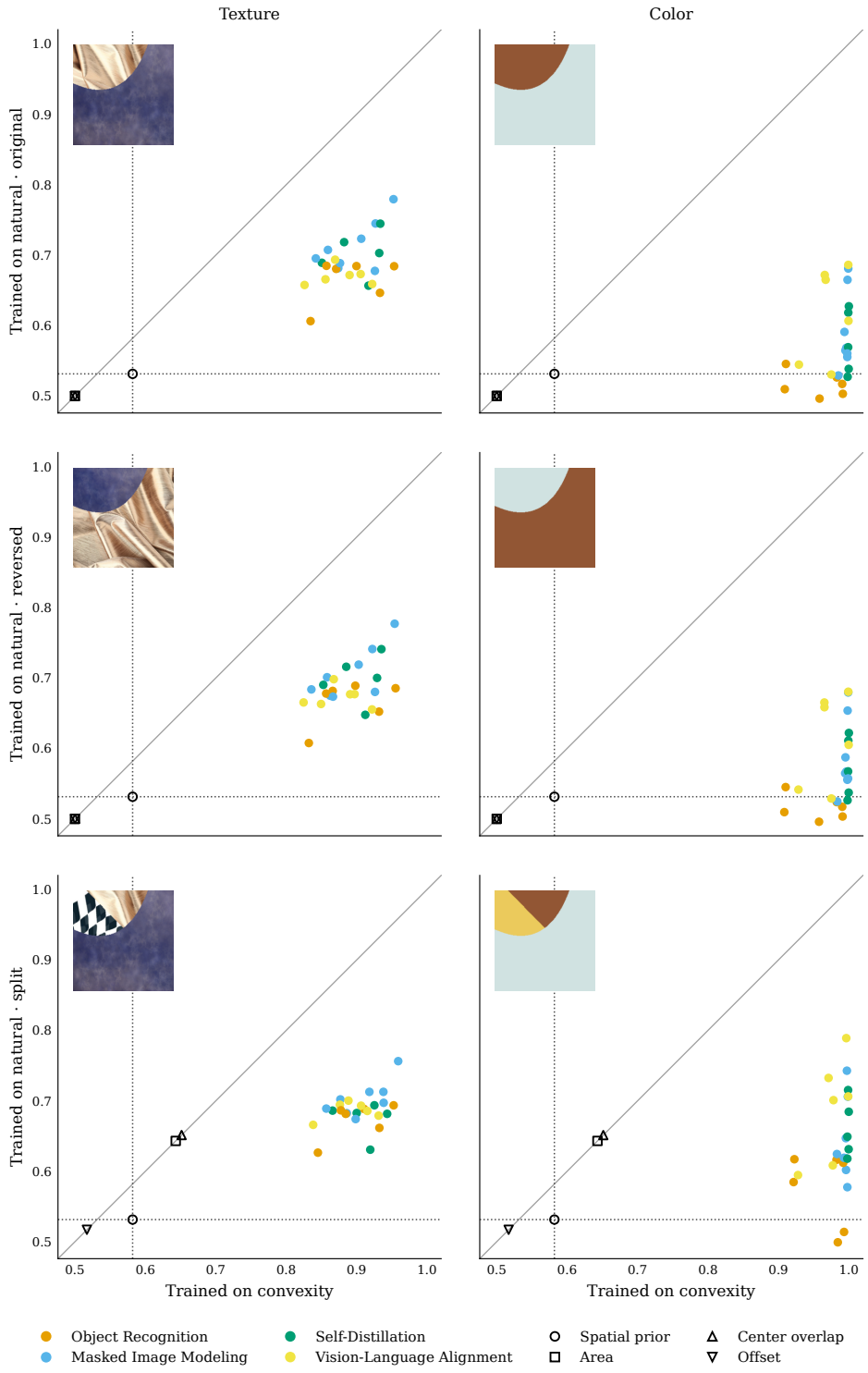

Figure 21: Performance comparison of probes directly trained on the convexity stimuli (x-axis) with zero-shot probes trained on natural images (y-axis).

A.5 Symmetry



Figure 22: Example images from the symmetry condition. Each image is divided into columns, with one of the columns having mirror symmetric contours.

Table 5: Probe performances on the symmetry test sets, averaged over all four variants (texture and color variants of the original and reversed stimuli).

