How Far Is Too Far? Object Recognition Declines Monotonically with Semantic Distance

Domenic Bersch
Department of Computer Science
Goethe University
Frankfurt, Germany
bersch@em.uni-frankfurt.de

Hai Van Tran
Department of Computer Science
Goethe University
Frankfurt, Germany
s7671493@stud.uni-frankfurt.de

Gemma Roig
Department of Computer Science
Goethe University
Frankfurt, Germany
roignoguera@em.uni-frankfurt.de

Abstract

Vision models are sensitive to object–background context, yet it remains unclear how recognition changes as the semantic distance between objects and backgrounds increases. We introduce ImageNet-OOC1k, a benchmark enabling graded control of object–background relationships along a continuous consensus visual-semantic dissimilarity axis. This formulation reveals a highly consistent monotonic decline in recognition accuracy across 35 models spanning convolutional neural networks, Vision Transformers, and vision–language models. Internal analyses of Vision Transformers reveal patterns consistent with aggregation failure, where object evidence is present at the patch level but not consolidated into the final prediction. Using this characterization as a diagnostic probe, a simple spatial re-ranking strategy recovers a subset of these errors and improves in-sample accuracy by up to 5.7 percentage points. These results establish context sensitivity as a continuous and predictable property of modern vision systems, with Vision Transformer analyses suggesting that aggregation limitations may contribute to some failure types even when object evidence remains available.

1 Introduction

Object recognition in natural images is often framed as an object-centric task [6, 27]. In practice, however, scenes contain strong statistical regularities between objects and their surrounding context, and modern vision models exploit these regularities extensively [3, 10, 43]. Existing benchmarks study this dependence using binary or coarse manipulations, such as contrasting congruent and incongruent object–background pairings [19, 28, 40]. While these approaches show that context matters, they do not provide a continuous notion of semantic relatedness. This raises the central question of what happens to recognition as object–background mismatch increases.

We address this limitation by introducing **ImageNet-OOC1k**[1], a benchmark for graded, parametrically controlled manipulation of object–background relationships. We define a visual-semantic axis over ImageNet classes from a consensus of classifier representations across multiple pretrained

[1] Code and dataset information: https://github.com/ToastyDom/ImageNet-OOC1k

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

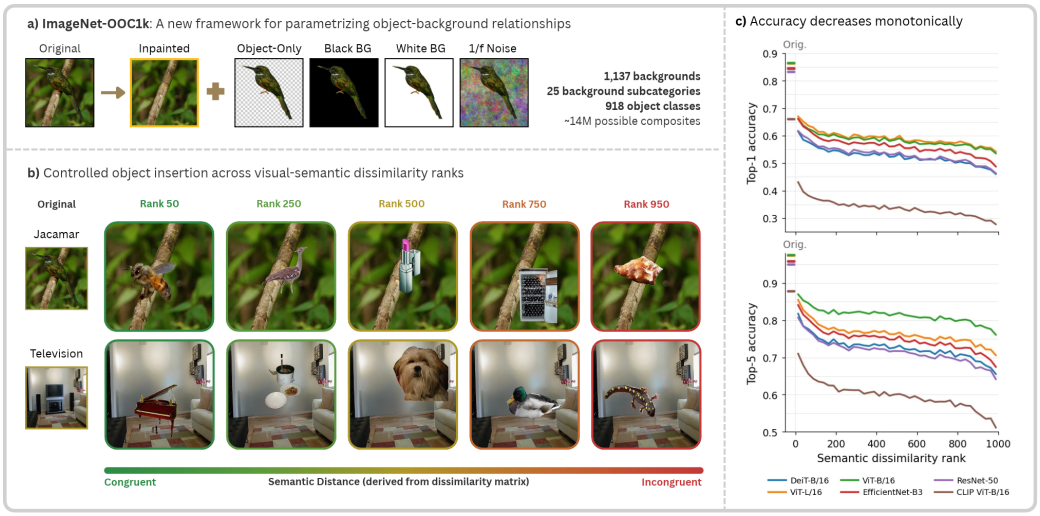


Figure 1: **ImageNet-OOC1k enables controlled recomposition along a visual-semantic axis. (a)** Dataset overview. ImageNet-S masks [9] are used to construct original images, inpainted backgrounds, and object-only variants, including crops and composites on black, white, and $1/f$ noise backgrounds. **(b)** Objects are inserted at increasing semantic distance ranks relative to the original class. The example backgrounds (*jacamar*, *television*) use the same ranks (50, 250, 500, 750, 950), but different inserted objects because distance is defined relationally. **(c)** Top-1 accuracy decreases monotonically with semantic distance rank (bin width 25) for representative models (ViT-B/16, ViT-L/16, DeiT-B/16, ResNet-50, EfficientNet-B3, CLIP ViT-B/16). Colored markers (*Orig.*) show original-image accuracy. Top-5 accuracy follows the same trend, indicating that semantic mismatch affects the broader class ranking.

models, whose classifier heads can be interpreted as class prototypes in feature space that encode semantic structure [36]. This allows recognition to be analyzed directly as a function of semantic distance.

We then use this classifier-derived axis to measure how recognition changes across 35 models. Across convolutional neural networks, Vision Transformers, and vision–language models, accuracy follows a monotonic function of semantic object–background distance (Figure 1c). The closely aligned trajectories indicate that context sensitivity follows a shared and structured pattern rather than arising from isolated failures.

To investigate the mechanisms underlying this behavior, we analyze internal representations in Vision Transformers, which allow patch-level inspection of class evidence. Using ViT-B/16 as a case study, we show that a subset of failures arise despite spatially distributed patch-level evidence for the correct object, consistent with incomplete consolidation of competing object and background signals into the final prediction. We examine their recoverability with a spatial re-ranking probe that leverages intermediate spatial structure, improving in-sample accuracy by +5.7 percentage points while maintaining a 9.4:1 found-but-lost (FBL) recovery-to-breakage ratio. This indicates that relevant information remains available beyond what is reflected in the final prediction.

Our contributions are threefold: (1) we introduce ImageNet-OOC1k, a benchmark for graded control over object–background relationships along a continuous visual-semantic axis; (2) we identify a highly consistent monotonic relationship between semantic distance and recognition accuracy across diverse vision models; and (3) we identify a failure mode under context shifts that suggests limitations in how Vision Transformers aggregate available object evidence, and demonstrate that a substantial fraction of these errors can be recovered through spatial re-ranking.

2 Related Work

We discuss prior work most related to ours, focusing on context dependence in visual recognition, datasets for studying object–background relationships, and methods for analyzing model internals.

Object–background context in visual recognition. A large body of work shows that recognition models rely on contextual information beyond the object itself. Backgrounds alone can carry significant class signal [21, 43], and models often fail when objects appear in atypical environments [3, 26]. Controlled evaluations further show substantial performance drops when contextual cues are altered or removed [2, 40]. These effects are often framed as shortcut learning, where models exploit spurious correlations in the data, including correlations between object classes and backgrounds [10, 28].

Datasets for context sensitivity. Existing datasets study context through binary or coarse manipulations, such as contrasting congruent and incongruent pairings or introducing controlled correlations between objects and environments [28, 40]. Other approaches explore unusual object–scene relationships or induce distribution shifts by placing objects in atypical environments [3, 13, 14, 26]. Recent generative and inpainting-based pipelines increase contextual diversity [19, 42], while synthetic datasets allow controlled semantic or physical violations [4]. However, these approaches lack precise continuous control over semantic relatedness and therefore cannot characterize how recognition varies with the degree of mismatch.

Mechanistic interpretability in Vision Transformers. Within the growing literature on vision model interpretability, our work builds most directly on logit lens methods that decode intermediate predictions and reveal how class evidence evolves across layers [8, 22, 36]. Related work highlights limitations of the CLS token, showing that it can fail to capture object information in complex scenes [15]. However, these tools have not been applied to controlled context manipulations. We connect them to context sensitivity by testing whether failures under controlled object–background mismatch reflect limitations in aggregating distributed object evidence.

3 ImageNet-OOC1k: Parametrizing Object–Background Relationships

Understanding how recognition changes under object–background mismatch requires precise control over the semantic relationship between objects and scenes. Existing benchmarks mainly use binary or coarse manipulations, which do not allow performance to be analyzed as a continuous function of mismatch. To address this, we introduce ImageNet-OOC1k, a benchmark that treats object–background context as a quantitative variable and enables recognition to be analyzed across the spectrum from compatible to strongly mismatched pairings (Figure 1a,b). Full dataset details are provided in Appendix A.

3.1 Dataset Construction

ImageNet-OOC1k builds on ImageNet-S [9], which provides high-quality segmentation masks for images from the ImageNet validation set [6, 27]. Using these annotations, each image is decomposed into a foreground object and a background obtained via human-supervised inpainting. We use the `IOPaint` toolkit [29] with the MAT model [16], resulting in 1,137 background-only scenes.

The same masks are used to construct a reusable object library of foreground crops aligned with ImageNet class labels, spanning 918 classes, of which 915 contain at least four valid object instances. The 1,137 backgrounds are grouped into nature scenes (558 images, 12 subcategories), human-related scenes (533 images, 13 subcategories), and miscellaneous scenes (46 images) containing close-ups, ambiguous scenes, or backgrounds that do not fit either category. Example backgrounds are shown in Appendix Figure 8, and detailed dataset statistics are provided in Appendix Table 3.

ImageNet-OOC1k contains inpainted backgrounds, object masks, composite specifications, and metadata describing object classes, scene categories, and generation parameters. Object-only control images are described in Appendix A.3. Because the images derive from ImageNet/ImageNet-S, image-derived components are released in forms compatible with source-data terms, as detailed in Appendix A.4.

3.2 Controlled Composite Generation

The key property of ImageNet-OOC1k is the explicit decoupling of objects and backgrounds, enabling combinatorial recomposition of scenes. For each background, object classes are combined across

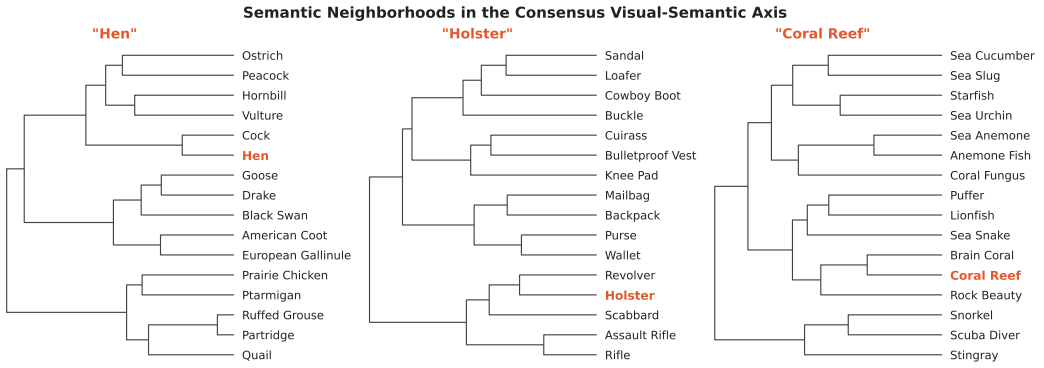


Figure 2: **Semantic neighborhoods defined by the consensus visual-semantic axis.** Each panel shows the 15 nearest neighbors of a focal class (highlighted), clustered using Ward linkage. Classes form coherent semantic groups (e.g., *hen* with birds, *holster* with weapons, *coral reef* with marine organisms), illustrating that classifier representations encode rich semantic structure. Similar high-level groupings are consistently observed across individual classifiers (Appendix Figures 10–12), indicating that this structure is broadly shared across models.

the full range of relative positions along the visual-semantic axis (Section 4), spanning a range from highly compatible to strongly mismatched pairings.

This defines approximately 14 million possible object–background combinations using individual object instances, enabling dense sampling across visual-semantic space rather than relying on a small set of predefined conditions. Object sampling, spatial placement, and scale follow fixed, reproducible procedures, and all generation parameters are recorded. Composite-generation details, including sampling, scaling, and placement, are provided in Appendix A.1. The automated generation pipeline is available at `https://github.com/ToastyDom/ImageNet-OOC1k`, together with metadata, composite specifications, and documentation of the source-data constraints. This treats object–background context as a graded experimental variable, allowing recognition to be measured as a function of visual-semantic distance rather than under fixed congruent or incongruent conditions (Figure 1b).

4 A Consensus Visual-Semantic Axis

To parameterize object–background relationships in ImageNet-OOC1k, we define semantic distance directly in the representational space of pretrained classifiers. The final linear layer of modern models can be interpreted as a set of class prototypes, where each row corresponds to a class embedding in feature space [36]. The relative arrangement of these prototypes forms a structured embedding in which visually and semantically related classes lie nearby. This organization emerges from visual discrimination and is consistently observed across architectures, as shown in Figure 2 and Appendix B (Figures 9–12).

