Extended Abstract Track

The Binding Problem in Vision Models: Geometric, Functional, and Behavioral Approaches

Editors: List of editors' names

Abstract

Existing studies of neural networks have focused largely on *compositionality*—whether individual features can be linearly decoded and reused—while overlooking the equally important issue of *binding*, i.e., how features are linked together to form coherent objects. This leaves a gap in understanding whether models truly represent feature conjunctions rather than mere unstructured feature bags. We propose a geometric and functional framework for quantifying binding, introducing a binding score based on principal angles between concept subspaces and validating it with linear or non-linear probes. To complement this, we design a behavioral diagnostic dataset in which pairs of images share identical feature bags but differ in how those features are bound into objects. Together, these frameworks highlight binding as a distinct and measurable dimension of representation, providing tools to diagnose where current vision models succeed—and where they fail—in capturing object structure.

Keywords: binding, compositionality, linear representation

1. Introduction

Any object in a neural system—biological or artificial—is represented by a collection of features distributed across neurons, tokens, or layers. What makes these features an object is not just their presence but their *binding*: the fact that they belong together as a coherent whole. In fact, failure to bind features to corresponding objects is a common failure mode of vision models (see Appendix A for qualitative examples) (Campbell et al., 2025; Lewis et al., 2024; Zhang et al., 2024; Yuksekgonul et al., 2023; Assouel et al., 2025).

However, the need to encode both feature compositionality *and* their binding gives rise to a fundamental trade-off for vision models. On one side, a network could encode every possible conjunction explicitly, akin to a one-hot encoding, which ensures perfect binding but introduces massive redundancy and discards compositional reuse. On the other, it could represent features in isolation, ignoring object structure and allowing spurious cross-object interactions. In practice, models such as Vision Transformers (ViTs) must navigate between these extremes, allocating limited capacity to directions that support binding *and* compositionality. Our aim is to provide a mathematical framework and concrete tools to measure this allocation.

Existing interpretability work largely focuses on whether individual concepts can be linearly decoded, and in some cases on whether the corresponding directions are orthogonal. However, orthogonality between concept directions does not capture *binding*, but only measures the correlation of features often caused by dataset imbalance (Uselis et al., 2025). On the other hand, binding concerns whether conjunctions of features require *new* directions beyond the linear combinations of the individual concepts.

In this paper, we develop a framework to directly measure binding in neural embeddings. Our approach compares the subspace of a joint probe (predicting feature conjunctions) to the union of single-feature subspaces: directions required by the joint but orthogonal to the union serve as evidence of binding. We formalize this geometrically with principal angles and introduce a binding score. We complement this with a functional test based on the relative performance of a joint probe and a union probe, which allows non-linear decoding.

To evaluate binding behaviorally, we further design a diagnostic dataset in which image pairs share identical *bags* of feature but differ in how features are bound together into objects. Success requires encoding conjunctions rather than unstructured feature bags. Our contributions are as follows:

1. A geometric and functional framework for quantifying binding in neural embeddings.
2. A novel behavioral dataset that evaluates binding.

2. Measuring Binding

We study how information about multiple concepts is represented in neural embeddings. Our key idea is to test whether a joint probe (trained to decode concept pairs or tuples) relies on directions in representation space that are *not* captured by the linear span of the single-concept probes. If so, we interpret these extra directions as evidence of *binding*.

It is important to distinguish binding from mere correlations between concepts. Correlations typically arise from imbalances in the dataset (e.g., "red is more often a square than a circle"), and are reflected as directions lying within the span of the single-concept subspaces (Uselis et al., 2025). In contrast, a true binding direction is orthogonal to the single-concept subspaces: it cannot be reconstructed from any linear combination of single-concept features. At the extreme, imagine a one-hot representation for each possible (color, shape) pair that discards all compositional structure. Each such vector would be entirely orthogonal to the color and shape subspaces, reflecting pure binding without composition.