	Trained on natural					Trained on symmetry				
Model	ACC ↓	BCE	IG	IGE	mIoU	ACC	BCE	IG	IGE	mIoU
PE-B	0.567	1.781	-0.700	-0.648	0.425	0.723	0.537	0.157	0.226	0.625
DINOv2-L	0.552	1.217	-0.147	-0.137	0.394	0.818	0.372	0.321	0.463	0.739
PE-L	0.543	1.747	-0.667	-0.618	0.359	0.776	0.455	0.239	0.344	0.697
DINOv3-L	0.532	1.720	-0.569	-0.494	0.351	0.800	0.394	0.300	0.432	0.719
EVA-02-L	0.531	1.573	-0.423	-0.368	0.368	0.745	0.547	0.147	0.211	0.640
CLIP-L	0.528	1.448	-0.299	-0.260	0.353	0.726	0.537	0.156	0.225	0.606
CLIP-B	0.526	1.253	-0.080	-0.068	0.340	0.666	0.605	0.089	0.128	0.522
DINOv3-B	0.525	1.766	-0.615	-0.534	0.328	0.772	0.447	0.246	0.355	0.687
DINOv2-B	0.521	1.459	-0.389	-0.363	0.345	0.745	0.497	0.197	0.284	0.639
SigLIP-2-L	0.521	1.566	-0.414	-0.360	0.326	0.746	0.529	0.165	0.237	0.632
DeiT-3-B	0.518	1.876	-0.703	-0.600	0.317	0.653	0.609	0.085	0.122	0.504
BEiT-L	0.514	1.611	-0.438	-0.374	0.326	0.792	0.403	0.291	0.420	0.714
DeiT-3-L	0.511	2.128	-0.955	-0.814	0.305	0.742	0.527	0.166	0.240	0.632
SigLIP-2-B	0.510	1.575	-0.402	-0.343	0.305	0.651	0.619	0.075	0.108	0.500
BEiT-B	0.510	1.673	-0.500	-0.426	0.314	0.743	0.521	0.173	0.249	0.638
EVA-02-B	0.509	1.911	-0.762	-0.663	0.315	0.693	0.586	0.107	0.155	0.559
BEiT-3-B	0.508	2.009	-0.836	-0.713	0.303	0.760	0.494	0.200	0.288	0.666
DINO-B	0.507	1.824	-0.651	-0.555	0.302	0.648	0.640	0.054	0.078	0.502
FlexiViT-L	0.506	1.976	-0.814	-0.700	0.290	0.680	0.612	0.082	0.118	0.538
AugReg-L	0.505	1.760	-0.587	-0.501	0.304	0.659	0.606	0.087	0.126	0.509
AugReg-B	0.505	1.756	-0.583	-0.497	0.297	0.607	0.655	0.039	0.056	0.448
MAE-B	0.505	2.344	-1.171	-0.999	0.300	0.758	0.502	0.192	0.276	0.656
FlexiViT-B	0.504	1.723	-0.560	-0.482	0.281	0.610	0.668	0.026	0.037	0.451
MAE-L	0.502	2.371	-1.198	-1.021	0.296	0.794	0.443	0.250	0.361	0.710
BEiT-3-L	0.501	1.657	-0.484	-0.412	0.338	0.802	0.431	0.262	0.378	0.722
Spatial prior	0.501	1.124	0.000	0.000	0.306	0.476	0.694	0.000	0.000	0.437
Area	0.523	–	–	–	–	0.523	–	–	–	–
Center overlap	0.506	–	–	–	–	0.506	–	–	–	–
Offset	0.523	–	–	–	–	0.523	–	–	–	–

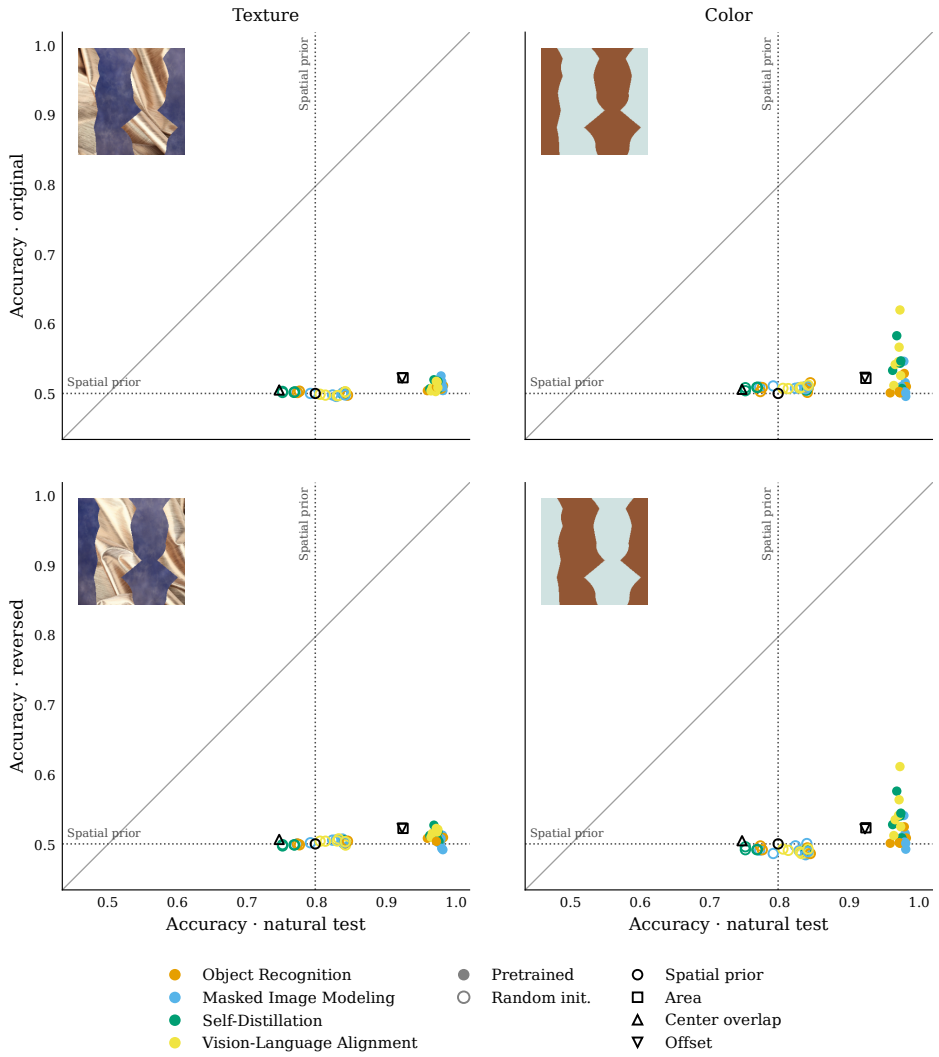

Figure 23: Comparison of the probe performances on the i.i.d. test set of natural images (x-axis) and applied zero-shot to the stimuli isolating symmetry (y-axis).