To obtain a stable and model-agnostic estimate, we aggregate across diverse architectures. We compute semantic distances from ten pretrained ImageNet classifiers, including Vision Transformers (ViT-B/16 at 224 and 384 resolution, ViT-S/16, ViT-L/16, ViT-B/32) and convolutional networks (ResNet-50, ResNet-101, EfficientNet-B3, EfficientNet-B5, DenseNet-121). For each model m, let $U^{(m)} \in \mathbb{R}^{C\times d}$ denote the classifier weight matrix over C classes and d-dimensional class embeddings, with row $u_i^{(m)}$ representing class i. Rows are L2-normalized and pairwise cosine dissimilarity is computed,

$$D_{ij}^{(m)} = 1 - \frac{u_i^{(m)}}{\|u_i^{(m)}\|_2} \cdot \frac{u_j^{(m)}}{\|u_j^{(m)}\|_2}. \tag{1}$$

The resulting matrices are symmetrized via $(D + D^\top)/2$ to correct minor numerical asymmetries. To account for scale differences across the M models, we z-score the off-diagonal entries of each symmetrized matrix to obtain $\tilde{D}^{(m)}$. Averaging the normalized matrices gives the consensus dissimilarity

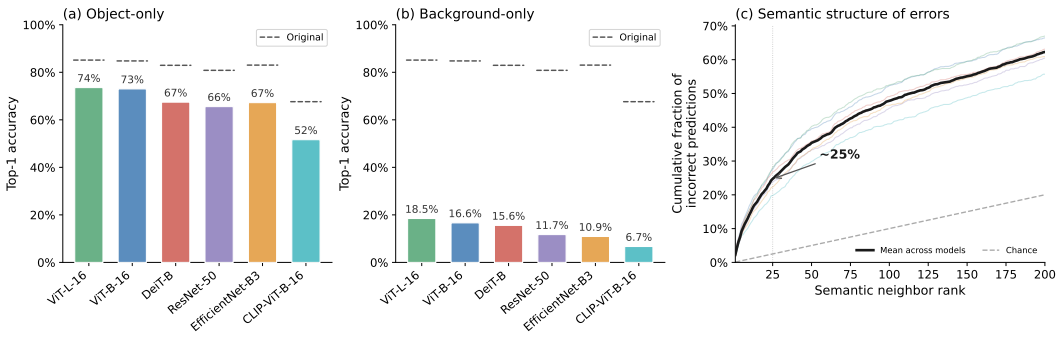


Figure 3: **Objects and backgrounds provide separable recognition signals. (a)** Top-1 accuracy for object-only images across representative models. Bars show performance averaged across three neutral background conditions (black, white, and $1/f$ noise; see Appendix A.3). Dashed lines indicate accuracy on original images. **(b)** Top-1 accuracy for background-only images across the same models. Dashed lines again indicate original-image accuracy. **(c)** Cumulative semantic rank of incorrect background-only predictions. The x-axis shows semantic neighbor rank relative to the ground-truth class, where 1 denotes the nearest semantic neighbor. The y-axis shows the cumulative fraction of incorrect predictions whose predicted class falls within that rank. Thin lines correspond to the models shown here, and the thick line shows their mean. The dashed line indicates chance-level accumulation. Results for all 35 models are shown in Appendix C.1.

matrix,

$$D_{ij}^{\text{cons}} = \frac{1}{M} \sum_{m=1}^{M} \tilde{D}_{ij}^{(m)}. \tag{2}$$

The consensus matrix is then symmetrized, and its off-diagonal entries are min–max rescaled to $[0, 1]$. Diagonal entries are fixed at zero throughout, preserving $D_{ii}^{\text{cons}} = 0$. The resulting consensus axis summarizes this shared class structure while reducing model-specific noise. Low values correspond to related classes within this discriminative space, while high values reflect increasing mismatch. This axis defines the continuous variable used to parameterize object–background relationships in ImageNet-OOC1k. We assess potential circularity through model-overlap and own-axis controls and comparisons with axes derived from independently pretrained representations (Appendix C.4). The decline generalizes across the tested discriminative axes, while retrieval, linguistic, and taxonomic distances produce different curves.

5 Behavioral Analysis of Context Dependence

Having established a continuous visual-semantic axis over object–background relationships, we now analyze how context influences recognition. We ask three questions. First, do backgrounds alone induce structured, class-specific predictions? Second, how does recognition change as a function of semantic mismatch? Third, do these effects depend on background type?

5.1 Backgrounds Induce Structured Semantic Predictions

We begin with a control analysis that separates object-derived from background-derived evidence. Prior work has shown that backgrounds can carry class signal [21, 43]. Here, we quantify this effect within the same semantic space used throughout the paper, using object-only images composited onto neutral backgrounds (black, white, and $1/f$ noise), with an example in Figure 1a and details in Appendix A.3.

Removing natural context reduces object-only performance by approximately 12–16 percentage points across models (Figure 3a). This indicates that models rely primarily on object information, but still benefit from contextual support. Performance varies little across the three neutral background conditions, is consistent across all 35 models (Table 4; Appendix C.1, Figure 15), and shows similar trends under top-5 evaluation. In contrast, background-only images yield substantially lower but still surprisingly high top-1 accuracy (Figure 3b), reaching 13.3% on average across the six representative

models despite the complete absence of the object. Under top-5 evaluation, background-only accuracy reaches up to $\sim$40% across the full 35-model set, with the highest values among BEiT and ConvNeXt models and lower values for CLIP and the evaluated CNNs (Appendix C.1).

Crucially, these predictions are not arbitrary. Among background-only images with an incorrect top-1 prediction, approximately 25% of predicted classes fall within the 25 nearest visual-semantic neighbors of the ground-truth class (Figure 3c). Across the full 35-model set, the corresponding analysis yields an 8.9-fold enrichment over a 2.9% permutation baseline ($p < 0.001$, Appendix C.1). Thus, background signals reliably place the model in the correct semantic neighborhood, even when the exact class is not recovered.

Together, these results show that backgrounds provide structured, class-relevant information rather than noise. We next ask how recognition changes when object- and background-derived signals are systematically pulled apart along the visual-semantic axis.

5.2 Recognition Degrades with Semantic Distance

As object and background semantics are pulled apart, accuracy decreases monotonically with semantic object–background distance across representative models (Figure 1c). The gap to original-image accuracy should not be interpreted as part of this decline across composite bins, since the lowest-rank bin already averages composites from the 25 nearest semantic neighbors and may include recomposition costs from scaling, insertion, and residual artifacts. This pattern generalizes across all 35 models, despite differences in architecture, training objective, and baseline performance (Appendix C.2). Across controlled composites (Appendix A.1), top-1 accuracy drops by approximately 9–19 percentage points from the most similar to the most dissimilar pairs. Image-level logistic regressions confirm that this trend is not an artifact of binned visualization, with strictly negative slopes and 95% confidence intervals below zero for every model (Appendix C.2). Negative slopes persist across all 35 models after controlling for placement, scale, and background and object-instance effects, and after additionally controlling for boundary discontinuities on a separate 30,000-composite sample (Appendix A.2). The trend also extends to top-5 accuracy, showing that semantic mismatch reshapes the broader class ranking rather than only the top prediction. CLIP [25] follows the same trend despite multimodal training, indicating that language grounding does not eliminate reliance on visual context.

While the overall relationship is monotonic, backgrounds separate into two regimes. Most induce decreases in accuracy, while a smaller subset shows facilitation, where accuracy increases with semantic distance. These regimes are structured, with nature scenes primarily driving interference and human-related environments more often producing facilitation (Appendix C.3). This may reflect the greater visual consistency of nature scenes, which impose stronger contextual expectations than more heterogeneous human-related environments. A mechanistic analysis of these background-specific regimes is beyond the scope of this paper, but the pattern shows that context type modulates object–background interactions in distinct ways.

Finally, this monotonic decline depends on the visual-semantic space used to define distance. Classifier heads and linear probes on independently pretrained DINOv1 [5], DINOv2 [23], and MAE features [12] reproduce the decline, whereas retrieval, CLIP-text, and WordNet [20] distances produce qualitatively different curves (Appendix C.4). Thus, the result is not a generic consequence of ranking classes by any semantic measure, but reflects discriminative visual class geometry.

6 Representational Competition and Aggregation Failure

Section 5.2 showed that recognition accuracy declines monotonically as semantic object–background distance increases, suggesting that failures reflect structured interactions between object and background signals. We therefore ask whether these failures arise because object information is missing, or because competing object- and background-derived signals are not successfully combined.

We study this question in ViT-B/16, where the CLS token serves as a global aggregation token that collects information from spatial patch tokens for the final classification decision. This architecture enables patch-level inspection via the logit lens [22, 36], allowing us to track how object- and background-derived class evidence evolves across layers before it is consolidated into the final

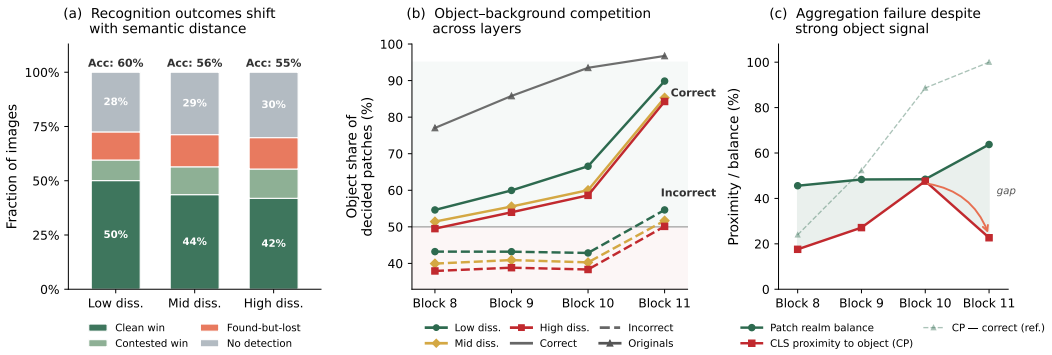


Figure 4: **Recognition outcomes reveal increasing object–background competition and aggregation failure.** All panels use the 75,000-composite stratified subset analyzed for ViT-B/16. **(a)** Outcome composition across all composites, grouped into low (1–350), mid (351–700), and high (701–1,000) semantic dissimilarity rank bands. Stacked bars show the fraction of images assigned to each outcome category at post-norm. **(b)** Object share of decided patches (realm balance) across blocks 8–11, computed over all composites and split by correctness and dissimilarity group. Realm balance captures the competition between object- and background-aligned patch evidence. Correct predictions transition to object dominance at block 11, whereas incorrect predictions remain balanced or background-leaning. **(c)** Aggregation dynamics restricted to found-but-lost cases. The green line shows patch-level realm balance, and the red line shows CLS classification proximity (CP) to the object class. Patch evidence increasingly favors the object, while CLS proximity to the object drops at the final layer.

prediction. The analysis is performed on a stratified subset of 75,000 composites sampled uniformly across 20 semantic distance bins (Appendix A.1).

6.1 Semantic Distance Increases Object–Background Competition

We decompose predictions into four outcome types based on correctness and patch-level object evidence at the final layer. Clean wins are correct predictions with strong object dominance across patches. Contested wins are correct predictions with weaker or mixed patch evidence. Found-but-lost cases are incorrect predictions in which many patches nevertheless predict the object class. No-detection cases are incorrect predictions with little patch-level object evidence. Formal definitions of the outcome taxonomy are given in Appendix D.2. The metrics used to define them, including semantic pull, realm balance, and classification proximity, are described in Appendix D.1.

We first ask how the frequency of these outcome types changes with semantic distance. Figure 4a shows that clean wins decrease as object–background distance increases, while no-detection cases account for most of the performance drop. Contested wins and found-but-lost (FBL) cases change more modestly in frequency, but frequency alone does not capture how the underlying evidence changes. We therefore quantify object–background competition using *realm balance*, the fraction of decided patches favoring the object rather than the background. Figure 4b shows that increasing semantic distance shifts predictions toward background dominance, especially for incorrect cases, which remain background-leaning until the final layer. At the final layer, correct predictions transition sharply to object dominance, whereas incorrect predictions remain unresolved. Successful recognition is therefore associated with object evidence overcoming sustained background competition at the aggregation stage.

6.2 Found-but-Lost Cases Reveal Aggregation Failure

FBL cases represent one specific error type in which the object is clearly detected at the patch level, yet the model still predicts the wrong class. They provide a targeted diagnostic for aggregation failure, while the causal mechanism remains unresolved. The spatial structure of these cases shows that the key difference is not whether the object is found, but how object evidence evolves across the patch grid (Figure 5). In clean wins, object evidence spreads across most of the grid by the final layer and

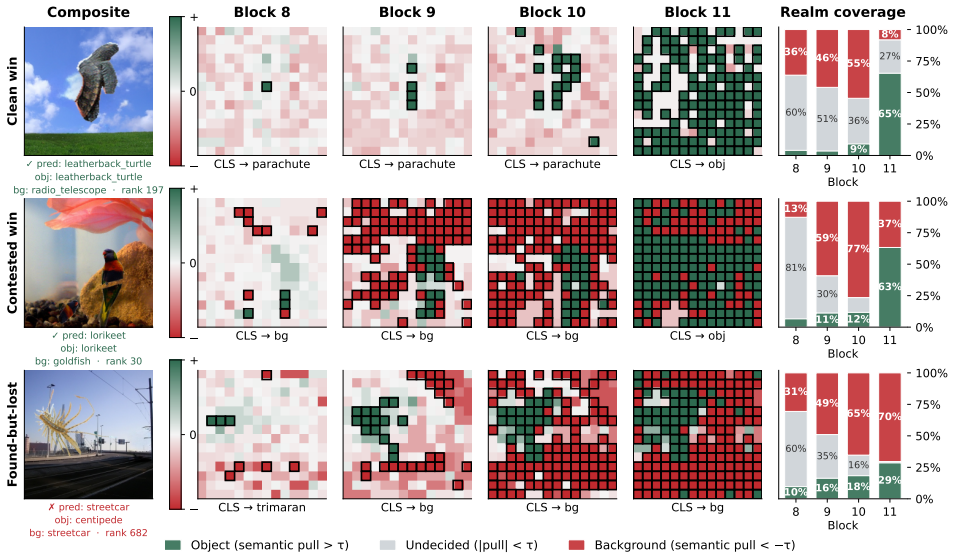


Figure 5: **Patch-level semantic pull heatmaps across outcomes.** Each row shows an example composite and patch-level predictions across the final transformer blocks (8–11). Green indicates alignment with the object class and red with the background class, with intensity reflecting semantic proximity. Black borders denote exact class matches. The rightmost column shows the fraction of patches assigned to each region. Rows correspond to clean win, contested win, and found-but-lost cases, illustrating full, partial, and stalled expansion of object-aligned patches.

achieves global dominance. In contested wins, this expansion is partial. In FBL cases, expansion stalls at an intermediate stage, leaving object evidence spatially localized, while no-detection cases show little to no expansion. Appendix Figure 21b quantifies this pattern by measuring the expansion of exact object-class patches across layers.

Figure 4c shows how this incomplete expansion relates to the CLS representation. Up to block 10, patch-level evidence and the CLS representation evolve in parallel toward the object class. At the final layer, however, this alignment breaks. Patch-level evidence continues to strengthen toward the object, while the CLS representation reverses direction, with classification proximity (Appendix D.1) dropping from approximately 48% to 23%. Thus, FBL cases reveal a failure to consolidate available object evidence into the final prediction.

This failure is quantitatively a near-miss. In 98.7% of FBL cases, the correct class appears within the CLS top-5 at post-norm with a median rank of 2, compared to 37% and a median rank of 9 for no-detection cases (Appendix Figure 21a). This shows that the correct class is often available in the model's representation, but not selected.

The incomplete expansion also leaves a measurable spatial signature. In FBL cases, patches supporting the runner-up class, often the true object class, form a compact cluster, while patches supporting the predicted class are more dispersed, especially at block 10 where aggregation remains unresolved (Appendix Figure 21c). While not all errors are of this type, these results identify a clear and interpretable failure mode characterized by incomplete expansion and consolidation of available object evidence into the final decision.

6.3 Recoverability of Aggregation Failure

The compact spatial signature identified above suggests that some aggregation failures may be recoverable from intermediate representations. We exploit this structure as a diagnostic probe, rather than as a deployable inference-time correction method. FBL cases exhibit three properties that make this probe possible. The correct class often remains near the top of the CLS ranking, object evidence remains spatially compact at intermediate layers, and aggregation is uncertain near the final decision.