To formalize this intuition, we treat each probe as defining a subspace of the representation space: single-concept probes span subspaces $\mathcal{H}_{C_i}$, while a joint probe spans $\mathcal{H}_S$ for a set of concepts S. Binding then corresponds to the extent to which $\mathcal{H}_S$ requires directions *outside* the union of the single-concept subspaces. In the following sections, we develop the geometry and metrics needed to quantify this relationship. A detailed derivation is left in Appendix B, with Figure 3 illustrating our framework.

Concept and joint subspaces. For each concept C_i with k_i classes, we train a linear probe with weight matrix

$$W_{C_i} \in \mathbb{R}^{k_i \times d},$$

and define the concept subspace

$$\mathcal{H}_{C_i} = \text{rowspan}(W_{C_i}) \subseteq \mathbb{R}^d.$$

For example, if C_i corresponds to *color* with classes {red, blue, green}, the probe weights W_{color} span a subspace $\mathcal{H}_{\text{color}}$ of $\mathbb{R}^d$. Similarly, if C_j corresponds to *shape* with classes {circle, square, triangle}, we obtain a subspace $\mathcal{H}_{\text{shape}}$.

For a set of concepts $S \subseteq \{C_1, \ldots, C_n\}$, the joint probe predicting their Cartesian product has weights

$$W_S \in \mathbb{R}^{K_S \times d}, \qquad K_S = \prod_{C_i \in S} k_i,$$

with joint subspace $\mathcal{H}_S = \text{rowspan}(W_S)$. Continuing the example, the joint probe for (color, shape) has $K_S = 3 \times 3 = 9$ classes such as "red-square" or "blue-circle."

Union versus joint. The union subspace of the single-concept probes in S is

$$\mathcal{H}_{\cup S} = \sum_{C_i \in S} \mathcal{H}_{C_i}.$$

If $\mathcal{H}_S \subseteq \mathcal{H}_{\cup S}$, then the joint probe uses no information beyond the individual concepts. If $\mathcal{H}_S$ requires new directions outside $\mathcal{H}_{\cup S}$, these extra components are candidates for binding.

Binding score. Let $U_{\cup S}$ and V_S be orthonormal bases for $\mathcal{H}_{\cup S}$ and $\mathcal{H}_S$, respectively, and set

$$M_S = U_{\cup S}^\top V_S.$$

The singular values σ_i of M_S are related to the *principal angles* ϕ_i between the two subspaces by

$$\sigma_i = \cos \phi_i, \qquad i = 1, \ldots, q, \quad q = \min(r_{\cup S}, r_S).$$

We define the binding score as

$$B_S = 1 - \frac{1}{r_S} \sum_{i=1}^{q} \sigma_i^2 = 1 - \frac{1}{r_S} \sum_{i=1}^{q} \cos^2 \phi_i,$$ [1]

which measures the fraction of the joint subspace unexplained by the union. A value of $B_S = 0$ indicates no binding, while larger values indicate stronger binding.

Dimension mismatch. When $\dim(\mathcal{H}_S) > \dim(\mathcal{H}_{\cup S})$, some nonzero binding score is inevitable. To ensure that binding is not trivially inflated by excess rank, we repeat the analysis after truncating the joint probe to match the dimensionality of the union and report the discarded variance. Appendix D contains preliminary results on the binding score and associated tail energies.

Functional validation. Geometry alone can be misleading. We therefore complement B_S with a functional test: train one joint probe on the raw representations,

$$g_S^{\text{raw}} : h \mapsto \hat{y}_S,$$

and another on the union features,

$$g_S^{\cup} : U_{\cup S}^\top h \mapsto \hat{y}_S.$$

The difference in classification accuracy,

$$\Delta\text{Acc} = \text{Acc}(g_S^{\text{raw}}) - \text{Acc}(g_S^{\cup}),$$

1. Equivalently $B_S = 1 - \frac{1}{r_S}\|M_S\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm. This avoids the need to find the singular values of M_S using SVD.

indicates whether the raw representations contain additional predictive information beyond the union. We interpret binding as present only when both $B_S > 0$ after rank-matching and $\Delta\text{Acc} > 0$.