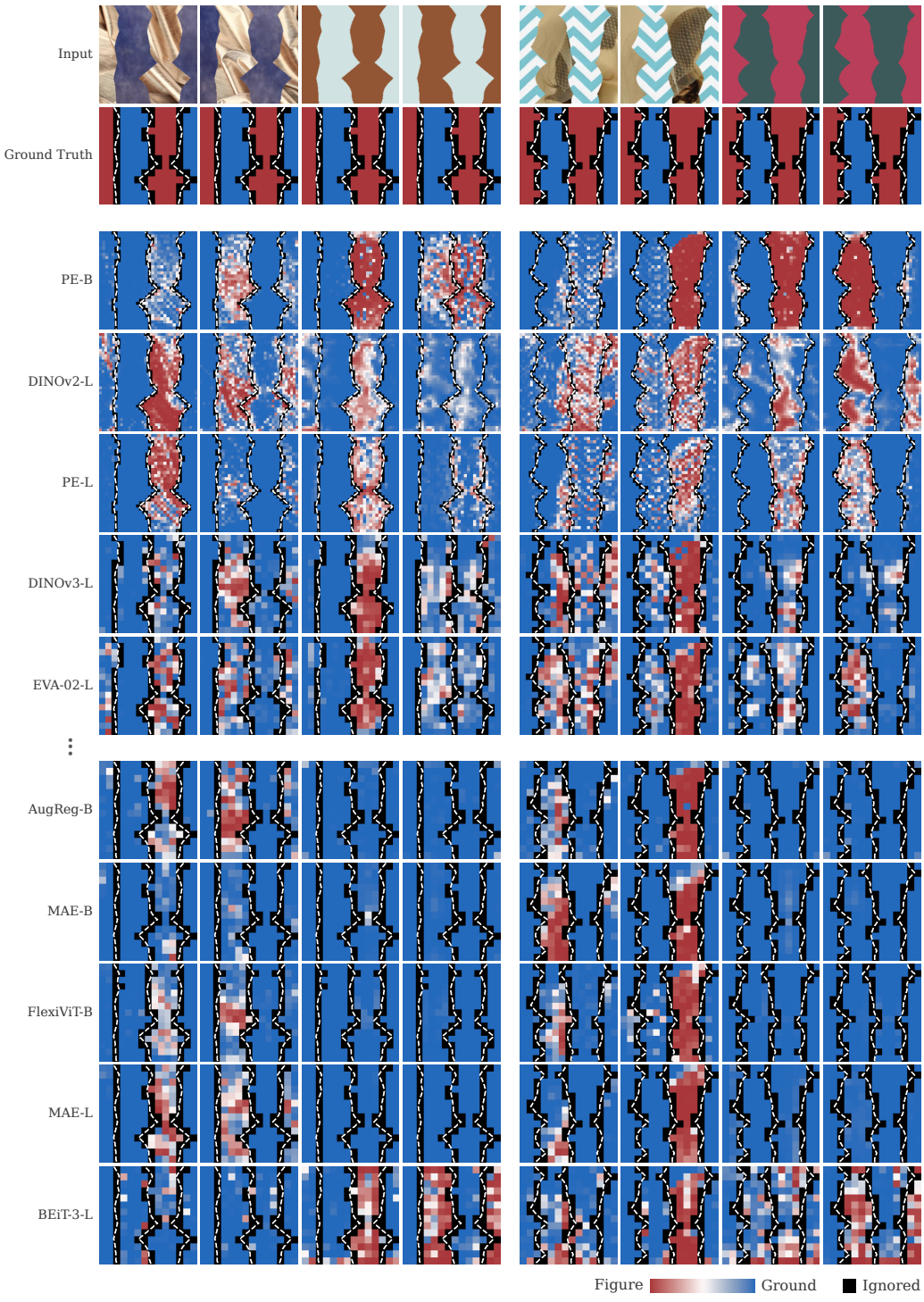

Figure 24: Probe predictions for the best and worst models in the symmetry condition.