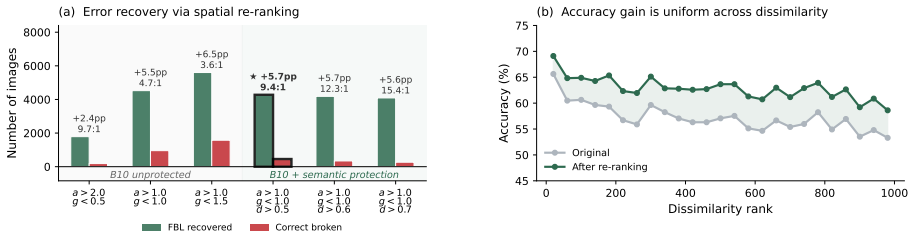


Figure 6: **Spatial re-ranking probes the recoverability of aggregation failures in ViT-B/16. (a)** Recovery and breakage counts at block 10 across re-ranking rules, measured as FBL cases recovered and originally correct predictions broken. The left panel uses spatial compactness advantage and logit gap (a, g); the right additionally constrains semantic distance between the top-1 and runner-up classes (d) to reduce near-neighbor swaps. The star marks the selected operating point used in the main analysis. **(b)** Top-1 accuracy before and after applying the selected re-ranking rule across semantic distance ranks (bin width 40), evaluated on the 75,000-image stratified ViT-B/16 diagnostic subset, which explains the slight baseline difference from the full composite benchmark in Figure 1c. Re-ranking improves accuracy across the full range, with a net gain of approximately +5.7 percentage points.

These observations motivate a constrained re-ranking strategy that compares the predicted class with the runner-up class using logit uncertainty, spatial compactness, and semantic distance.

The re-ranking rule combines CLS uncertainty, measured by the logit gap g, spatial compactness advantage between the top-1 and runner-up predictions a, and semantic distance between the competing classes d (Appendix E). Across the threshold sweep in Figure 6a, adding the semantic-distance constraint improves the recovery-to-breakage tradeoff, and the selected operating point recovers 4,277 FBL cases while breaking 456 correct predictions. This yields a 9.4:1 FBL recovery-to-breakage ratio. Including incidental corrections of non-FBL errors, the same rule produces an in-sample net accuracy improvement of +5.7 percentage points on the ViT-B/16 diagnostic subset, with gains observed across the full semantic distance range (Figure 6b).

Each component reflects the aggregation-failure signature identified above. The logit gap identifies uncertain aggregation, spatial compactness captures localized object evidence, and semantic distance prevents near-neighbor swaps. To test whether this signature is specific to ViT-B/16 or to CLS-token pooling, we repeated the outcome taxonomy, patch-expansion analysis, and fixed-rule recovery probe across ViT-S/16, ViT-L/16, DeiT-B/16, and ViT-RelPos-B/16, a no-CLS ViT with average pooling. Across models, FBL cases remained near-miss errors, with the correct class appearing in the top-5 in 94.7–98.7% of cases. The fixed rule produced positive net accuracy changes across architectures, including in the average-pooling model, with recovery magnitude varying with the strength of the spatial compactness signal (Appendix F.1–F.3). Together, these results are consistent with the interpretation that a substantial subset of failures reflects how information is combined, rather than missing object representations.

7 Discussion and Conclusion

We introduced ImageNet-OOC1k, a framework for graded, model-grounded control over object–background relationships. This setting reveals that recognition accuracy decreases monotonically with visual-semantic object–background distance across diverse vision models, showing that context sensitivity is a predictable property rather than a discrete failure under incongruence.

The results also show that backgrounds provide structured, class-relevant signals that actively shape predictions. Recognition can therefore be understood as a competition between object-derived and background-derived evidence, whose balance changes systematically with visual-semantic distance. Individual backgrounds nevertheless differ, with some showing facilitation rather than interference (Appendix C.3). This graded structure is central to the benchmark. By separating object–background relationships along this axis, ImageNet-OOC1k exposes failure signatures that binary congruent/incongruent manipulations would likely conflate. In Vision Transformers, we identify one such signature, where a subset of failures points to limitations in how competing signals are

aggregated. Even when the correct class is strongly represented across patches, the final prediction can fail to consolidate this evidence. The recovery experiment supports this interpretation, showing that a substantial fraction of FBL errors can be corrected using intermediate spatial structure. While this probe is not a complete solution, it suggests that aggregation is a potential bottleneck and points to future architectural mechanisms for better resolving object–background competition.

Our work has several limitations. The dataset relies on inpainting and compositing, which may introduce artifacts, and object cutouts inherited from ImageNet-S are not always perfectly coherent. Placement, scale, and appearance changes are sampled independently of the visual-semantic axis. The decline persists after controlling for placement, scale, and background and object-instance effects, and after additionally controlling for boundary discontinuities on a separate sample (Appendix A.2). Pair-specific visual compatibility between object and scene, however, cannot be fully separated from semantic mismatch. The single-target setting also leaves multi-object interactions for future investigation. Why some backgrounds hinder recognition while others help remains unclear. Differences in object plausibility and scene diversity may play a role. While the behavioral result spans all 35 models, the mechanistic evidence is restricted to the tested Vision Transformers. Generalization to CNNs remains an open question requiring an analogous spatial diagnostic. The axes remain tied to ImageNet categories, including through the labels used to train the independent probes, and their alignment with human similarity judgments has not been tested. Finally, the recovery analysis requires access to intermediate representations and is intended as a diagnostic probe rather than a deployable inference-time method. It does not establish the causal mechanism of FBL errors.

Overall, ImageNet-OOC1k shows that controlled context manipulations can link recognition errors to patterns of object and background evidence within models. By turning object–background context into a graded experimental variable, the dataset can support future work on anticipating failures, comparing how architectures resolve competing evidence, and designing models that more reliably integrate objects within the scenes in which they appear.

Acknowledgments

This work was partly funded by the BMFTR neuroAI grant GROWUPAI, DFG project 5368 (Abstract REpresentations in Neural Architectures, ARENA), and DFG project, RO 6458/5-1 (Learning from the Environment Through the Eyes of Children, LEECHI). We gratefully acknowledge support from the Hessian Center for Artificial Intelligence and Goethe University (NHR Center NHR@SW) for providing the computing and data-processing resources needed for this work. The authors declare no competing interests.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf`.

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[4] Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 255–264, October 2021.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pages 9630–9640. IEEE, 2021.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

[8] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024.

[9] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7457–7476, 2022.

[10] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 15979–15988. IEEE, 2022.

[13] Klara Janouskova, Cristian Gavrus, and Jiri Matas. Robust context-aware object recognition, 2025. URL https://arxiv.org/abs/2510.00618.

[14] P Kattakinda and S Feizi. Focus: Familiar objects in common and uncommon settings. In *International Conference on Machine Learning (ICML)*, 2022.

[15] Tarun Khajuria, Braian Olmiro Dias, Marharyta Domnich, and Jaan Aru. Interpreting the structure of multi-object representations in vision encoders. In *World Conference on Explainable Artificial Intelligence*, pages 359–382. Springer, 2025.

[16] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[18] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[19] H.S. Malik, M. Huzaifa, M. Naseer, S. Khan, and F.S. Khan. Objectcompose: Evaluating resilience of vision-based models on object-to-background compositional changes. In M. Cho, I. Laptev, D. Tran, A. Yao, and H. Zha, editors, *Computer Vision – ACCV 2024*, volume 15476 of *Lecture Notes in Computer Science*, Singapore, 2025. Springer. doi: 10.1007/978-981-96-0917-8_23.

[20] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

[21] Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19087–19097, 2022.

[22] nostalgebraist. Interpreting gpt: The logit lens. AI Alignment Forum, August 2020. URL `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`. Posted August 31, 2020. Last accessed April 9, 2026.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL `https://arxiv.org/abs/2304.07193`.

[24] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/radford21a.html`.

[26] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[28] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

[29] Quing Sanster. Iopaint: Open-source inpainting and outpainting tool. `https://github.com/Sanster/IOPaint`, 2024. GitHub repository.

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[31] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[32] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[33] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021.

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[35] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13684 of *Lecture Notes in Computer Science*, pages 516–533, Cham, 2022. Springer. doi: 10.1007/978-3-031-20053-3_30.

[36] Martina G. Vilas, Timothy Schaumlöffel, and Gemma Roig. Analyzing vision transformers for image classification in class embedding space. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 40030–40041. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7dd309df03d37643b96f5048b44da798-Paper-Conference.pdf.

[37] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 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*, pages 19175–19186, 2023.

[38] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[39] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142, June 2023.

[40] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021.

[41] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546, 2019. URL http://arxiv.org/abs/1905.00546.

[42] Tianze Yang, Tyson Jordan, Ruitong Sun, Ninghao Liu, and Jin Sun. Common inpainted objects in-n-out of context, 2025. URL https://arxiv.org/abs/2506.00721.

[43] Zhuotun Zhu, Lingxi Xie, and Alan Yuille. Object recognition with and without objects. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3609–3615, 2017. doi: 10.24963/ijcai.2017/505. URL https://doi.org/10.24963/ijcai.2017/505.

A ImageNet-OOC1k Dataset Details

ImageNet-OOC1k is derived from ImageNet-S [9], which provides segmentation masks for ImageNet validation images [6, 27]. We use these masks to separate foreground objects from their original backgrounds. Inpainting was performed under human supervision using `IOPaint` [29] with the MAT model [16], producing background-only scenes while retaining the surrounding scene structure. For each image, we interactively selected the region to remove, repeated the inpainting when necessary, and retained only backgrounds that passed visual inspection. We adopted this procedure because automated mask-based inpainting often produced duplicated textures, implausible structures, or residual object content. We also provide a reusable library of segmented foreground objects. The resulting dataset contains 1,137 background-only scenes and an object library spanning 918 ImageNet classes.

This appendix provides the dataset statistics, composite-generation details, controls for compositing artifacts, object-only control construction, release information, and compute resources supporting the main experiments.

A.1 Composite Generation and Experimental Subsets

The main composite experiments use the decoupled structure of ImageNet-OOC1k to sample a finite set of object–background pairings along the real-valued visual-semantic axis, with distance defined relative to each background's original class. Table 3 summarizes the dataset composition, including background statistics, object library coverage, and the subsets used in the behavioral and mechanistic experiments. The full pool contains 1,137 backgrounds and 12,321 individual object instances, yielding 14,008,977 possible object–background pairings. This count refers to individual object instances rather than unique classes and does not include additional variations in placement or scale.

For the behavioral experiments (§5.2), we construct a large composite set from a stratified subset of 260 backgrounds, sampled as 10 backgrounds from each of 26 background groups: 12 nature subcategories, 13 human-related subcategories, and one miscellaneous group containing close-ups, ambiguous scenes, and backgrounds that do not fit either category.

For each object class, up to four instances are sampled from the object library when available. The object library spans 918 ImageNet classes. Classes with fewer than four valid instances are still included but contribute fewer composites. Objects are initially scaled by setting their dominant dimension, width for wide objects and height for tall objects, to a uniformly sampled 50–60% of the corresponding background dimension while preserving aspect ratio. Each side is then clamped independently to a minimum of 100 pixels and a maximum of half the corresponding background dimension. Objects are placed at random positions with a 5% margin from image borders.

Object brightness is scaled by a factor sampled uniformly from $[0.9, 1.1]$ and colour saturation by a factor sampled uniformly from $[0.95, 1.05]$. With probability 0.3, objects are rotated by an angle sampled uniformly from $[-10°, 10°]$. Placement, sampled scale, and appearance transformations are generated without reference to visual-semantic rank, and generation settings and per-composite metadata are recorded for reproducibility. This results in 953,862 composites covering the full visual-semantic distance range. Controls for potential compositing artifacts are described in Appendix A.2.

For the mechanistic analysis (§6), we use a stratified subset of 75,000 composites sampled uniformly across visual-semantic distance and background category (20 bins $\times$ 3 categories). Sampling is performed without balancing for correctness, preserving the natural distribution of model behavior within each condition.

The large underlying composite pool (Table 3) enables controlled sampling across visual-semantic space while maintaining diversity in object identity and scene structure.

A.2 Robustness to compositing artifacts

Placement and scale are sampled independently of visual-semantic rank, and each background spans nearly the full rank range. Nevertheless, realized object size, cutout properties, and compositing artifacts could still vary systematically with distance. We therefore test whether these properties explain the accuracy decline using the release dataset of 953,862 composites and all 35 models.

We first collect 18 generation properties from the composite metadata, generation code, and object masks, covering placement, scale, size constraints, and cutout geometry. Eccentricity measures distance from the image centre, while border distance measures the smallest gap between the bounding box and an image edge. Mask area is the fraction of the background covered by object pixels, and mask fill is the fraction of the cutout's bounding box occupied by those pixels. Outline complexity is the mask perimeter divided by the square root of its area. Aspect distortion measures the absolute log change in aspect ratio introduced by the size constraints.

For each property, we correlate its values across composites with their semantic ranks. This tests whether, for example, objects become systematically smaller or more peripheral as semantic distance increases. We also compare the property's mean between the first and last distance bins. To estimate the associated accuracy change, we multiply this difference in means by the univariate least-squares slope of accuracy on that property, using accuracy averaged across models. Table 1 reports these measurements. Correlations near zero and similar first- and last-bin means indicate little systematic variation with distance, while the estimated contributions show the magnitude of the associated accuracy changes relative to the observed decline.

Table 1: **Generation properties across semantic distance.** First and last bins span ranks 1–25 and 976–1000. Contributions are univariate estimates in percentage points, with shares calculated relative to the observed change of -14.46 percentage points. Negative contributions follow the decline, while positive contributions oppose it and therefore have negative shares.

Property	ρ	First-bin mean	Last-bin mean	Contribution (pp)	Share (%)
Placement					
Horizontal centre	≈ 0	0.502	0.500	$+0.001$	-0.01
Vertical centre	≈ 0	0.498	0.497	-0.005	$+0.03$
Eccentricity	$+0.002$	0.173	0.174	-0.063	$+0.44$
Border distance	-0.001	0.132	0.132	$+0.001$	-0.01
Rotation flag	-0.001	0.301	0.296	$+0.009$	-0.06
Scale and size constraints					
Sampled scale factor	-0.001	0.550	0.550	-0.005	$+0.03$
Relative width	-0.013	0.457	0.448	-0.223	$+1.55$
Relative height	$+0.006$	0.454	0.456	$+0.144$	-1.00
Bounding-box area	-0.006	0.205	0.202	-0.383	$+2.65$
Mask area	$+0.012$	0.103	0.103	$+0.108$	-0.75
Aspect ratio	-0.014	1.250	1.208	$+0.235$	-1.63
Aspect distortion	$+0.002$	0.136	0.138	-0.044	$+0.31$
Minimum-size constraint	-0.004	0.102	0.101	$+0.025$	-0.18
Maximum-width constraint	-0.013	0.689	0.654	-0.087	$+0.60$
Maximum-height constraint	$+0.010$	0.607	0.632	$+0.199$	-1.38
Both maximum constraints	-0.002	0.310	0.301	-0.100	$+0.69$
Cutout geometry					
Outline complexity	-0.027	6.875	6.743	$+0.260$	-1.80
Mask fill ratio	$+0.022$	0.526	0.537	$+0.362$	-2.51

The measured properties vary little with semantic rank, with the largest absolute correlation reaching 0.027. Their individual estimated contributions are also small relative to the observed 14.46-percentage-point decline. Bounding-box area gives the largest estimated decrease at 0.383 percentage points, or 2.65% of the observed change. These estimates describe individual associations rather than formal bounds and cannot be summed because the properties overlap.