We note that our geometric analysis is conducted under the assumption that features are represented linearly, an assumption that is common but not entirely realistic. In contrast, the functional probes here can be implemented as **non-linear** classifiers, allowing them to capture binding structure that may only be accessible through non-linear decision boundaries. This flexibility provides a practical complement to the linear subspace analysis.

Summary. Our framework provides a simple operational test: if a joint probe uses directions outside the span of single probes *and* achieves higher accuracy than a union probe, we conclude that the representation contains dedicated features for binding concepts together.

3. A Behavioral Dataset for Binding

A central challenge in vision models is determining whether models encode *bindings*—the conjunction of features into coherent objects—or whether they merely maintain unstructured bags of features. To test this, we design a behavioral diagnostic dataset tailored for binding, inspired by the classic "superposition catastrophe" in the binding problem (Greff et al., 2020).

Dataset construction. Each datapoint in our dataset is a *pair* of images. Every image contains two non-overlapping objects, each defined by a color and a shape. Crucially, the two images in a pair always contain the *same bag of features*: the set of colors and the set of shapes across the two images are identical (Figure 4). What differs is the *binding* of these features. In pairs labeled "same", the two images contain the same bound objects (e.g., a red circle and a blue square in both images). In pairs labeled "different", the bindings are swapped across images while preserving the feature marginals (e.g., red circle + blue square in one image versus red square + blue circle in the other). Solving this task requires the model to represent the conjunction of features into objects: a representation that discards binding and stores only feature bags cannot succeed above chance.

We further introduce several baseline variants of the dataset—such as fixing locations, changing only one feature, or using entirely different objects—to provide simplified controls where binding is less critical (see Appendix for details).

Probing procedure. To quantify whether binding information is present in intermediate representations of a Vision Transformer (ViT), we extract the `[CLS]` activations at a given layer for each image in a pair. We then train a lightweight "CLIP-style" probe: each image is projected through a learned linear mapping into a shared latent space, and the similarity of the two images in a pair is computed via their dot product (Radford et al., 2021). The dot product is interpreted as a *logit*, which is passed through a sigmoid to yield a probability. A binary cross-entropy loss is then applied. Formally, letting $h_A, h_B \in \mathbb{R}^d$ denote the `[CLS]` activations of the two images in a pair, and $W \in \mathbb{R}^{d\times k}$ the learned projection, the probe computes

$$s = \langle W^\top h_A,\ W^\top h_B \rangle, \qquad p = \sigma(s), \qquad \ell = -\big(y \log p + (1-y)\log(1-p)\big),$$

where $y \in \{0, 1\}$ denotes the label ("different" vs. "same"). High probe accuracy at a given layer indicates that binding information is encoded in the corresponding representation. We include preliminary results in Appendix 7.

References

Rim Assouel, Declan Campbell, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models, 2025. URL https://arxiv.org/abs/2506.15871.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor W. Webb. Understanding the limits of vision language models through the lens of the binding problem, 2025. URL https://arxiv.org/abs/2411.00238.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks, 2020. URL https://arxiv.org/abs/2012.05208.

Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models, 2024. URL https://arxiv.org/abs/2212.10537.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Arnas Uselis, Andrea Dittadi, and Seong Joon Oh. Does data scaling lead to visual compositional generalization?, 2025. URL https://arxiv.org/abs/2507.07102.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. URL https://arxiv.org/abs/2210.01936.

Yizhe Zhang, He Bai, Ruixiang Zhang, Jiatao Gu, Shuangfei Zhai, Josh Susskind, and Navdeep Jaitly. How far are we from intelligent visual deductive reasoning?, 2024. URL https://arxiv.org/abs/2403.04732.