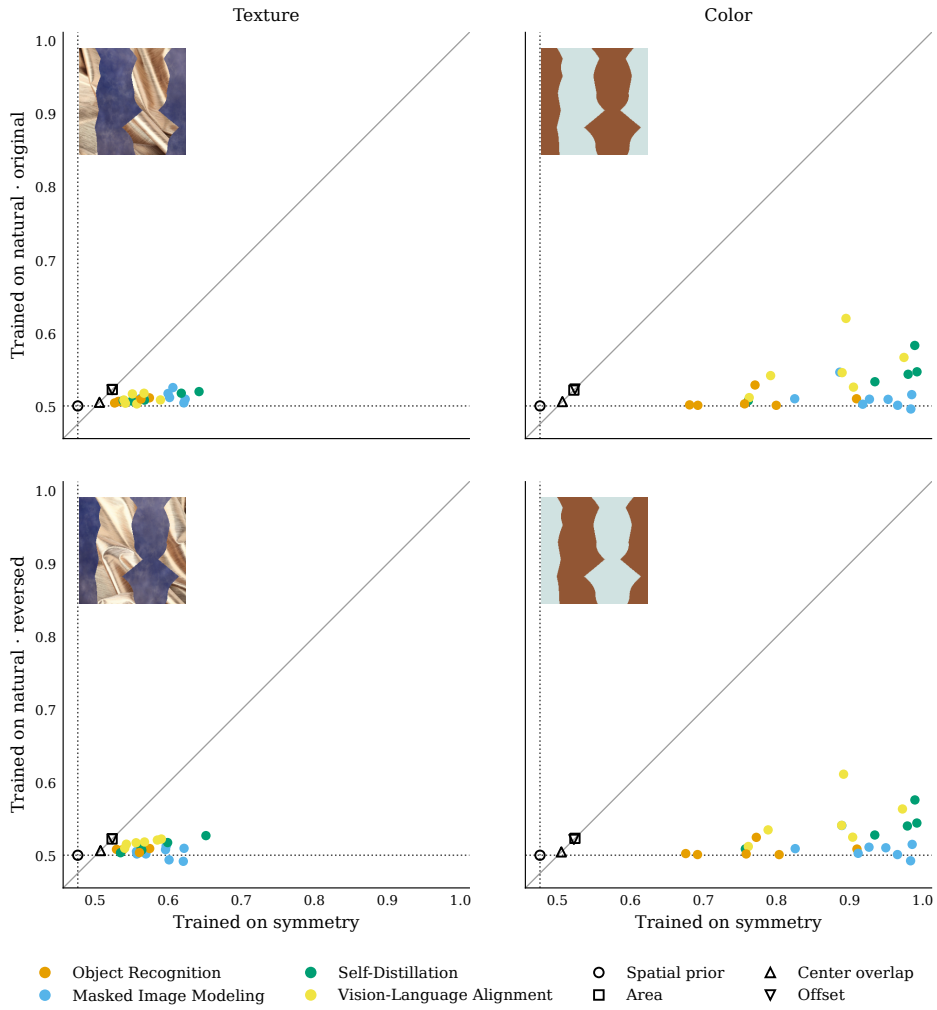

Figure 25: Performance comparison of probes directly trained on the symmetry stimuli (x-axis) with zero-shot probes trained on natural images (y-axis).

A.6 Parallelism



Figure 26: Example images from the parallelism condition, showing random stripes of constant width. Foreground stripes and background are filled with random textures or colors.

Table 6: Probe performances on the parallelism test sets, averaged over all four variants (texture and color variants of the original and reversed stimuli).

	Trained on natural					Trained on parallelism				
Model	ACC ↓	BCE	IG	IGE	mIoU	ACC	BCE	IG	IGE	mIoU
DINOv2-L	0.652	0.841	0.212	0.201	0.504	0.895	0.235	0.449	0.657	0.841
BEiT-3-L	0.635	1.089	0.041	0.036	0.478	0.918	0.213	0.469	0.688	0.875
PE-L	0.613	1.249	-0.189	-0.178	0.455	0.909	0.216	0.467	0.683	0.864
PE-B	0.611	1.279	-0.216	-0.203	0.465	0.863	0.315	0.368	0.539	0.803
EVA-02-L	0.599	1.163	-0.055	-0.049	0.440	0.893	0.273	0.409	0.600	0.837
DINOv3-L	0.592	1.137	-0.022	-0.019	0.418	0.916	0.201	0.481	0.705	0.872
CLIP-L	0.592	1.115	-0.006	-0.006	0.438	0.844	0.358	0.325	0.476	0.757
BEiT-L	0.583	1.190	-0.060	-0.053	0.408	0.908	0.217	0.465	0.682	0.858
DINOv2-B	0.574	1.107	-0.054	-0.052	0.403	0.862	0.305	0.378	0.554	0.796
CLIP-B	0.570	1.027	0.103	0.091	0.400	0.808	0.420	0.262	0.384	0.705
EVA-02-B	0.558	1.378	-0.270	-0.243	0.382	0.842	0.363	0.320	0.469	0.755
BEiT-B	0.546	1.461	-0.331	-0.293	0.360	0.872	0.298	0.384	0.563	0.801
DeiT-3-L	0.535	2.049	-0.919	-0.813	0.350	0.881	0.282	0.400	0.587	0.806
SigLIP-2-L	0.535	1.298	-0.182	-0.163	0.357	0.858	0.334	0.348	0.510	0.782
AugReg-L	0.532	1.607	-0.477	-0.422	0.341	0.837	0.358	0.324	0.475	0.731
BEiT-3-B	0.532	1.725	-0.596	-0.527	0.339	0.873	0.307	0.375	0.549	0.808
DINOv3-B	0.531	1.537	-0.421	-0.377	0.335	0.887	0.253	0.429	0.629	0.836
AugReg-B	0.530	1.546	-0.417	-0.369	0.336	0.843	0.346	0.336	0.492	0.743
DINO-B	0.527	1.466	-0.337	-0.298	0.340	0.870	0.312	0.370	0.542	0.798
MAE-B	0.526	1.904	-0.774	-0.685	0.341	0.839	0.362	0.320	0.469	0.768
MAE-L	0.525	1.970	-0.840	-0.743	0.335	0.866	0.306	0.376	0.551	0.802
DeiT-3-B	0.525	1.864	-0.734	-0.650	0.334	0.852	0.330	0.352	0.516	0.766
SigLIP-2-B	0.521	1.410	-0.280	-0.248	0.338	0.792	0.449	0.233	0.341	0.684
FlexiViT-L	0.514	1.820	-0.697	-0.621	0.309	0.854	0.363	0.320	0.468	0.762
FlexiViT-B	0.498	1.629	-0.506	-0.451	0.289	0.784	0.468	0.215	0.314	0.662
Spatial prior	0.496	1.093	0.000	0.000	0.314	0.561	0.683	0.000	0.000	0.370
Area	0.627	–	–	–	–	0.627	–	–	–	–
Center overlap	0.581	–	–	–	–	0.581	–	–	–	–
Offset	0.627	–	–	–	–	0.627	–	–	–	–