We next test whether the accuracy decline with semantic distance persists after accounting for object placement, size, and the difficulty of particular backgrounds and cutouts. The idea is to account for these properties when estimating the relationship between distance and accuracy. For each model, we fit a linear regression predicting top-1 correctness from semantic rank, then progressively add the controls shown in Table 2. We first include 11 position and size predictors, comprising the two centre coordinates, eccentricity, border distance, sampled scale, mask area, aspect ratio, and the four size-constraint flags. We then account for some backgrounds being generally easier than others by estimating a separate baseline for each background. Finally, we add a separate baseline for each object instance to account for corresponding differences between cutouts. These background

and object-instance fixed effects capture average differences in difficulty while retaining a common accuracy–distance slope. As an alternative to object-instance effects, we use each model's object-only accuracy for that cutout, averaged over black, white, and $1/f$ noise backgrounds, while retaining the preceding controls. Standard errors are clustered by background and object instance.

We also examine the shape of the relationship after adjustment. Rather than assuming that accuracy changes along a straight line, we divide semantic ranks into 40 bins of width 25 and estimate a separate adjusted accuracy change for each bin relative to the first. We retain the same position and size predictors and background and object-instance effects. The estimated changes are then averaged across all 35 models and compared with the unadjusted changes in observed accuracy.

Specification	Median slope	Retained (%)	CI below 0
Rank only	-0.929	100.0	35/35
+ Placement and scale	-0.979	105.5	35/35
+ Background effects	-0.977	104.4	35/35
+ Object-instance effects	-0.608	66.5	35/35
Object-only alternative	-0.782	87.1	35/35

Table 2: **Accuracy–distance slopes after adjustment.** Each + row adds controls to those above it. The final row replaces object-instance effects with object-only accuracy, retaining the other controls. Slopes are medians across 35 models, in percentage points per 100 ranks. "Retained" is the median percentage of each model's rank-only slope remaining after adjustment. "CI below 0" counts models with entirely negative 95% confidence intervals.

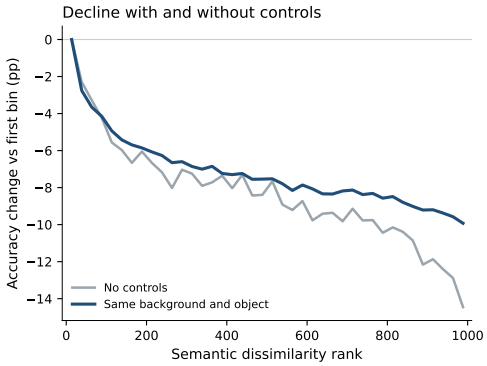


Figure 7: **Accuracy decline after adjustment.** Curves show accuracy changes relative to the first distance bin, averaged across all 35 models. The adjusted curve accounts for placement and scale measurements, background effects, and object-instance effects.

All 35 models retain negative slopes with 95% confidence intervals below zero under every specification (Table 2). Placement and scale controls and background effects leave the slopes largely unchanged. Adding object-instance effects retains a median of 66.5% of the original slope, while the object-only alternative retains 87.1%. The adjusted curve also preserves the decreasing relationship (Figure 7), with a 9.93-percentage-point decline compared with 14.46 points without controls and a Spearman correlation of $\rho = -0.993$ with bin index.

Finally, we test whether discontinuities at object boundaries explain the remaining decline. We reconstruct a separate sample of 30,000 composites covering 3,000 cutouts and measure luminance and colour differences across the object edge, together with local luminance gradients. Each measure is expressed relative to the corresponding region of the background alone. Adding these measures to the regression with placement and scale controls, background effects, and object-instance effects changes the median slope on this sample from -0.585 to -0.583 percentage points per 100 ranks. All 35 models retain confidence intervals below zero. Together, these controls show that the accuracy decline persists after accounting for measured generation properties and average differences between backgrounds and cutouts.

A.3 Object-Only Image Generation

To isolate object-derived recognition signals, we construct object-only images by compositing segmented foreground objects onto neutral backgrounds. Foreground objects are extracted using the corresponding segmentation masks, which are binarized with a threshold of 127 to obtain a hard separation between foreground and background.

Each object is placed onto one of three background conditions: a uniform black background, a uniform white background, or a $1/f$ noise background. For the uniform conditions, all background

pixels are set to a constant intensity value. For the $1/f$ noise condition, backgrounds are generated in the frequency domain by sampling random phase and scaling amplitudes by the inverse of spatial frequency in the amplitude spectrum, followed by an inverse Fourier transform. Noise is generated independently per channel using a deterministic seed derived from the image filename, ensuring reproducibility.

These images are used as controls in §5.1 and evaluated in Figure 3a, with full results across models reported in Appendix C.1. Example object-only images are shown in Figure 1a.

A.4 Release, Licensing, and Ethical Considerations

ImageNet-OOC1k is derived from ImageNet validation images and ImageNet-S segmentation masks. Inpainted backgrounds are generated using `IOPaint` with the MAT inpainting model. Because the underlying images come from ImageNet, ImageNet-OOC1k inherits the access and redistribution constraints of ImageNet/ImageNet-S.

Code, metadata, semantic-distance matrices, and generation specifications are available at `https://github.com/ToastyDom/ImageNet-OOC1k`. Image-derived components are released subject to the access and redistribution terms of ImageNet/ImageNet-S. The accompanying documentation describes component provenance, access requirements, and inherited restrictions. Image-derived components, including backgrounds, object crops, masks, and composites, will be released in forms compatible with the access conditions and redistribution terms of the source datasets. We will not redistribute ImageNet-derived image content except in forms permitted by the source-data terms. The release documents the provenance of each component, access requirements, inherited restrictions, and the distinction between newly introduced assets and components derived from existing datasets.

ImageNet-OOC1k does not involve newly collected human-subject data, crowdsourcing, or participant interaction. The benchmark is intended for diagnostic evaluation of object–background context sensitivity in image recognition models, not for deployment decisions about people or safety-critical systems. Its main potential impact is to support more reliable evaluation of visual recognition models by exposing predictable context-dependent failures. At the same time, improved visual recognition can affect downstream applications whose societal consequences depend on the deployment context, including sensitive settings such as surveillance. The release will document intended use and known limitations, including the inherited scope and biases of ImageNet-style object categories.

A.5 Compute Resources

All experiments were conducted on an internal university HPC cluster using a single NVIDIA A100-SXM4-80GB GPU per job, without multi-GPU parallelization. The evaluated vision models were not trained or fine-tuned. All experiments use publicly available pretrained weights loaded via `timm` (v1.0.26) and OpenAI CLIP. The software stack consisted of Python 3.11.15, PyTorch 2.10.0, CUDA 12.8, and cuDNN 9.10.02.

The compute reported here reflects the specific experimental sampling used in this paper, rather than an intrinsic requirement of the benchmark. In particular, we materialized and evaluated approximately 954K composite images to densely sample semantic distance across backgrounds and object classes. Batched inference across the 35 pretrained models used in the behavioral benchmark, together with 9 control conditions comprising approximately 55K additional images, required approximately 86 GPU-hours using a batch size of 128. The mechanistic analyses required additional extraction of patch-level logit-lens representations for the 75K-image stratified subset across five ViT variants: ViT-S/16, ViT-B/16, ViT-L/16, DeiT-B/16, and ViT-RelPos-B/16. ViT-RelPos-B/16 was included only in this mechanistic robustness analysis and is not part of the 35-model behavioral benchmark. This required approximately 11 GPU-hours.

Dataset generation was comparatively lightweight. Constructing the composite image set used in this paper required approximately 3.5 CPU-hours. Computing the consensus dissimilarity matrix from 10 pretrained classifiers took under 2 minutes on CPU. Background inpainting for the 1,137 source images was performed using `IOPaint`.

Storage requirements were modest and depend on the number of composites materialized. The ImageNet-OOC1k dataset instance used in this paper occupies approximately 98 GB, dominated by composite images stored in HDF5 format (86 GB). The remaining 12 GB includes original images,

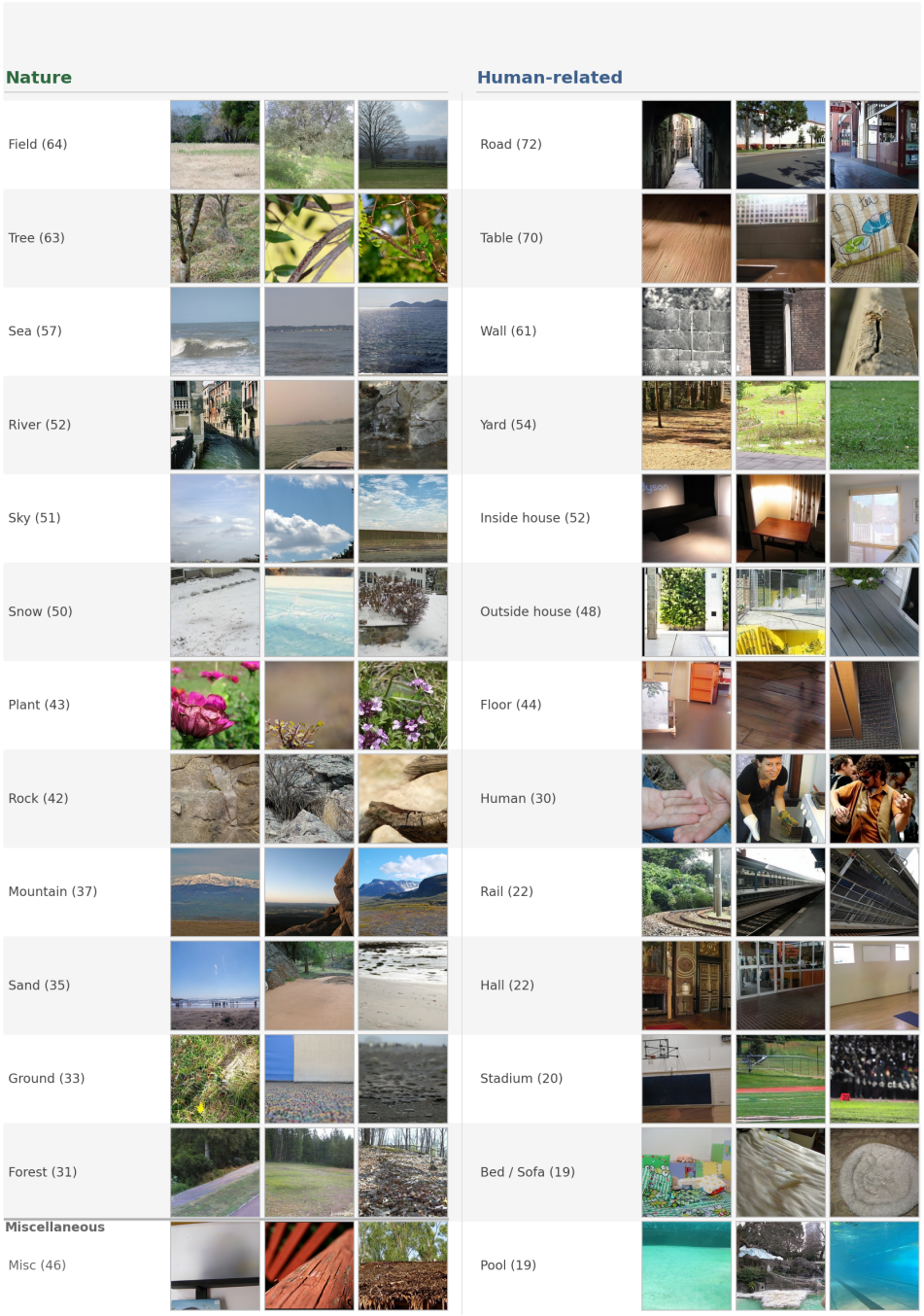


Figure 8: **Background scene subcategories in ImageNet-OOC1k.** Representative inpainted backgrounds organized by category and subcategory, with the number of scenes per subcategory shown in parentheses. The figure shows *nature* environments (12 subcategories, 558 backgrounds), *human-related* environments (13 subcategories, 533 backgrounds), and *miscellaneous* scenes (46 backgrounds). All images show background-only scenes after removal of the original foreground object via inpainting. Subcategories are ordered by decreasing frequency within each category.

inpainted backgrounds, segmentation masks, and the object library. The consensus dissimilarity matrix and per-model matrices occupy 34 MB. Experimental outputs add approximately 10 GB, including prediction CSVs (5.6 GB) and patch-level activation files for five ViT variants (4.7 GB). Peak disk usage across the dataset and all experimental artifacts was approximately 109 GB.

Table 3: **ImageNet-OOC1k dataset and composite generation summary.** Counts are reported for the full dataset, the subset used in behavioral experiments (§5.2), and the subset used for mechanistic analysis (§6).

Component	Detail	Count
Backgrounds		
Nature	field	64
	tree	63
	sea	57
	river	52
	sky	51
	snow	50
	plant	43
	rock	42
	mountain	37
	sand	35
	ground	33
	forest	31
Category total		**558**
Human-related	road	72
	table	70
	wall	61
	yard	54
	inside_house	52
	outside_house	48
	floor	44
	human	30
	rail	22
	hall	22
	stadium	20
	bed_sofa	19
	pool	19
Category total		**533**
Misc	–	46
Total backgrounds		**1,137**
Object Library		
Object image–mask pairs	Total	12,321
Object classes	ImageNet classes	918
	Classes with at least four instances	915
Composites (Behavioral Experiments, §5.2)		
Generated composites	Used in experiments	953,862
Backgrounds used	Stratified, 10 per background group	260
Possible pairings	Full object–background pool, upper bound	$\sim$14.0M
Subset (Mechanistic Analysis, §6)		
Total composites	Transformer analysis	75,000
Sampling strategy	20 bins $\times$ 3 categories $\times$ 1,250/cell	–

In total, the reported experiments required approximately 97 GPU-hours and 4 CPU-hours. Preliminary experiments, including pipeline debugging, parameter tuning, and exploratory analyses, used the same inference-only setup and required an additional estimated 30–50 GPU-hours.

B Semantic Structure of Classifier Representations

Section 4 defines the consensus visual-semantic axis from classifier-head representations. Here, we visualize the class-level geometry underlying this axis and compare it with individual model geometries. The goal is to show that the consensus matrix summarizes a visual-semantic structure already present across models, rather than combining unrelated class organizations.