Appendix A. Examples of Binding Failure in Vision Models

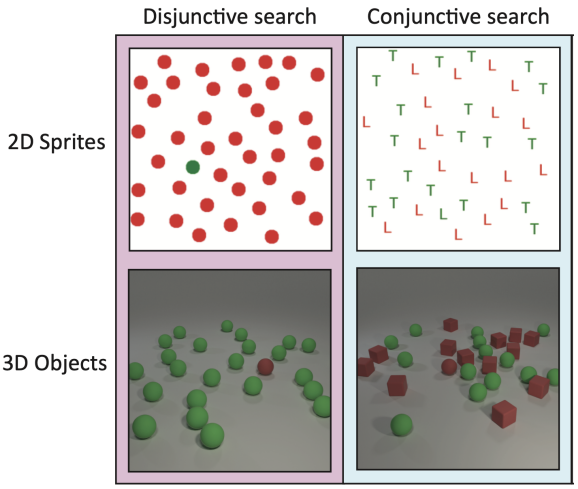


Figure 1: Campbell et al. (2025) demonstrate that Vision-Language Models (VLMs) struggle with conjunctive search tasks, particularly as the total number of objects increases.

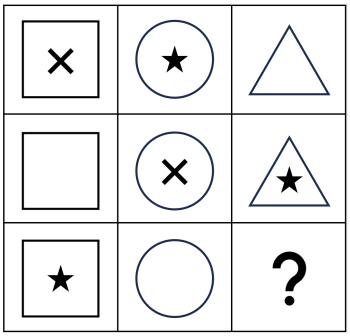


Figure 2: Zhang et al. (2024) prompt VLMs to describe grid patterns. For the middle-right block, the model outputs "a triangle with an X inside," apparently combining elements from adjacent blocks (middle-center and middle-right). They further show that segmenting the grid into separate images before passing them in significantly reduces such errors.

Appendix B. Mathematical Formulation of Binding

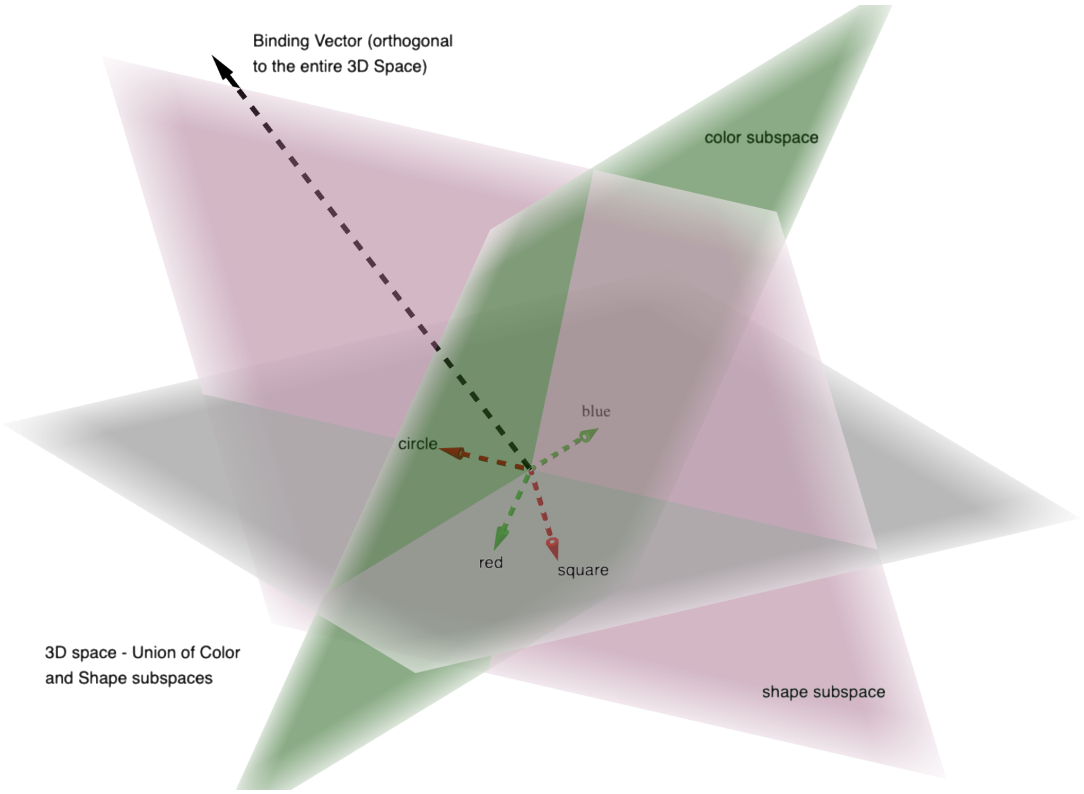


Figure 3: Illustration of the subspaces and binding when concepts are shape and color, each of which contains two classes.