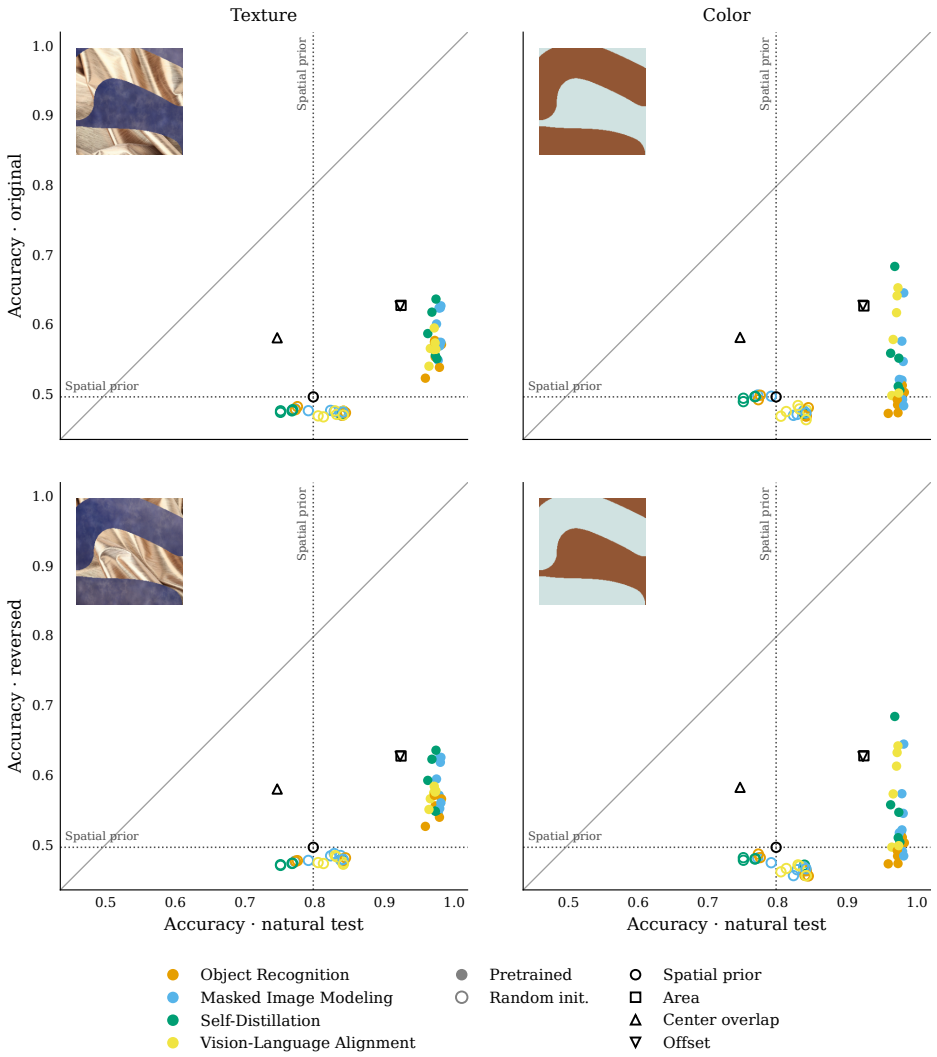

Figure 27: Comparison of the probe performances on the i.i.d. test set of natural images (x-axis) and applied zero-shot to the stimuli isolating parallelism (y-axis).