For the consensus and selected individual models, we show the full class-level dissimilarity matrix and local neighborhoods for three representative classes (hen, holster, coral reef). Across supervised classifiers, semantically related classes form coherent local groups. Although fine-grained neighbor order varies, broad semantic neighborhoods recur across architectures, supporting the use of a consensus axis for the main experiments. For these classifiers, dissimilarities are computed from classifier-head representations as described in Section 4. For comparison, we encode ImageNet class-name prompts, `"a photo of a {class}"`, with the CLIP text encoder, L2-normalize the embeddings, and compute cosine dissimilarity.

B.1 Consensus Representation

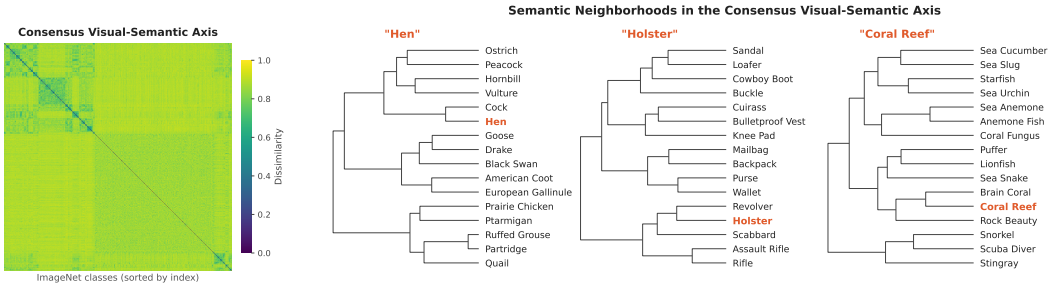


Figure 9: **Consensus visual-semantic structure.** Left: dissimilarity matrix over ImageNet classes derived from the consensus visual-semantic axis. Right: local neighborhoods for the same focal classes as in Figure 2, showing the 15 nearest neighbors clustered using Ward linkage. The consensus aggregates classifier-head representations across the ten pretrained models described in Section 4 and defines the semantic axis used in the main experiments.

B.2 Individual Model Representations

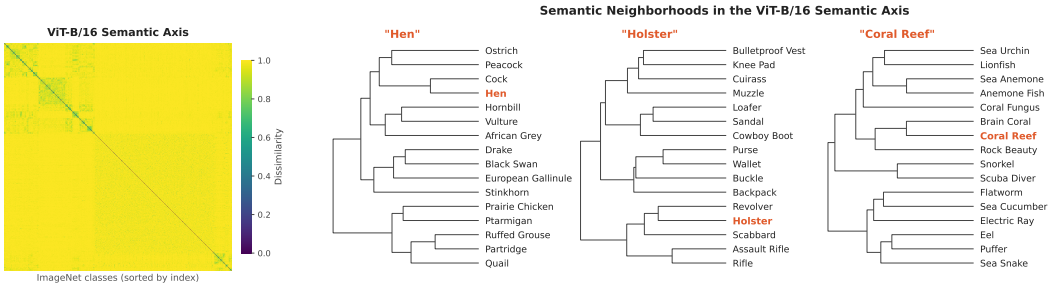


Figure 10: **Semantic structure for ViT-B/16.** Left: dissimilarity matrix over ImageNet classes derived from classifier-head representations. Right: local neighborhoods for the same focal classes as in Figure 2, showing the 15 nearest neighbors clustered using Ward linkage. Semantic groupings closely match the consensus structure, indicating that coherent class neighborhoods already emerge in an individual transformer model.

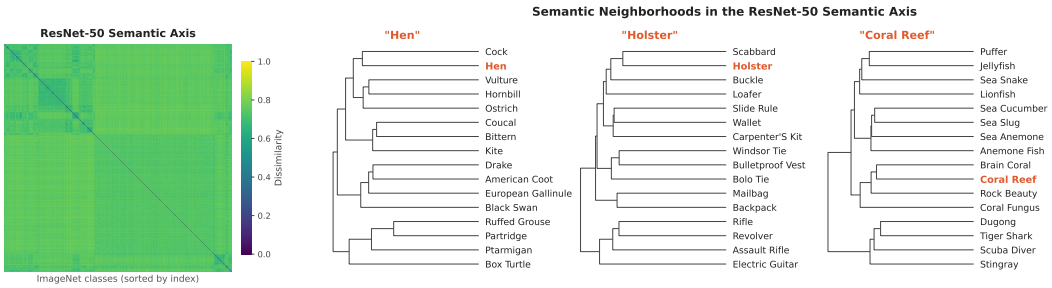


Figure 11: **Semantic structure for ResNet-50.** Left: dissimilarity matrix over ImageNet classes derived from classifier-head representations. Right: local neighborhoods for the same focal classes as in Figure 2, showing the 15 nearest neighbors clustered using Ward linkage. Similar semantic groupings emerge despite architectural differences, consistent with the consensus structure.

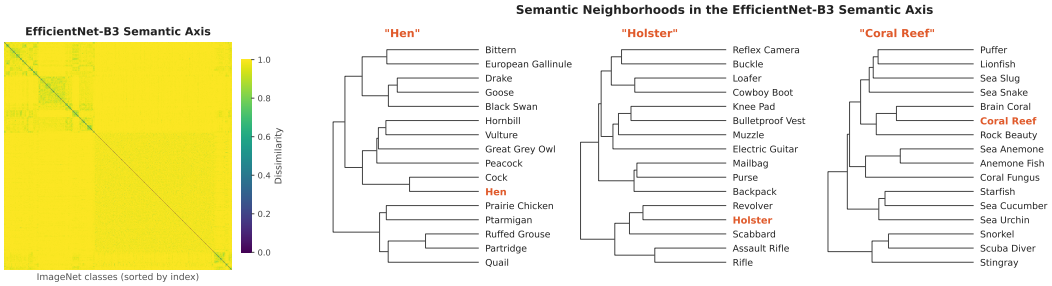


Figure 12: **Semantic structure for EfficientNet-B3.** Left: dissimilarity matrix over ImageNet classes derived from classifier-head representations. Right: local neighborhoods for the same focal classes as in Figure 2, showing the 15 nearest neighbors clustered using Ward linkage. High-level semantic organization is preserved, with variation primarily in fine-grained neighbor ordering.

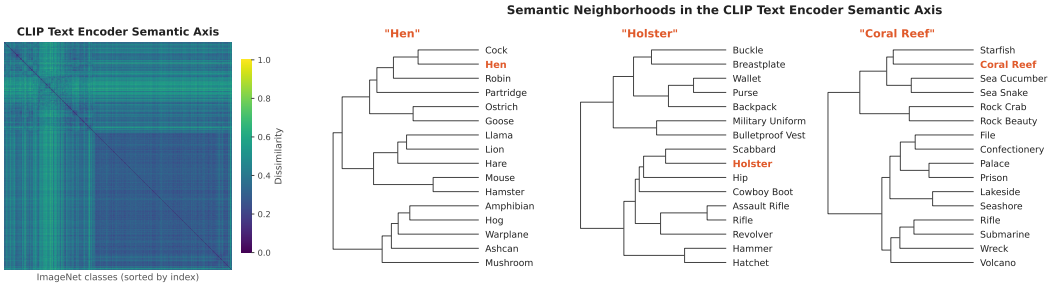


Figure 13: **Semantic structure for CLIP text embeddings.** Left: dissimilarity matrix over ImageNet classes derived from CLIP text encoder embeddings of class-name prompts. Right: local neighborhoods for the same focal classes as in Figure 2, showing the 15 nearest neighbors clustered using Ward linkage. The neighborhoods differ from those derived from classifier-head representations.

Table 4: **Model set used in this study.** Models are grouped into architecture families used throughout the behavioral analyses. Models marked with ⋆ correspond to the six representative models shown in the main-text behavioral figures.

Model	Family	Reference
ViT-S/16 (vit_small_patch16_224)	ViT	[7]
ViT-B/16⋆	ViT	[7]
ViT-B/32 (vit_base_patch32_224)	ViT	[7]
ViT-B/16@384 (vit_base_patch16_384)	ViT	[7]
ViT-L/16⋆	ViT	[7]
DeiT-S/16 (deit_small_patch16_224)	DeiT	[34]
DeiT-B/16⋆	DeiT	[34]
DeiT-III-S (deit3_small_patch16_224)	DeiT	[35]
DeiT-III-M (deit3_medium_patch16_224)	DeiT	[35]
DeiT-III-B (deit3_base_patch16_224)	DeiT	[35]
DeiT-III-L (deit3_large_patch16_224)	DeiT	[35]
BEiT-B (beit_base_patch16_224)	BEiT	[1]
BEiT-L (beit_large_patch16_224)	BEiT	[1]
BEiT-v2-B (beitv2_base_patch16_224)	BEiT	[24]
BEiT-v2-L (beitv2_large_patch16_224)	BEiT	[24]
BEiT-3-B (beit3_base_patch16_224)	BEiT	[37]
BEiT-3-L (beit3_large_patch16_224)	BEiT	[37]
ResNet-34	CNN	[11]
ResNet-50⋆	CNN	[11]
ResNet-101	CNN	[11]
ResNet-50 (SSL)	CNN	[41]
Inception-v3	CNN	[30]
Inception-v4	CNN	[31]
EfficientNet-B2	EfficientNet	[32]
EfficientNet-B3⋆	EfficientNet	[32]
EfficientNet-B4	EfficientNet	[32]
EfficientNetV2-S	EfficientNet	[33]
EfficientNetV2-M	EfficientNet	[33]
ConvNeXt-S	ConvNeXt	[18]
ConvNeXt-B	ConvNeXt	[18]
ConvNeXt-L	ConvNeXt	[18]
ConvNeXt-XL	ConvNeXt	[18]
ConvNeXtV2-B	ConvNeXt	[39]
ConvNeXtV2-L	ConvNeXt	[39]
CLIP ViT-B/16⋆	CLIP	[25]

C Robustness and Generalization of Behavioral Results

This section provides the robustness analyses supporting the behavioral results in Section 5. The goal is to show that the main effects are not specific to the representative models shown in the main text, to a small subset of backgrounds, or to the construction of the consensus visual-semantic axis.

We first define the full model set used throughout the behavioral analyses (Table 4). We then extend the object-only and background-only controls from Section 5.1 to all models, showing that object- and background-derived signals remain separable across architectures. Next, we show that the monotonic accuracy decline with semantic distance holds across the full model set and examine how this global trend decomposes across individual backgrounds. Finally, we compare the consensus axis with alternative similarity spaces to verify that the observed effects are not a generic consequence of ranking classes by any semantic measure.

C.1 Object and Background Signals Across Models

This section extends Figure 3 to the full model set (Table 4). We evaluate both background-only and object-only controls under top-1 and top-5 accuracy to test whether the separation between background-derived and object-derived signals generalizes across architectures.

Background-only accuracy is consistently above chance (0.1%) for all models under top-1 evaluation (Figure 14), confirming that backgrounds alone carry class-specific information. Although these backgrounds come from the ImageNet validation split, recurring object–context associations, such as aquatic animals with water, provide contextual cues that models can learn during training. Under top-5 evaluation, performance increases further, showing that the correct class frequently appears among the top predictions even when the object is absent. The highest background-only accuracies are observed for BEiT and ConvNeXt models, while CLIP and the evaluated CNNs show lower but still above-chance performance.

We next test whether the semantic-neighbor enrichment of incorrect top-1 predictions on background-only images could be explained by non-uniform class density in the consensus space. To do so, we computed a permutation null by shuffling object class labels across backgrounds while preserving the empirical distribution of model predictions. Across 1,000 permutations, the expected fraction of incorrect top-1 predictions falling within the top-25 semantic neighbors of the ground-truth class was 2.86%, compared to 2.50% under a uniform class baseline. The observed fraction was 25.48%, corresponding to an 8.9-fold enrichment over the permutation null. None of the permutations reached the observed value ($p < 0.001$). This enrichment was consistent across models, ranging from 7.2-fold to 10.9-fold at $K = 25$.

Figure 16 shows that semantic-neighbor enrichment extends across architectures and prediction positions. For incorrect top-1 predictions (panel a), the mean fraction within the ground-truth class's 25 nearest visual-semantic neighbors is 25.6%, with family means ranging from 19.8% for CLIP to 28.5% for BEiT. For incorrect classes at each of the top-five prediction positions (panel b), this fraction ranges from 25.5–29.0%, compared with a uniform baseline of 2.5%. Even when the ground-truth class is absent from the entire top five, 16.6% of predicted classes fall within this neighborhood. Predictions outside these 25 neighbors can describe visible background content, such as *seashore* for coastal scenes or *valley* for mountains.

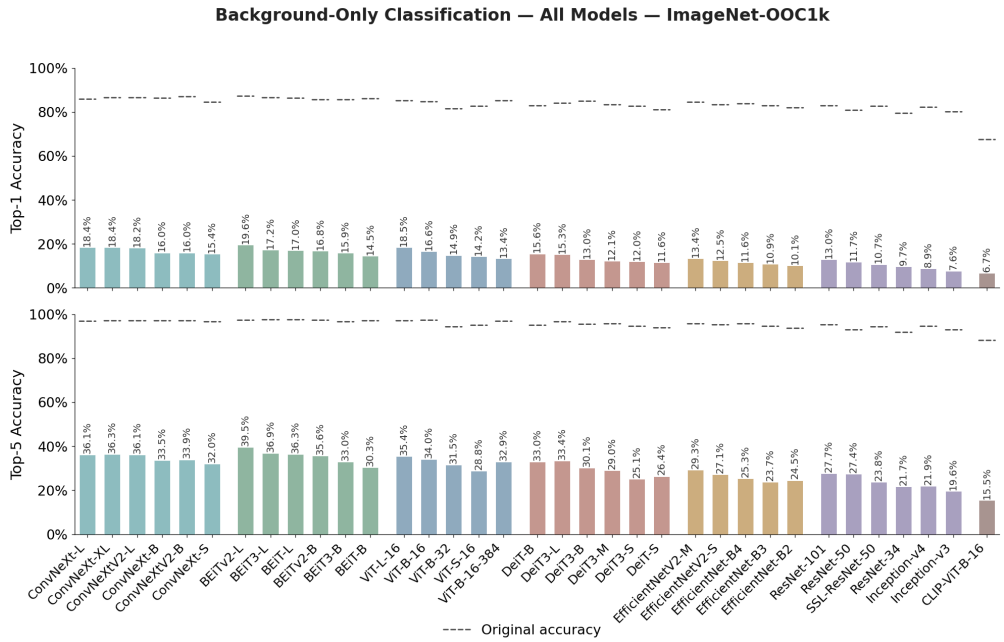


Figure 14: **Background-only accuracy across all models.** Rows show top-1 and top-5 accuracy for background-only images, with models grouped by architecture family and ordered by mean performance. All models perform above their respective chance baselines (0.1% for top-1 and 0.5% for top-5), showing that backgrounds alone carry class-specific signal. Top-5 accuracy increases substantially across models, indicating that the correct class often appears among the leading predictions even when the object is absent. The highest accuracies are observed for BEiT and ConvNeXt models, while CLIP and the evaluated CNNs show lower performance.