B.1. Setup and Notation

We work in a d-dimensional representation space $\mathbb{R}^d$. Let the set of concepts be denoted by $\{C_1, \ldots, C_n\}$, where each concept C_i has k_i possible classes. For each concept C_i, we train a multi-class linear probe with weight matrix

$$W_{C_i} \in \mathbb{R}^{k_i \times d}.$$

The probe defines a linear subspace associated with the concept,

$$\mathcal{H}_{C_i} := \text{rowspan}(W_{C_i}) \subseteq \mathbb{R}^d, \qquad r_{C_i} := \text{rank}(W_{C_i}),$$

which we refer to as the *concept subspace*.

More generally, for any subset of concepts $S \subseteq \{C_1, \ldots, C_n\}$, we consider a joint probe that predicts the Cartesian product of their classes. The corresponding probe weight matrix

is

$$W_S \in \mathbb{R}^{K_S \times d}, \qquad K_S = \prod_{C_i \in S} k_i,$$

and the associated subspace is

$$\mathcal{H}_S := \text{rowspan}(W_S), \qquad r_S := \text{rank}(W_S).$$

The row span of W_S thus captures all directions in representation space that are linearly exploited by the probe to decode the joint labels across the concepts in S.

B.2. Orthonormal Bases

For each concept subspace $\mathcal{H}_{C_i}$ defined by the row span of the probe weights W_{C_i}, we construct an orthonormal basis in order to work with a numerically stable and basis-invariant representation of the subspace.

Given a probe weight matrix

$$W_{C_i} \in \mathbb{R}^{k_i \times d}, \qquad r_{C_i} = \text{rank}(W_{C_i}),$$

we compute an orthonormal basis

$$U_{C_i} \in \mathbb{R}^{d \times r_{C_i}}$$

such that the columns of U_{C_i} span $\mathcal{H}_{C_i}$.

This can be achieved in two equivalent ways. One option is to apply an economy QR decomposition to the transpose:

$$W_{C_i}^\top = Q_i R_i, \qquad U_{C_i} = Q_i(:, 1 : r_{C_i}),$$

where the first r_{C_i} columns of Q_i form an orthonormal basis. Alternatively, we may compute a singular value decomposition

$$W_{C_i} = A_i \Sigma_i V_i^\top, \qquad U_{C_i} = V_i(:, 1 : r_{C_i}),$$

where the first r_{C_i} right singular vectors of W_{C_i} define the basis.

The same procedure is applied to the joint probe W_S, yielding an orthonormal basis

$$V_S \in \mathbb{R}^{d \times r_S}$$

whose columns span the joint subspace $\mathcal{H}_S$. In all subsequent analysis, we work with these orthonormalized representations of the subspaces rather than the raw weight matrices, ensuring that comparisons between subspaces are invariant to arbitrary reparameterizations of the probes.

B.3. Union Subspace

Given a collection of concepts $S \subseteq \{C_1, \ldots, C_n\}$, we define the *union subspace* as the linear span of their individual concept subspaces,

$$\mathcal{H}_{\cup S} := \sum_{C_i \in S} \mathcal{H}_{C_i} = \text{span}\{\, U_{C_i} : C_i \in S \,\}.$$

To construct an orthonormal basis for $\mathcal{H}_{\cup S}$, we first concatenate the individual orthonormal bases,

$$X_S := \left[U_{C_{i_1}} \;\; U_{C_{i_2}} \;\; \cdots \;\; U_{C_{i_{|S|}}} \right],$$

and then apply a QR decomposition or singular value decomposition,

$$X_S = Q_{\cup S} R_{\cup S}.$$

The orthonormal basis of the union subspace is taken as

$$U_{\cup S} := Q_{\cup S}(:, 1 : r_{\cup S}),$$

where

$$r_{\cup S} = \dim(\mathcal{H}_{\cup S}).$$

Thus $U_{\cup S}$ provides a numerically stable representation of the union subspace that will be used for all subsequent comparisons with joint probes.