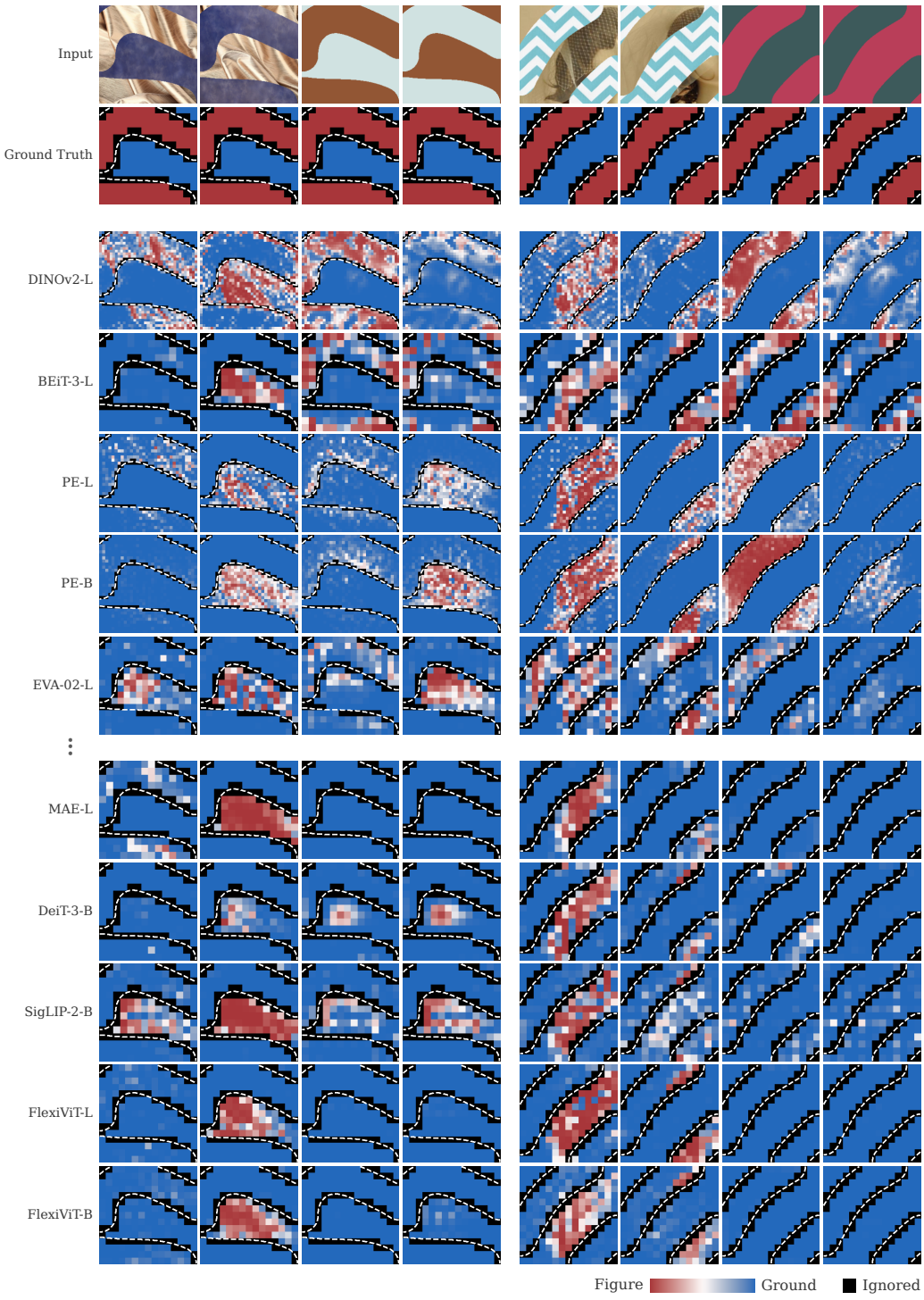

Figure 28: Probe predictions for the best and worst models in the parallelism condition.

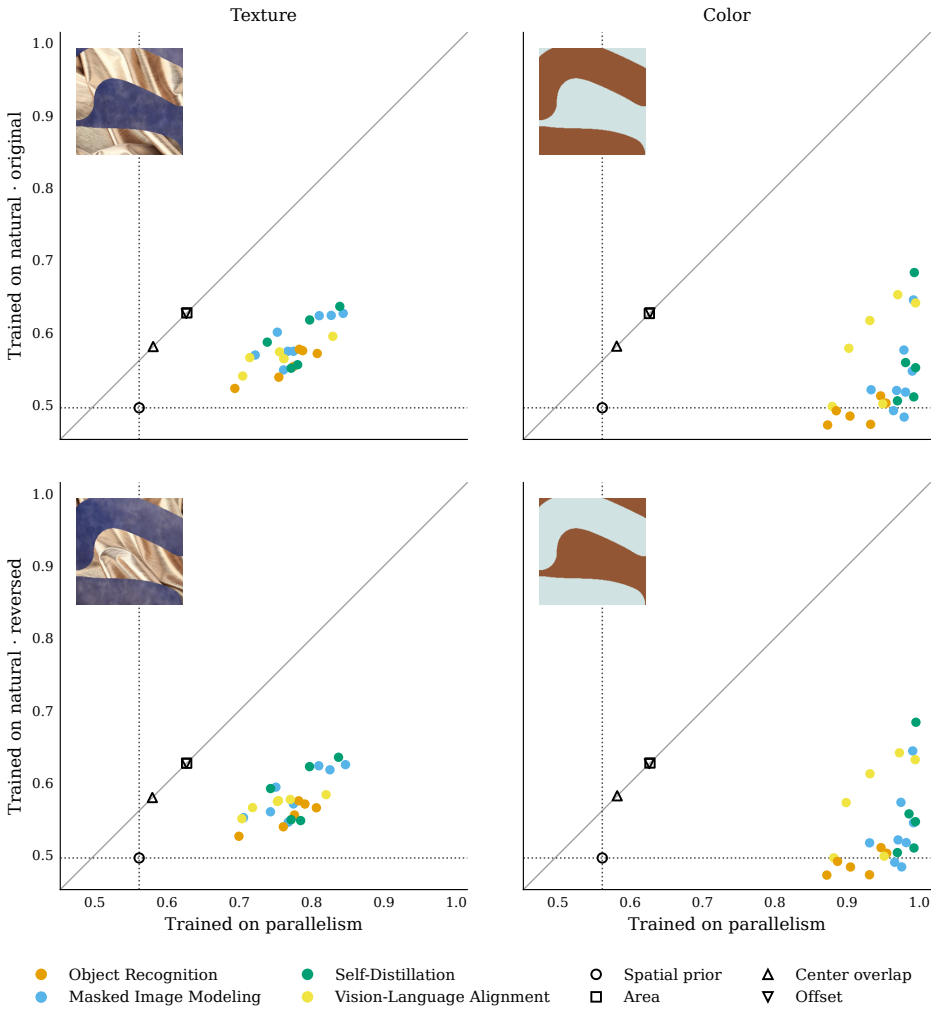

Figure 29: Performance comparison of probes directly trained on the parallelism stimuli (x-axis) with zero-shot probes trained on natural images (y-axis).