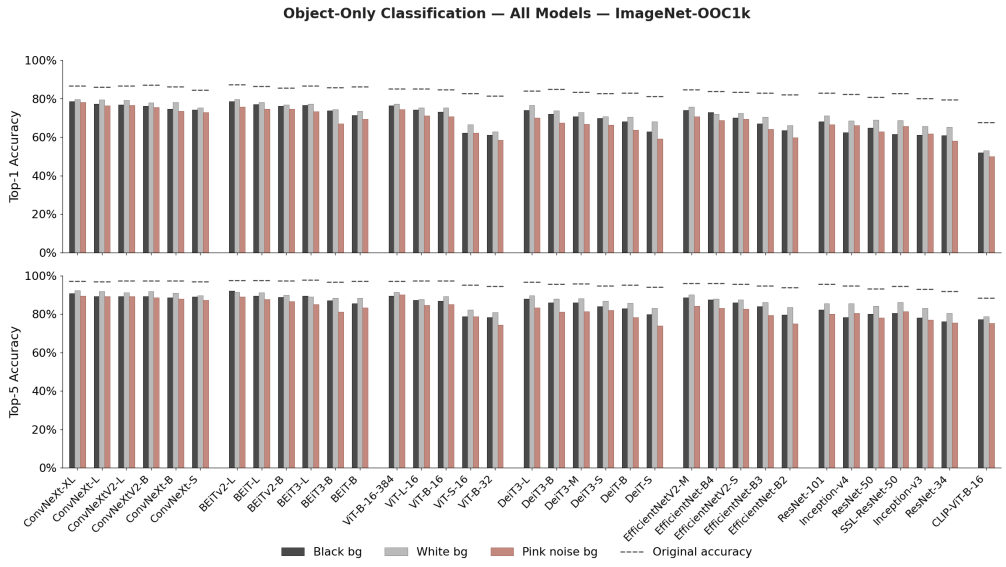


Figure 15: **Object-only accuracy across all models and neutral background conditions.** Rows show top-1 and top-5 accuracy for objects composited onto black, white, and $1/f$ noise backgrounds. Accuracy varies little across neutral background conditions in both evaluations, indicating that object-derived evidence is largely preserved under these manipulations. The same pattern holds across the full model set.

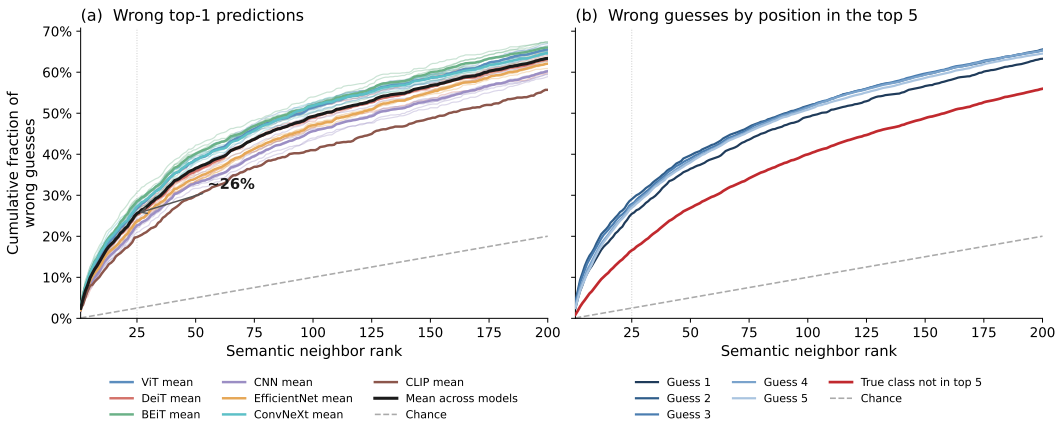


Figure 16: **Semantic structure of background-only predictions across all 35 models. (a)** Cumulative visual-semantic neighbor-rank curves for incorrect top-1 predictions. Thin colored lines show individual models, thicker colored lines show architecture-family means, and the black line shows the mean across models. **(b)** Corresponding curves for incorrect classes ranked first through fifth in the model predictions. The red curve pools all five predictions from images where the ground-truth class is absent from the top five. In both panels, the ground-truth class is excluded from the neighbor ranking, so rank 1 denotes its nearest other class. Dashed lines show the uniform baseline over the remaining 999 classes.

C.2 Universality Across Models

To test whether the accuracy–distance pattern observed in Figure 1c generalizes beyond the representative models shown in the main text, we repeat the same analysis across the full 35-model set listed in Table 4. Figure 17 shows that all models exhibit the same monotonic decline in recognition accuracy as semantic object–background distance increases. Although models differ substantially in baseline accuracy and in the magnitude of performance degradation, the functional form remains closely aligned across convolutional networks, vision transformers, and CLIP.

Across models, top-1 accuracy decreases by approximately 9–19 percentage points from the most similar to the most dissimilar composites. To verify that this trend is not an artifact of binned visualization, we fit an image-level logistic regression for each model, predicting correctness from semantic distance rank. All 35 models yielded strictly negative slopes, with all 95% confidence intervals below zero (mean $\beta = -0.00038$ per rank; range $[-0.00048, -0.00028]$). This corresponds to an odds ratio of approximately 0.963 per 100-rank increase, confirming that recognition probability decreases continuously with semantic distance at the image level.

C.3 Heterogeneity Across Backgrounds

To understand how the global accuracy–distance relationship arises, we analyze how it varies across individual backgrounds. For each of the 260 backgrounds, we compute a per-background interference score that summarizes the relationship between semantic distance and recognition accuracy. We collect all composites generated for a given background across the full range of semantic dissimilarity ranks and evaluate top-1 accuracy using five supervised models (ViT-B/16, ViT-L/16, DeiT-B/16, ResNet-50, and EfficientNet-B3). CLIP is excluded from this analysis because of its substantially lower baseline accuracy in this setting, although it shows the same qualitative trend. To reduce noise from sparse samples at individual ranks, we group ranks into bins of width 50, average accuracy within each bin across images and then across models, and compute the Spearman correlation ρ between binned dissimilarity rank and accuracy. This yields one scalar interference score per background. Strongly negative values indicate decreasing accuracy with semantic distance, values near zero indicate no systematic effect, and positive values indicate facilitation, where accuracy increases with semantic distance.

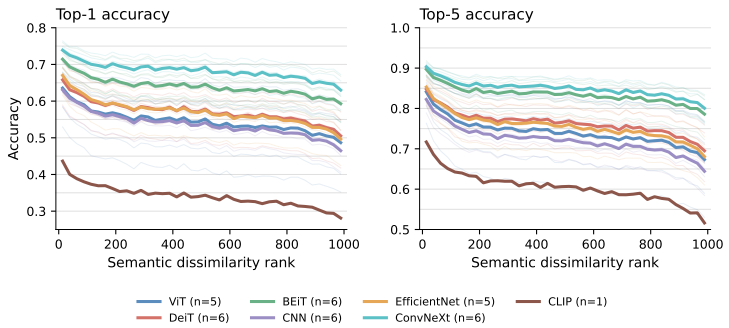


Figure 17: **Recognition accuracy as a function of semantic distance across all models.** Thin lines show individual models and bold lines show architecture family means. All 35 models exhibit a consistent monotonic decline for both top-1 and top-5 accuracy as semantic object–background distance increases. Distances are grouped into bins of width 25 ranks for visualization.

The resulting scores form a broad distribution (mean $\rho = -0.32$, median $\rho = -0.49$, $\sigma = 0.52$). For visualization, we partition backgrounds into four descriptive bins using fixed thresholds on ρ, corresponding to strong negative ($\rho \leq -0.50$, $n = 129$), mid-negative ($-0.50 < \rho \leq 0.00$, $n = 55$), mid-positive ($0.00 < \rho \leq 0.50$, $n = 51$), and strong positive ($\rho > 0.50$, $n = 25$) relationships. These bins separate backgrounds with strong decreasing, weak decreasing, weak increasing, and strong increasing accuracy–distance trends.

Most backgrounds exhibit negative correlations and contribute to the aggregate monotonic decline, while a smaller subset shows weak or positive correlations, corresponding to reduced or reversed sensitivity to semantic mismatch (Figure 18). The monotonic decline therefore refers to accuracy averaged across backgrounds rather than a relationship that holds for every individual background. This variation is structured by scene type. Strong-interference backgrounds are predominantly nature scenes, whereas weak or positive effects occur more often in human-related environments.

The reasons for these background-specific differences remain unclear. An approximate grouping based on WordNet hypernyms shows that human-made and natural objects account for 49.3% and 46.4% of library images, respectively. Every background is paired with the same 918 object classes, so human-related backgrounds were not assigned a broader overall object pool than nature backgrounds. However, these observations do not establish which object–background pairings are plausible. Nature scenes may impose stronger expectations about which objects belong in them, whereas more heterogeneous human-related environments may support a broader range of plausible objects. This remains a hypothesis, and determining how object plausibility and scene diversity contribute to the observed trends requires further investigation.

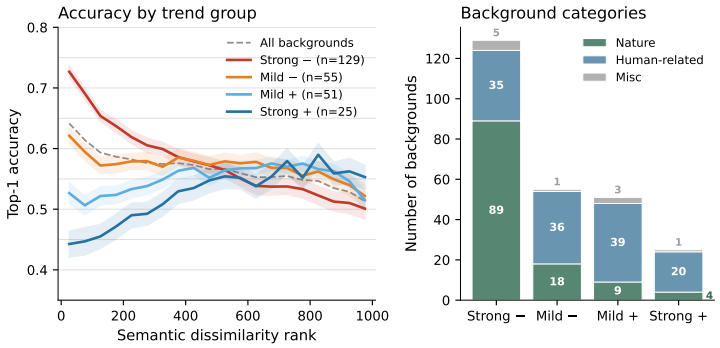


Figure 18: **Background-dependent modulation of semantic distance effects. (a)** Accuracy as a function of semantic distance, grouped by background-specific interference score ρ. Distances are grouped into bins of width 50 ranks for visualization. Shaded bands show ± 1 standard error of the mean across models. **(b)** Distribution of background categories across interference groups. Most backgrounds show negative accuracy–distance trends, while a smaller subset exhibits weak or reversed effects.

C.4 Validation of the Visual-Semantic Axis

This section supports the axis construction in Section 4 and the accuracy–distance result in Section 5.2. To assess the robustness of our findings and address concerns about circularity, we examine whether the decline persists in non-contributing models and under axes derived from independently pretrained representations and alternative definitions of class similarity.

Model overlap. The consensus axis uses ten classifier heads, eight from evaluated models, which we call *contributors*. If direct model overlap drives the observed pattern, we would expect particularly steep declines among contributors. We therefore use the same consensus axis for all 35 evaluated models and compare their mean first-to-last-bin accuracy declines. Contributors are compared with non-contributing siblings from the ResNet and EfficientNet families and models from families not represented in the consensus (Table 5). Non-contributing siblings show the largest decline, while independent families also show a clear decline. The pattern therefore extends beyond contributors, and direct model overlap does not explain why the decline occurs across all 35 models.

Evaluated-model group	n	Mean decline (pp)
Axis contributors	8	15.4
Non-contributing siblings	6	17.2
Independent families	21	13.3
All models	35	14.5

Table 5: **Accuracy decline by participation in consensus construction.** Decline is first-bin minus last-bin top-1 accuracy, using bins of width 25 ranks and averaged across models within each group. Siblings are non-contributing models from the ResNet and EfficientNet families. Independent families comprise DeiT, BEiT, ConvNeXt, Inception, and CLIP. EfficientNet-B5 and DenseNet-121 contribute to the consensus but are not among the evaluated models.

Own-axis comparisons. If direct reuse of a model's classifier head drives the effect, we would expect that model to decline more strongly along its own axis. We therefore keep each model's predictions fixed and re-rank the composites using each contributing classifier-head axis. We compare the decline obtained from the model's own head with its mean decline under other heads (Table 6).

Among the seven contributors other than ResNet-101, only three decline more strongly along their own axis, and the mean difference is small at 0.3 percentage points. Their own classifier geometry therefore does not consistently produce a larger decline.

ResNet-101 appears to be an exception, declining by 24.3 percentage points along its own axis compared with 10.9 under other axes. However, applying the ResNet-101 axis to the other 34

evaluated models also produces a large mean decline of 20.5 percentage points. Its axis therefore produces steep declines for other models as well, so the large own-axis result alone does not demonstrate a selective match between ResNet-101 and its own classifier geometry.

Model	Own axis	Other axes	Difference
EfficientNet-B3	13.6	11.9	+1.7
ResNet-50	11.5	10.5	+1.0
ViT-B/16	9.8	8.8	+0.9
ViT-B/32	14.0	14.0	0.0
ViT-L/16	9.3	9.4	−0.1
ViT-S/16	13.6	14.2	−0.7
ViT-B/16@384	7.9	8.7	−0.9
Mean, excluding ResNet-101	11.4	11.1	+0.3
ResNet-101	24.3	10.9	+13.5

Table 6: **Own-axis versus other-axis accuracy decline.** All values are first-bin minus last-bin top-1 accuracy in percentage points, using bins of width 25 ranks. Own-axis decline uses distances from the evaluated model's classifier head. Other-axis decline averages the declines obtained under other contributing heads. For the seven contributors other than ResNet-101, this mean excludes both the model's own head and the ResNet-101 head, averaging over eight axes. For ResNet-101, it averages over the other nine heads. Differences are computed before rounding.

Visual class representations. A remaining concern is that the original heads all come from supervised ImageNet classifiers. We therefore construct axes from frozen DINOv1 ViT-B/16, DINOv2 ViT-B/14, and MAE ViT-B/16 representations, none of which contributed to the consensus or appeared in the behavioral benchmark. DINOv1 and MAE were pretrained on ImageNet images without labels, while DINOv2 used the broader LVD-142M dataset, which includes ImageNet-derived images. For each backbone, we train a linear layer on frozen pre-logit features using ImageNet labels and cross-entropy loss. We train each probe for 20 epochs with AdamW, a learning rate of 10^{-3}, and weight decay of 10^{-4}. We then construct axes from the probe weight rows using the same procedure as for the original classifier heads. The three probe axes are also combined using the consensus procedure from Section 4.

All three independently constructed probe axes reproduce the declining trend (Figure 19b). Their consensus yields a 12.0 percentage-point decline, compared with 14.5 for the original consensus. Across all 35 models, declines under the two consensuses correlate at $r = 0.969$, despite only moderate agreement between the individual probe axes and the original consensus ($\rho = 0.24$–0.56). Table 7 reports the individual results and held-out probe accuracies. DINOv1 and DINOv2 provide the cleaner tests because the MAE probe has substantially lower held-out accuracy.