B.4. Principal Angles and Binding Score

To compare the information captured by the union subspace $\mathcal{H}_{\cup S}$ with the joint subspace $\mathcal{H}_S$, we compute the principal angles between them. Let

$$U_{\cup S} \in \mathbb{R}^{d \times r_{\cup S}}, \qquad V_S \in \mathbb{R}^{d \times r_S}$$

denote the orthonormal bases spanning $\mathcal{H}_{\cup S}$ and $\mathcal{H}_S$, respectively. The overlap matrix is defined as

$$M_S := U_{\cup S}^\top V_S \in \mathbb{R}^{r_{\cup S} \times r_S}.$$

We compute the thin singular value decomposition

$$M_S = Q \Sigma R^\top,$$

where

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_q), \qquad q = \min(r_{\cup S}, r_S).$$

The singular values satisfy

$$\sigma_i = \cos \phi_i,$$

with ϕ_i the i-th principal angle between $\mathcal{H}_{\cup S}$ and $\mathcal{H}_S$.

To summarize the overall alignment, we define the binding score

$$B_S = 1 - \frac{1}{r_S} \sum_{i=1}^{q} \sigma_i^2.$$[2]

This score measures the average fraction of the joint probe subspace that lies outside the union subspace. A value of $B_S = 0$ indicates that the joint subspace is entirely contained in the union, and thus no additional binding information is present. A value of $B_S > 0$ indicates that some directions of the joint subspace are not explained by the union, reflecting evidence of binding. At the extreme, $B_S = 1$ corresponds to the case where the joint subspace is orthogonal to the union.

2. Equivalently, $B_S = 1 - \frac{1}{r_S} \|M_S\|_F^2$, since the squared singular values of M_S sum to the squared Frobenius norm. This avoids the need for SVD on M_S.

Dimension mismatch. When $r_S > r_{\cup S}$, at least $r_S - r_{\cup S}$ directions of the joint subspace necessarily fall outside the union. In this case, interpreting B_S requires caution, as some nonzero binding score arises purely from dimensionality mismatch rather than genuine binding structure.

B.5. Rank-Matching

A key consideration in interpreting the binding score is the potential effect of dimension mismatch between the union subspace $\mathcal{H}_{\cup S}$ and the joint subspace $\mathcal{H}_S$. In particular, when $r_S > r_{\cup S}$, it is inevitable that at least $r_S - r_{\cup S}$ directions of $\mathcal{H}_S$ lie outside the union. If left uncorrected, this effect can artificially inflate the binding score B_S, even when the apparent "extra" directions carry little or no meaningful information. To mitigate this problem, we adopt a strict rank-matching procedure.

Given the joint probe weight matrix $W_S \in \mathbb{R}^{K_S \times d}$, we compute its singular value decomposition

$$W_S^\top = \tilde{U}\Sigma\tilde{V}^\top,$$

where the diagonal entries of Σ contain the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{r_S} > 0$. To match the dimensionality of the union subspace, we retain only the top $r_{\cup S}$ right singular vectors, weighted by their singular values. Formally, we define

$$V_S^{(\text{cap})} := \tilde{V}(:, 1 : r_{\cup S}),$$

which spans the truncated joint subspace of dimension $r_{\cup S}$.

The amount of variance discarded by this truncation is quantified by the tail energy,

$$E_{\text{tail}} = \frac{\|\Sigma_{>(r_{\cup S})}\|_F^2}{\|\Sigma\|_F^2},$$

where $\Sigma_{>(r_{\cup S})}$ denotes the diagonal submatrix of singular values beyond the first $r_{\cup S}$. A small value of E_{tail} indicates that most of the energy of the joint probe is preserved within the top $r_{\cup S}$ directions, while a large value suggests that the truncated directions may be carrying nontrivial signal. After constructing $V_S^{(\text{cap})}$, we recompute the principal angles and the binding score using this rank-matched basis in place of V_S.