B Performance by Layer

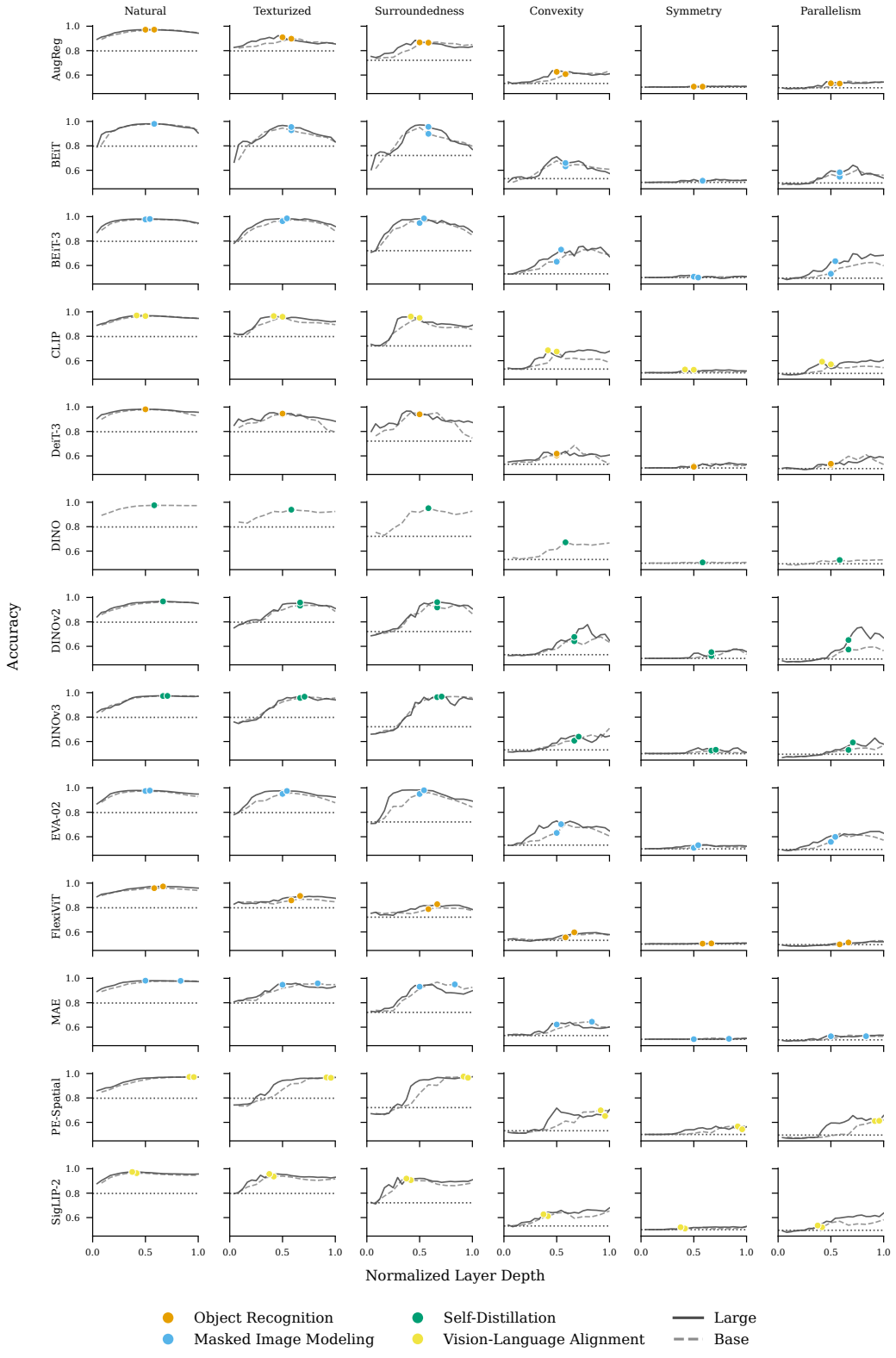

Figure 30: Individual probe performances for each model across layers.