We also construct retrieval axes from the same frozen DINOv2 representation. Instead of learning to distinguish classes, retrieval asks how visually similar their image examples are. Using 50 images per class from the ImageNet training split, we compare each image with images from another class and record either its highest similarity or the average similarity of its five closest matches. We average these values over images to obtain class similarities, which are then converted to symmetric dissimilarities. Both retrieval axes produce U-shaped curves ending above their starting accuracy, unlike the declining probe curve (Figure 19c).

Linguistic and taxonomic similarities. Finally, we examine how recognition accuracy varies when class similarity reflects linguistic or taxonomic relationships (Figure 20). CLIP text embeddings produce a different curve from the visual classifier axes. Mean accuracy initially decreases but then rises above its starting value, with most individual models recovering toward the far end. CLIP itself retains an overall decline along its text axis, consistent with its joint image–text training. WordNet shortest-path distance yields a U-shaped relationship across individual models, including CLIP, rather than a consistent decrease.

Together, these controls demonstrate that the decline generalizes beyond the original consensus and contributing models, reproducing across independently pretrained discriminative representations. This convergence supports a shared visual-semantic structure relevant to recognition under object–background conflict. Classifier heads and probes capture directions that distinguish competing classes, directly addressing our question of which class the model selects when object and background

Axis	Probe accuracy	Agreement ρ	Decline (pp)
Original consensus	—	1.00	14.46
DINOv1 probe	74.7%	0.25	8.11
DINOv2 probe	81.2%	0.24	10.19
MAE probe	40.7%	0.56	12.46
Probe consensus	—	0.50	12.00
Retrieval, best match	—	0.43	−2.24
Retrieval, top five	—	0.43	−2.09

Table 7: **Accuracy–distance relationships under probe and retrieval axes.** Probes are fitted using 90 images per class and evaluated on ten held-out images per class, all sampled from the ImageNet training split. Agreement is the Spearman correlation with the original consensus over unique off-diagonal class pairs. Decline is first-bin minus last-bin top-1 accuracy, averaged across all 35 models. Negative values indicate that accuracy ends above its starting value.

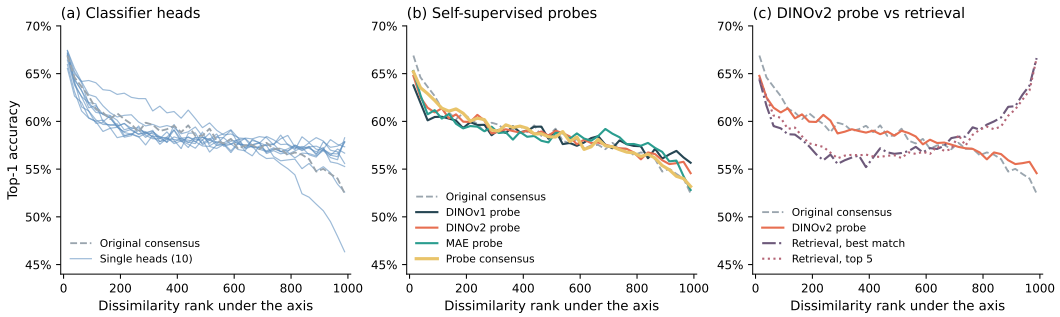


Figure 19: **Accuracy–distance relationships under visual classifier and retrieval axes.** Each curve shows mean top-1 accuracy across the same 35 models and 953,862 composites, re-ranked under the given axis and grouped into bins of width 25. Dashed gray curves show the original consensus. **(a)** The ten contributing classifier-head axes, each shown as a thin blue curve. **(b)** Axes derived from linear probes on frozen DINOv1, DINOv2, and MAE representations, together with their probe consensus. **(c)** The DINOv2 probe axis compared with best-match and top-five retrieval axes derived from the same frozen representation. Probe axes reproduce the declining trend, while retrieval axes produce non-monotonic curves.

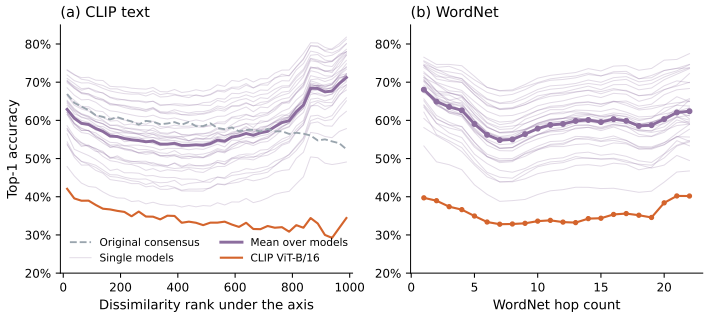


Figure 20: **Accuracy–distance relationships under linguistic and taxonomic similarities.** Thin lines show individual evaluated models, the bold purple line shows the mean across all 35 models, and orange highlights CLIP ViT-B/16. Both panels use the same 953,862 composites. **(a)** Distances derived from CLIP text embeddings of ImageNet class-name prompts, grouped into bins of width 25 ranks. The dashed gray curve shows mean accuracy under the original consensus axis. Most models recover at larger CLIP-text distances, while CLIP itself retains a declining trend. **(b)** Accuracy by WordNet shortest-path hop count, showing only hop counts containing at least 1,000 composites. The relationship is U-shaped across models, including CLIP.

evidence conflict. Retrieval, linguistic, and taxonomic measures describe other relationships between classes, so their different curves are compatible with this task-specific interpretation. These results support the consensus visual-semantic axis as a measure of shared discriminative visual class structure over ImageNet categories. Its alignment with human similarity judgments remains to be tested.

D Spatial Diagnostics and Recovery of Aggregation Failures

This section provides the definitions and additional analyses supporting the aggregation-failure results in Section 6. We first define the core metrics used throughout the analysis, including semantic pull, realm balance, and classification proximity. We then define the outcome taxonomy used to separate clean wins, contested wins, found-but-lost cases, and no-detection cases. Finally, we characterize the spatial structure of found-but-lost errors and describe the re-ranking strategy, operating regimes, and recovery results shown in Figure 6.

D.1 Core Metrics

We use three core metrics to track how object- and background-related evidence evolves across layers. Semantic pull measures whether a patch prediction is closer to the object class or the background class. Realm balance summarizes the fraction of decided patches favoring the object. Classification proximity measures how close the CLS prediction is to the true object class in the consensus space.

For each patch token p at layer ℓ, we obtain a logit-lens prediction $\hat{c}_p^{(\ell)}$. We define **semantic pull** as

$$s_p^{(\ell)} = D^{\text{cons}}\left(\hat{c}_p^{(\ell)}, c_{\text{bg}}\right) - D^{\text{cons}}\left(\hat{c}_p^{(\ell)}, c_{\text{obj}}\right),$$

where D^{cons} is the consensus dissimilarity matrix, c_{obj} is the ground-truth object class, and c_{bg} is the class originally associated with the background. Positive values indicate predictions closer to the object class than to the background class, while negative values indicate predictions closer to the background class.

A patch is considered decided if $|s_p^{(\ell)}| > \tau$. We set $\tau = 0.0945$, corresponding to the mean standard deviation of semantic pull across early layers (blocks 0–7), where patch predictions are uninformative. Object-realm patches satisfy $s_p^{(\ell)} > \tau$, and background-realm patches satisfy $s_p^{(\ell)} < -\tau$.

For image i, we define **realm balance** as the fraction of decided patches favoring the object,

$$\text{RB}_i^{(\ell)} = \frac{|\{p : s_p^{(\ell)} > \tau\}|}{|\{p : s_p^{(\ell)} > \tau\}| + |\{p : s_p^{(\ell)} < -\tau\}|}.$$

Values above 0.5 indicate that more decided patches favor the object than the background, while values below 0.5 indicate background dominance.

For the CLS token, we define **classification proximity** to the object class as

$$\text{CP}^{(\ell)} = 1 - D^{\text{cons}}\left(\hat{c}_{\text{CLS}}^{(\ell)}, c_{\text{obj}}\right),$$

where $\hat{c}_{\text{CLS}}^{(\ell)}$ is the logit-lens prediction of the CLS token at layer ℓ. Higher values indicate that the CLS prediction is closer to the true object class in the consensus space.

D.2 Outcome Taxonomy

We define a four-way outcome taxonomy to characterize model behavior at the image level. All definitions are evaluated at post-norm. For each image i, let e_i denote the number of patches (out of 196) whose argmax prediction exactly matches the object class, and let RB_i denote realm balance at post-norm.

Each image is assigned to one of four mutually exclusive categories:

- **Clean win:** correct prediction with $e_i \geq 20$ and $\text{RB}_i > 0.75$
- **Contested win:** correct prediction but not a clean win

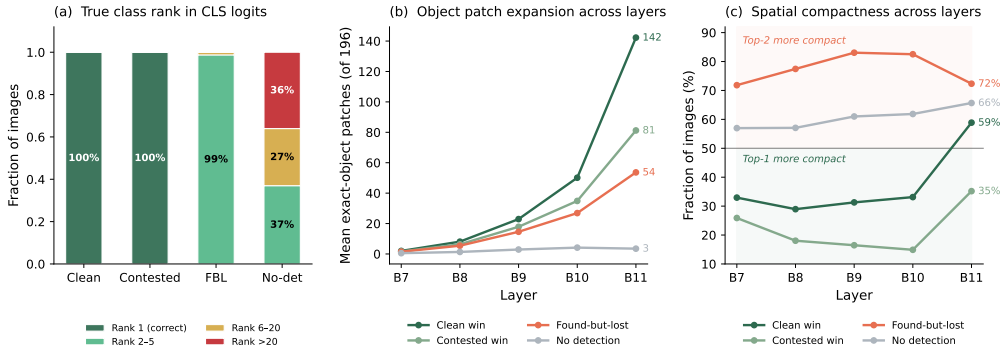


Figure 21: **Diagnostic signatures of aggregation failure.** All panels use the 75,000-composite stratified subset analyzed for ViT-B/16, split by outcome category. **(a)** Distribution of the true object-class rank in post-norm CLS logits. FBL cases are near-miss errors, with the true class typically ranked close to the final prediction. **(b)** Mean number of patches whose logit-lens argmax matches the object class across transformer blocks. Clean wins show broad expansion of object-aligned patches, FBL cases show partial expansion, and no-detection cases remain near baseline. **(c)** Fraction of images where the runner-up CLS class (top-2) is more spatially compact than the predicted CLS class (top-1). Values above 0.5 indicate that top-2 is more compact more often, while values below 0.5 indicate that top-1 is more compact more often. Compactness is measured as the mean distance of supporting patches from their centroid. FBL cases show a strong top-2 compactness advantage at intermediate layers, indicating a localized object signal that has not been fully aggregated.

- **Found-but-lost:** incorrect prediction with $e_i \geq 20$
- **No detection:** incorrect prediction with $e_i < 20$

The detection threshold $k = 20$ captures 98.9% of correct images; the remaining 1.1% of correct cases with $e_i < 20$ are included in the contested win category. Results are qualitatively stable for $k \in [20, 30]$.

D.3 Diagnostic Signatures of Aggregation Failure

Figure 21 extends the analysis in Section 6.2 by characterizing patterns in found-but-lost (FBL) cases that are consistent with aggregation limitations. All analyses are performed on ViT-B/16 using the logit lens on 75,000 composite images, with outcomes defined at post-norm using the taxonomy in Section D.2.

Panel (a) shows that FBL cases are near-miss errors. For each image, we examine the rank of the true object class in the CLS logits at post-norm. While clean and contested wins are rank 1 by definition, FBL cases typically keep the correct class highly ranked despite the incorrect final prediction. The true class lies within the top-5 predictions in 98.7% of FBL cases, with median rank 2, compared to 37% and median rank 9 for no-detection cases. This confirms that FBL cases often contain the correct answer in the CLS representation, but fail to select it.

Panel (b) shows that this near-miss structure is accompanied by partial spatial expansion. Clean-win cases expand from approximately 2 to 142 exact object-class patches across layers, indicating that object evidence spreads broadly across the grid before the final prediction. FBL cases also detect the object, but expand only to around 54 patches. No-detection cases remain near baseline, increasing only from approximately 0.5 to 3.5 patches. Thus, the difference is not simply whether the object is found, but whether object evidence expands enough to dominate the spatial representation.

Panel (c) asks whether incomplete expansion leaves a recoverable spatial signature. The intuition is that object evidence in FBL cases may remain spatially localized, forming a compact cluster of supporting patches, while the competing predicted class is more spatially dispersed. We measure spatial spread as the mean Euclidean distance of supporting patches from their centroid on the 14×14 grid, as formalized in Section E. Lower spread indicates a more compact class-specific patch cluster.

The plot reports the fraction of images in which the runner-up class (top-2) is more compact than the predicted class (top-1). Values above 0.5 therefore indicate that the runner-up is more compact more often, while values below 0.5 indicate that the predicted class is more compact more often. In FBL cases, the runner-up is often the true object class, and it is more compact than the predicted class at intermediate layers, especially blocks 9–10. At block 10, this occurs in approximately 82% of FBL cases, compared to about 33% of clean wins. Clean-win cases show the opposite pattern, with the predicted class forming the more compact cluster. Thus, FBL cases contain a compact and localized object signal that has not been selected by the CLS token. The gap collapses at post-norm to roughly 13 percentage points, 72% versus 59%, as representations become fully aggregated. This identifies block 10 as a stage where aggregation is still unresolved and object evidence remains recoverably structured.

E Spatial Re-Ranking as a Recoverability Probe

To test whether FBL errors contain recoverable structure, we evaluate simple re-ranking rules that swap the CLS top-1 prediction with the top-2 prediction under different gating conditions. This is intended as a diagnostic probe, not as a deployment-time correction method. Although the rules are designed around FBL cases, a swap can also incidentally correct other errors or break originally correct predictions. Table 8 reports these tradeoffs.

Spatial spread. Let $\hat{c}_1$ and $\hat{c}_2$ denote the top-1 and top-2 classes from the post-norm CLS logits. For a candidate class c at transformer block ℓ, we define its patch support as the set of patches whose logit-lens prediction equals c,

$$\mathcal{P}_c^{(\ell)} = \{p : \hat{c}_p^{(\ell)} = c\}.$$

The spatial spread of c is the mean Euclidean distance of its supporting patches from their centroid on the $H \times W$ patch grid,

$$\mathrm{spread}(c, \ell) = \frac{1}{|\mathcal{P}_c^{(\ell)}|} \sum_{p \in \mathcal{P}_c^{(\ell)}} \left\| \mathbf{x}_p - \bar{\mathbf{x}}_c^{(\ell)} \right\|_2, \qquad \bar{\mathbf{x}}_c^{(\ell)} = \frac{1}{|\mathcal{P}_c^{(\ell)}|} \sum_{p \in \mathcal{P}_c^{(\ell)}} \mathbf{x}_p,$$

where $\mathbf{x}_p = (r_p, q_p)$ denotes the patch-grid coordinates. If $|\mathcal{P}_c^{(\ell)}| = 0$, spread is set to the maximum grid diagonal, $\sqrt{(H-1)^2 + (W-1)^2}$. If $|\mathcal{P}_c^{(\ell)}| = 1$, spread is set to zero.