This rank-matching step ensures that binding scores are not trivially inflated by dimensionality mismatch, and that any evidence of binding reflects additional structure rather than excess capacity of the joint probe.

B.6. Functional Evidence for Binding

While geometric measures provide a principled way to quantify the overlap between union and joint subspaces, geometry alone can be misleading. In particular, apparent binding may arise due to correlations in the training data or as an artifact of mismatched probe capacity. To address this issue, we complement the geometric analysis with functional tests that directly evaluate probe performance.

We first train a joint probe on the raw representations,

$$g_S^{\text{raw}} : h \mapsto \hat{y}_S,$$

where $h \in \mathbb{R}^d$ denotes the hidden representation and $\hat{y}_S$ the predicted joint label across the concepts in S. We then train a second probe that only has access to the union subspace features,

$$g_S^{\cup} : U_{\cup S}^{\top} h \mapsto \hat{y}_S,$$

where $U_{\cup S}$ is the orthonormal basis of the union subspace constructed in Section 3. Both probes are trained with identical architectures and regularization so that their performance can be fairly compared.

Let $\mathrm{Acc}(\cdot)$ denote the classification accuracy of a probe. We define the accuracy gap as

$$\Delta\mathrm{Acc} = \mathrm{Acc}(g_S^{\mathrm{raw}}) - \mathrm{Acc}(g_S^{\cup}).$$

When $\Delta\mathrm{Acc} \approx 0$, the union subspace suffices to recover the joint labels, suggesting little evidence for binding. In contrast, when $\Delta\mathrm{Acc} > 0$, the raw representations contain additional predictive information that is not captured by the union subspace, pointing to the presence of binding features.

In practice, best results are obtained by carefully controlling the probe capacity and ensuring a balanced dataset over the Cartesian product of classes. This avoids spurious improvements due to probe expressivity or label correlations. We adopt a conservative interpretation: binding is taken to be present only when both the geometric score B_S is strictly positive after rank-matching, and the functional accuracy gap $\Delta\mathrm{Acc}$ is positive.

Appendix C. Dataset Baseline Variations

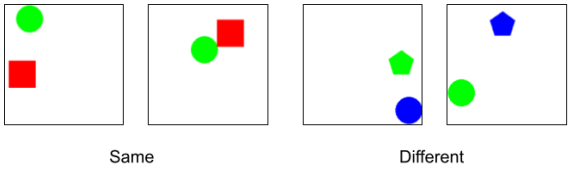


Figure 4: Sample datapoints from the original dataset.

We introduce several controlled baseline variations of our dataset to contextualize the binding task. In one variation, object locations are fixed within each pair, and "same" pairs are created by optionally swapping the two positions between images, which increases the potential for positional confusion (Figure 5 Red and Green). In another variation, "different" pairs are generated such that one object remains the same across both images while the other object changes only one feature (e.g., color but not shape), providing a simplified comparison where binding may not be strictly required (Figure 5 Yellow). Finally, in the most simplified variant, the "different" pair is constructed so that the two objects in one image are completely different from those in the other image of the pair, a case where binding is unlikely to be necessary for successful discrimination (Figure 5 Blue).