C Detailed significance results for Figure 5

Table 7: Summary statistics per condition. W: Wilcoxon signed-rank statistic (base vs. large, matched by model family); r: effect size $|Z|/\sqrt{n}$, with Z the normal approximation of W and n the number of non-zero pairs; H: Kruskal-Wallis statistic (pretraining task); $\epsilon^2 = H/(N-1)$. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Condition	N	N pairs	W	p(W)	r	H	df	p(H)	ϵ^2
Natural	25	12	4.0	0.003**	0.793	11.744	3	0.008**	0.489
Texturized	25	12	6.0	0.007**	0.747	10.588	3	0.014*	0.441
Surroundedness	25	12	13.0	0.042*	0.589	9.116	3	0.028*	0.380
Convexity	25	12	15.0	0.064	0.543	10.691	3	0.014*	0.445
Symmetry	25	12	30.0	0.519	0.204	11.290	3	0.010*	0.470
Parallelism	25	12	1.0	<0.001***	0.861	6.380	3	0.095	0.266

Table 8: Dunn post-hoc p-values (Holm-corrected) for all pretraining tasks. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Pair	Natural	Texturized	Surroundedness	Convexity	Symmetry	Parallelism
MIM vs OR	0.231	0.030*	0.073	0.025*	1.000	0.220
OR vs SD	1.000	0.242	0.062	0.094	0.123	0.206
SD vs VLA	1.000	1.000	1.000	1.000	1.000	1.000
MIM vs SD	0.082	1.000	1.000	1.000	0.123	1.000
OR vs VLA	0.672	0.020*	0.073	0.027*	0.058	0.168
MIM vs VLA	0.008**	1.000	1.000	1.000	0.058	1.000

D Licenses

We use the DTD dataset [16], which does not provide a standard license but "is made available to the computer vision community for research purposes." (https://www.robots.ox.ac.uk/ vgg/data/dtd/, 2026-05-06).

Similarly, the MSRA-10K dataset [15] does not provide a standard license but requires researchers to cite their paper (https://mmcheng.net/msra10k/, 2026-05-05).

We further use shapes from the Infinite DSprites dataset [21] which is licensed under the MIT license (https://github.com/sbdzdz/idsprites/, 2026-05-05).

The licenses and model cards for all models evaluated in our work are listed in Table 9.

Table 9: Vision Transformer checkpoints used in this work and their corresponding Hugging Face model cards and licenses.

Model	**Checkpoint**	**License**
AugReg-B	timm/vit_base_patch16_224.augreg_in21k_ft_in1k	Apache-2.0
AugReg-L	timm/vit_large_patch16_224.augreg_in21k_ft_in1k	Apache-2.0
BEiT-3-B	timm/beit3_base_patch16_224.in22k_ft_in1k	MIT
BEiT-3-BO	timm/beit3_base_patch16_224.pt	MIT
BEiT-3-L	timm/beit3_large_patch16_224.in22k_ft_in1k	MIT
BEiT-3-LO	timm/beit3_large_patch16_224.pt	MIT
BEiT-B	timm/beit_base_patch16_224.in22k_ft_in22k	MIT
BEiT-L	timm/beit_large_patch16_224.in22k_ft_in22k	MIT
CLIP-B	timm/vit_base_patch16_clip_224.openai	MIT
CLIP-L	timm/vit_large_patch14_clip_224.openai	MIT
DINO-B	timm/vit_base_patch16_224.dino	Apache-2.0
DINOv2-B	timm/vit_base_patch14_reg4_dinov2.lvd142m	Apache-2.0
DINOv2-L	timm/vit_large_patch14_reg4_dinov2.lvd142m	Apache-2.0
DINOv3-B	timm/vit_base_patch16_dinov3.lvd1689m	DINOv3 License
DINOv3-L	timm/vit_large_patch16_dinov3.lvd1689m	DINOv3 License
DeiT-3-B	timm/deit3_base_patch16_224.fb_in22k_ft_in1k	Apache-2.0
DeiT-3-L	timm/deit3_large_patch16_224.fb_in22k_ft_in1k	Apache-2.0
EVA-02-B	timm/eva02_base_patch14_224.mim_in22k	MIT
EVA-02-L	timm/eva02_large_patch14_224.mim_in22k	MIT
FlexiViT-B	timm/flexivit_base.1200ep_in1k	Apache-2.0
FlexiViT-L	timm/flexivit_large.1200ep_in1k	Apache-2.0
MAE-B	timm/vit_base_patch16_224.mae	CC-BY-NC-4.0
MAE-L	timm/vit_large_patch16_224.mae	CC-BY-NC-4.0
PE-B	timm/vit_pe_spatial_base_patch16_512.fb	Apache-2.0
PE-L	timm/vit_pe_spatial_large_patch14_448.fb	Apache-2.0
SigLIP-2-B	timm/vit_base_patch16_siglip_224.v2_webli	Apache-2.0
SigLIP-2-L	timm/vit_large_patch16_siglip_256.v2_webli	Apache-2.0

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide experimental results for all claims in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated section on the limitations of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not contribute theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We carefully describe all details needed to reproduce our results in the method section, including details on stimulus generation, probe training and evaluation. Additionally, we publish all code for stimulus generation, probe training and evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All code and data for this is publicly available under the link provided in the abstract.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The methods section describes our experimental setup in great detail.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We evaluate and report statistical significance for our claims regarding the influence of model size and pre-training task. Evaluating the statistical significance of the model ranking for each condition is prohibitive given the compute requirements. We therefore do not make any strong claims about detailed model rankings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the GPU type used (NVIDIA L40S) and the total time required for fitting the probes (120 GPU hours).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: We study the internal processing of published models from the perspective of research of human vision. We don't see any potential negative societal impacts of this work.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We don't see any such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all papers for datasets and models as required by the respective authors. Furthermore, we list licenses and model cards in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not provide a new model and did not collect a new dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work does not include experiments with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work does not include experiments with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: The core method of our work does not involve LLM usage.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.