Gating signals. Each rule combines up to three signals motivated by the FBL signature described in Section 6.2. The spatial compactness advantage a is measured at the penultimate transformer block ℓ_{pen}, block 10 for ViT-B/16,

$$a = \mathrm{spread}(\hat{c}_1, \ell_{\mathrm{pen}}) - \mathrm{spread}(\hat{c}_2, \ell_{\mathrm{pen}}).$$

Positive values indicate that the runner-up class forms a tighter patch cluster than the top-1 class. The logit gap is

$$g = z_{\hat{c}_1} - z_{\hat{c}_2},$$

where $z_{\hat{c}_1}$ and $z_{\hat{c}_2}$ are the post-norm CLS logits. Smaller values indicate a more uncertain final decision. The semantic distance is

$$d = D^{\mathrm{cons}}_{\hat{c}_1, \hat{c}_2},$$

computed from the consensus dissimilarity matrix.

Re-ranking rules. A swap replaces the top-1 prediction with the top-2 prediction when the selected gating conditions are satisfied. Unprotected rules apply the swap when

$$a > a_{\min} \qquad \text{and} \qquad g < g_{\max}.$$

Protected rules add a semantic-distance constraint,

$$d > d_{\min},$$

to suppress near-neighbor swaps and reduce breakage.

Results. The threshold settings in Table 8 illustrate the recovery–breakage tradeoff. Stricter unprotected settings yield high precision but recover fewer FBL cases, while relaxed settings recover more FBL cases at the cost of more broken correct predictions. Adding the semantic-distance constraint improves this tradeoff. The selected rule, $a > 1.0$, $g < 1.0$, and $d > 0.5$, recovers 4,277 FBL cases, incidentally corrects 428 non-FBL errors, and breaks 456 correct predictions. This yields a 9.4:1 FBL recovery-to-breakage ratio and a net gain of +5.7 percentage points over the 75,000-image ViT-B/16 diagnostic subset. Since thresholds are selected on the same subset used for the diagnostic analysis, this gain should be interpreted as an in-sample estimate of recoverability. Increasing $d_{\min}$ further improves the recovery-to-breakage ratio while preserving a similar net gain.

Together, these results show that FBL errors leave a structured spatial signature that can be detected from intermediate representations. The recovery analysis supports the interpretation that, in this subset of failures, object information is present but not selected by the final aggregation step.

Table 8: **Re-ranking strategies and performance tradeoffs.** Each rule specifies thresholds on spatial compactness advantage (a), logit gap (g), and optionally semantic distance (d). The swap from the top-1 to the top-2 prediction is applied only when all conditions are satisfied. "FBL recovered" denotes found-but-lost errors corrected by the swap, "Other corrected" denotes non-FBL errors corrected incidentally, and "Broken" denotes originally correct predictions made incorrect. "Ratio" is computed as "FBL recovered" divided by "Broken". Net Δ is computed over all 75,000 images. See Figure 6 for the corresponding visualization.

Rule	FBL recovered	Other corrected	Broken	Ratio	Net Δ
$a > 2.0$, $g < 0.5$	1,786	171	185	9.7:1	+2.4pp
$a > 1.0$, $g < 1.0$	4,523	518	953	4.7:1	+5.5pp
$a > 1.0$, $g < 1.5$	5,601	833	1,574	3.6:1	+6.5pp
$\star$ $a > 1.0$, $g < 1.0$, $d > 0.5$	**4,277**	**428**	**456**	**9.4:1**	**+5.7pp**
$a > 1.0$, $g < 1.0$, $d > 0.6$	4,184	397	341	12.3:1	+5.7pp
$a > 1.0$, $g < 1.0$, $d > 0.7$	4,089	363	265	15.4:1	+5.6pp

F Cross-Architecture Robustness of Aggregation-Failure Diagnostics

This section tests whether the aggregation-failure diagnostics identified in ViT-B/16 are specific to that architecture or to CLS-token pooling. We repeat the core analyses across four additional transformer variants: ViT-S/16, ViT-L/16, DeiT-B/16, and ViT-RelPos-B/16 [17, 38]. ViT-RelPos-B/16 uses average pooling rather than CLS-token pooling, allowing us to test whether similar signatures arise when the final prediction is computed from pooled patch representations rather than a dedicated aggregation token. For this model, we compute the top-1 and top-2 classes from the mean of the final-layer patch logits. FBL cases therefore identify errors in which object-aligned patches are present but the averaged logits favor another class.

F.1 Outcome Composition Across Architectures

We first compare the outcome taxonomy across architectures. Figure 22 shows that the same four outcome types appear across models, but their relative frequencies differ substantially. ViT-S/16 has the largest no-detection fraction and the smallest FBL fraction, indicating that many of its errors reflect weak or absent object evidence rather than recoverable aggregation failure. In contrast, ViT-B/16 and ViT-RelPos-B/16 show the largest FBL fractions, indicating a larger pool of errors where object evidence is present at the patch level but not selected by the final prediction.

These differences in error composition help explain why the recovery probe transfers unevenly across models. Models with larger FBL pools provide more opportunities for re-ranking to recover object evidence, whereas models dominated by no-detection errors leave less recoverable structure for the probe to exploit.

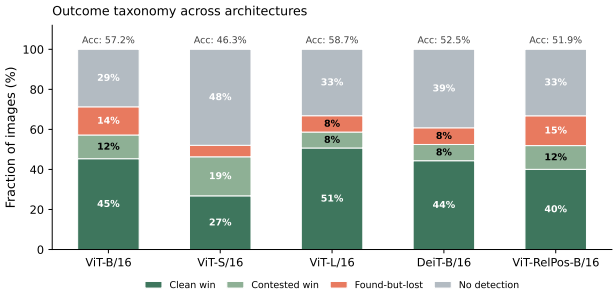


Figure 22: **Outcome taxonomy across architectures.** Each bar shows the fraction of composite images assigned to each outcome category: clean win, contested win, found-but-lost (FBL), and no detection. Outcomes are defined using patch-level predictions at post-norm as described in Appendix D.2. Baseline accuracy is shown above each bar. ViT-S/16 shows the highest no-detection rate and the lowest FBL rate, indicating that many of its errors reflect weak or absent object evidence rather than recoverable aggregation failure.

F.2 Patch Expansion Dynamics Across Architectures

We next ask whether FBL cases show similar spatial dynamics across architectures. Figure 23 tracks the number of patches whose logit-lens prediction matches the object class across the final transformer blocks and post-norm. Across models, clean wins show broad expansion of object evidence across the patch grid, FBL cases stall at an intermediate level, and no-detection cases remain near baseline. This shared pattern suggests that FBL cases capture a similar failure mode across model capacity, training procedure, and pooling mechanism.

F.3 Near-Miss Structure and Fixed-Rule Recovery

Across models, the central FBL signature is preserved. FBL cases remain near-miss errors, with the correct class appearing in the final top-5 predictions in 94.7–98.7% of cases across models. For ViT-B/16, the correct class has a median rank of 2. We then apply the same re-ranking rule used in Table 8 to all models, using thresholds $a > 1.0$, $g < 1.0$, and $d > 0.5$ at the penultimate block, without architecture-specific retuning. This fixed rule allows recovery performance to be compared across architectures without model-specific tuning.

Table 9 shows that the fixed rule yields net-positive accuracy changes in all models, although the recovery-to-breakage tradeoff varies substantially. Net Δ is computed over all images and therefore includes incidental corrections of non-FBL errors, whereas the recovery-to-breakage ratio refers specifically to FBL cases recovered relative to correct predictions broken. The rule transfers best to models with stronger spatial compactness signals and larger FBL pools, and less well to ViT-S/16, where FBL cases are rare and no-detection errors dominate. Descriptively, models with

Table 9: **Cross-architecture robustness of aggregation-failure diagnostics.** For each model, we report accuracy on the stratified composite set, FBL rate, the fraction of FBL cases where the correct class appears in the final top-5 predictions, the compactness gap measured at the penultimate block, and recovery performance under the fixed re-ranking rule from Table 8. The same rule, $a > 1.0$, $g < 1.0$, and $d > 0.5$, is applied uniformly across all models without architecture-specific retuning. "Net Δ" is computed over all images and includes incidental corrections of non-FBL errors. "FBL ratio" and "FBL rec." are computed only with respect to FBL recovery.

Model	Pool	Acc	FBL	Top-5	Gap	Net Δ	FBL ratio	FBL rec.
ViT-B/16	token	57.2%	14.1%	98.7%	49%	+5.7pp	9.4:1	40.6%
ViT-S/16	token	46.3%	5.7%	94.7%	17%	+0.9pp	0.7:1	13.8%
ViT-L/16	token	58.7%	8.2%	96.2%	20%	+1.1pp	2.3:1	16.4%
DeiT-B/16	token	52.5%	8.3%	96.9%	22%	+1.9pp	2.2:1	23.9%
ViT-RelPos-B/16	avg	51.9%	14.9%	95.1%	23%	+1.9pp	1.9:1	22.2%

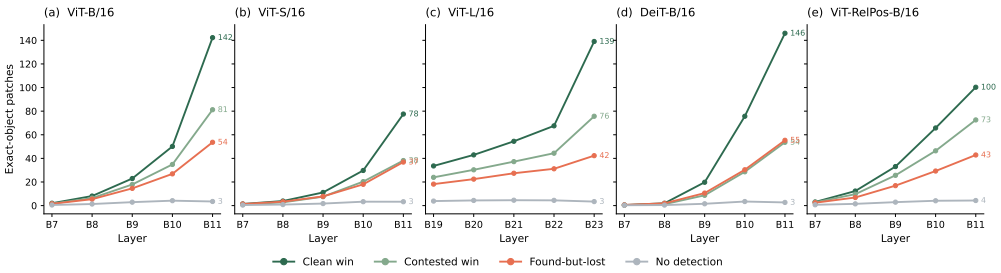


Figure 23: **Object patch expansion across ViT variants.** Each panel shows the mean number of patches, out of 196, whose argmax logit-lens prediction matches the object class, plotted across the final transformer blocks and post-norm, split by outcome category. Across architectures, clean wins show broad expansion, FBL cases stall at an intermediate level, and no-detection cases remain near baseline. This shared pattern supports the interpretation that FBL cases reflect incomplete propagation of object-aligned evidence rather than a failure unique to ViT-B/16.

larger compactness gaps broadly show larger recoverable fractions, but we treat this association as descriptive rather than inferential because the comparison includes only five architectures.

Together, these results show that FBL errors are near-miss errors with consistent spatial dynamics across transformer variants. The ViT-RelPos result further indicates that the FBL signature is not specific to CLS-token pooling. Even when final predictions are obtained by averaging patch tokens, object-aligned evidence can remain spatially localized and be outweighed by competing background-aligned evidence. The strength of recovery varies across models and is larger when the spatial compactness signal is stronger, suggesting that recoverability depends on how clearly object evidence remains localized before final aggregation.

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: The abstract and introduction summarize the main contributions: ImageNet-OOC1k, the consensus visual-semantic axis, the accuracy–distance relationship across 35 models, and the aggregation-failure analysis in Vision Transformers. The scope matches the evidence, with broad behavioral results in Section 5 and Appendix C, and mechanistic claims restricted to transformer analyses in Section 6 and Appendices D–F.

   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: Section 7 discusses limitations related to inpainting and compositing artifacts, the model-derived semantic axis, the focus of mechanistic analyses on Vision Transformers, and the diagnostic nature of the re-ranking analysis rather than a deployable inference-time method.

   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 is an empirical benchmark and mechanistic analysis paper and does not present theoretical results, theorems, or formal proofs.

   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: The paper provides the information needed to reproduce the main results, including dataset construction, semantic axis computation, composite generation, model set, evaluation metrics, sampling strategies, and mechanistic diagnostics in Sections 3–6 and Appendices A–F. Release plans and source-data constraints are described in Appendix A.4.

   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: Code, metadata, semantic-distance matrices, composite specifications, and generation parameters are available at `https://github.com/ToastyDom/ImageNet-OOC1k`, together with reproduction instructions. Access to image-derived components follows the terms of ImageNet/ImageNet-S.

   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 paper does not train new models, but specifies the experimental details needed to understand the results, including dataset construction, composite generation, semantic distance computation, model set, evaluation metrics, sampling strategies, and diagnostic thresholds in Sections 3–6 and Appendices A–F.

   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: The paper reports statistical tests for the main experimental claims. Appendix C.2 reports image-level logistic regressions with 95% confidence intervals, Appendix C.1 reports a permutation test for background-only semantic-neighbor enrichment, and Appendix A.2 reports fixed-effects regressions with standard errors clustered by background and object instance. Not every figure includes error bars.

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: Appendix A.5 reports the hardware, software stack, storage requirements, and compute used for dataset generation, behavioral inference, and mechanistic analyses. It also reports total compute for the presented experiments and additional compute for preliminary inference-only experiments.

   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: The research conforms to the NeurIPS Code of Ethics. The work does not involve human subjects, crowdsourcing, sensitive personal data, or high-risk data/model release. The planned ImageNet-OOC1k release will respect the access and licensing terms of the underlying ImageNet/ImageNet-S data.

   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: [Yes]

    Justification: Appendix A.4 discusses intended use, limitations, release constraints, and ethical considerations. The benchmark may support more reliable evaluation of vision models, while negative impacts are indirect and depend on downstream deployment contexts, including sensitive applications of visual recognition.

    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: The paper does not release high-risk models or newly scraped datasets. The planned dataset release is for robustness evaluation and will follow the access and licensing terms of the underlying ImageNet/ImageNet-S data.

    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 the datasets, pretrained models, and tools used in this work. ImageNet-OOC1k derives from ImageNet/ImageNet-S and will be released in accordance with their access terms, licensing conditions, and redistribution constraints, as described in Appendix A.4.

    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: [Yes]

    Justification: We introduce ImageNet-OOC1k and document its construction, metadata, generation procedure, intended use, limitations, and release conditions in Section 3 and Appendix A. The dataset and accompanying documentation are available subject to the access and redistribution terms of ImageNet/ImageNet-S.

    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: The paper does not involve crowdsourcing, direct interaction with human participants, or human-subject experiments.

    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: The paper does not involve crowdsourcing, direct interaction with human participants, or human-subject experiments, so IRB approval or equivalent review is not applicable.

    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: [N/A]

    Justification: The core methodology and experiments do not involve LLMs as components.

    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.