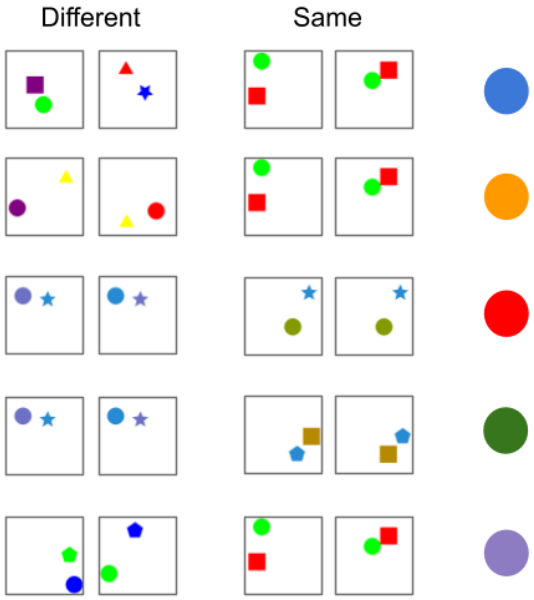


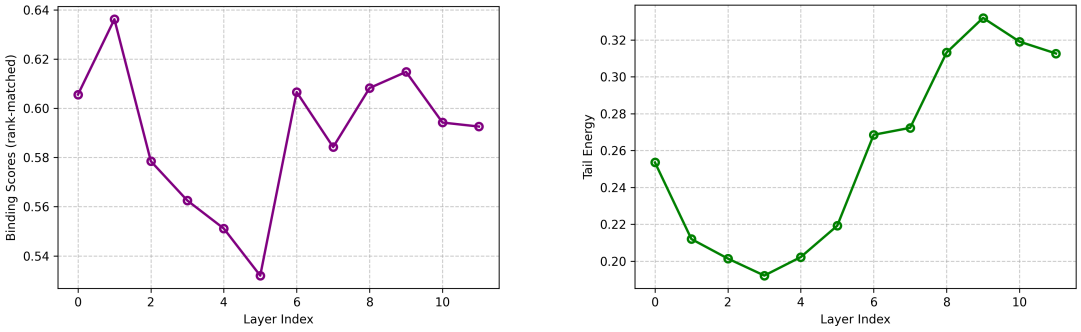
Figure 5: Samples from baseline datasets (last row is the original).

Appendix D. Preliminary Results on Binding Metrics

We perform preliminary experiments to compute the binding metrics. The plots in Figure 6 reveal a non-monotonic trajectory of binding across the layers of a DINO-pretrained Vision Transformer. Binding scores are high in the earliest layers, dip steadily through the middle layers, and then rise again at deeper layers, suggesting that early representations contain entangled low-level interactions, mid-layer representations are more compositional, and later layers reintroduce binding for higher-level integrated features. Tail energy shows a complementary trend: it is lowest in the middle layers and increases steadily toward deeper layers, indicating that the joint probe allocates more variance to directions outside the union as depth grows. However, we emphasize that our experimental results remain preliminary, and a systematic evaluation across different architectures, synthetic and natural datasets and their variations, as well as the aforementioned linear and non-linear probing experiments are needed to establish the robustness and generality of our findings.

Appendix E. Preliminary Results on the Binding Dataset

We perform probing on various dataset aforementioned and find that performance is generally high for most layers of a DINO-pretrained vision transformer. The location-swapping Green dataset (Figure 5) is, however, particularly difficult for layers 1 and 2. We plan on extending our dataset to **natural images with more complex features** (texture, semantic, etc.) and **more objects** to provide a more realistic testbed for binding in vision models.



(a) (b)

Figure 6: Binding scores and tail energies across layers of a DINO-pretrained Vision Transformer.

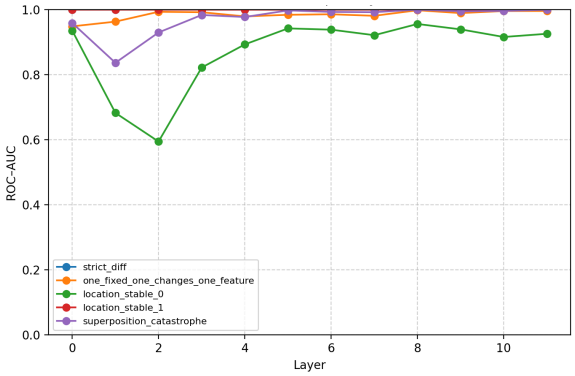


Figure 7: Testing ROC-AUC for each dataset. Curve colors correspond to the color coding of datasets in Figure 5.