LIMITATIONS OF RELATIONAL ENCODING IN VISUAL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Transformers (ViTs) provide dense, patch-level embeddings that support semantic segmentation, depth estimation, and correspondence. Richer visual understanding, however, requires representations to encode physical and spatial relations between entities. We analyze how much of this relational information can be extracted from frozen embeddings of recent foundation ViT models, pre-trained on static images only (DINOv2, DINOv3), on image-text pairs (`dino.txt`, PE-Core, PE-Spatial), and natively on videos (LeVJEPA). We use single-entity and pairwise probes, validating them on synthetic tasks and applying them to human-object interaction recognition in HICO-DET. To assess whether performance depends on proxy shortcuts, such as pose detection or spatial proximity, we evaluate out-of-distribution (OOD) generalization, position-related shortcuts, and counterfactual inpainting. We find a split between the two probe families. Probes on single entities can detect whether an interaction occurs, but fail to discern which entities actively participate: their accuracy is comparable to that of the best probe trained on positional encoding only. Pairwise probes retain an accuracy gain of 4–10 points over this positional baseline, and the OOD study shows a degree of generalization across objects. Increasing probe complexity through non-linearity does not yield consistent improvements, with MLP probes performing comparably on average to their linear or bilinear counterparts. Moreover, we find that models exposed to videos or text during training do not perform better on static images. These results suggest that the successes of ViTs in dense semantic extraction from local patches do not extend to the understanding of physical interactions in static images.

1 INTRODUCTION

Dense representations produced by Vision Transformers (ViTs) (Dosovitskiy et al., 2021) exhibit remarkable semantic structure (Amir et al., 2022). A new generation of self-supervised vision models encodes rich semantic information in local patch embeddings (Siméoni et al., 2025; Kuhn et al., 2026). This information contains extractable knowledge of object-centric properties, both visual (shape, color, texture), and semantic (class, part-whole membership). On downstream tasks such as segmentation (even in specialized fields like medical imaging (Yang et al., 2025)), visual correspondence and retrieval (Caron et al., 2021) and depth-estimation (El Banani et al., 2024) these models match or surpass the state-of-the-art. Mapping individual entities to local embeddings — a property generally found more in dense visual encoders than in masked autoencoders or standard CLIP models (Song et al., 2024; Lüddecke & Ecker, 2025; Siméoni et al., 2025) — is only a necessary condition for transforming images into actionable representations. A deep understanding of a scene also requires grasping the interactions between objects and their relations within the scene as a whole. This is essential for effective spatial and causal reasoning (Liu et al., 2023). Recent work has studied whether these representations encode relational cues directly within their internal embedding spaces, revealing that, in limited scopes, self-supervised transformers may implicitly learn forms of compositional scene understanding (Li et al., 2025; Lepori et al., 2024; Zhan et al., 2024). It remains an open question to what extent physical relations between objects are truly encoded by these visual models, especially considering that most models are trained solely on static images, where causal, physically driven interactions are obviously absent. We structure the experiments of our work around two central questions:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

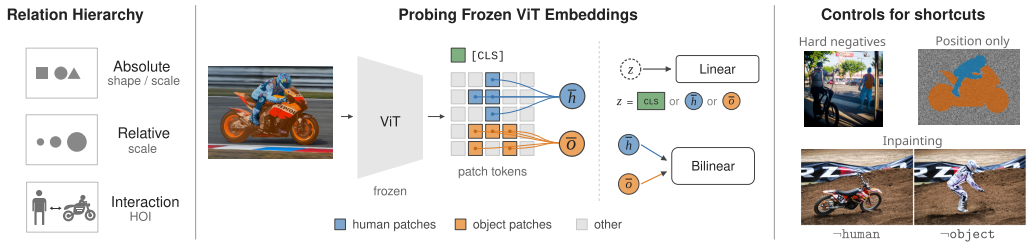


Figure 1: Overview. **(Left)** A scheme of the relation hierarchy, from same-entity features to relative comparisons to physical interactions. **(Center)** Images are processed into embeddings, from which we obtain the average representation of interacting entities, used as inputs for linear and bilinear probes. **(Right)** Examples of hard negatives and counterfactuals, including interaction-entity masks over noise and inpainted images.

i) Does relational understanding emerge from the self-supervised training of models on static images, and is it linearly accessible in the patch embeddings?

ii) Do alignment objectives beyond those of static images, such as the textual cues of `dino.txt` (Jose et al., 2025) or the PE family (Bolya et al., 2026) and the temporal causality in LeVJEPA (Kuhn et al., 2026), modify the internal representations of static images significantly with regard to physical interactions?

To answer these questions, we probe frozen representations across both synthetic geometric tasks and human-object interaction (HOI) detection in natural images. We deploy linear and bilinear probes, applied to single patches or pairs of patches (or averages thereof) respectively, to see whether the presence of certain interactions can be accurately extracted from the embeddings. We take care to avoid spurious results, such as pose matching, spatial proxies, and co-occurrence statistics, by building controlled datasets, studying out-of-distribution behaviors, and using counterfactual tasks emerging from inpainting (Bhattad et al., 2026).

We demonstrate that, for all models, probes can achieve high accuracy in detecting whether an image contains a specific interaction. However, the picture is more complex when the task is to discern which specific entities actively participate and which are bystanders. On the one hand, the single-patch linear probes reach accuracies comparable to the best probe trained on positional information only. On the other hand, bilinear probes retain a 4–10-point gain over this baseline and show some generalization to OOD objects. Replacing a bilinear with a one hidden-layer MLP over the concatenation of the patch embeddings does not further improve performance. Furthermore, we find no evidence of improved performance either when the model is natively trained on video data or when it is aligned with textual cues.

The core contributions of our paper are:

- We construct 2D and 3D synthetic datasets and augment the HICO-DET benchmark (Chao et al., 2018) with dense segmentation masks to better isolate the entities involved in the HOIs.
- We systematically test whether linear and bilinear probes can decode relations of varying complexity across models trained on images, videos, and text.
- We use counterfactual (inpainting) and out-of-distribution tests to distinguish relational encoding from possible task proxies.

2 Related Work

Li et al. (2025) investigate whether object binding ability, the property of grouping together objects' low-level features (e.g., color, shape, texture, etc.), emerges naturally in large pretrained ViTs. In particular, they focus on the *IsSameObject* property, which they define as the ability of ViTs to encode whether two patches belong to the same object in an image. Using a full-rank bilinear

probe, they find that this property is reliably decodable across models such as DINOv2, CLIP, and ImageNet-supervised ViTs, but absent in models trained with MAE. This suggests that the property emerges through training due to the specific optimization objectives of each SSL method, rather than being an architectural artifact. Lepori et al. (2024) employ mechanistic interpretability methods to reveal that pretrained ViTs, particularly CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024), when fine-tuned for solving same-different tasks, develop a two-stage processing pipeline: an early perceptual stage extracting disentangled object representations, followed by a relational stage performing abstract comparison operations.

In the field of spatial relation, Wen et al. (2024) propose RelatiViT, a purpose-built ViT-based architecture for spatial relation prediction tasks, which surpasses the performance of many Vision-Languange Models (VLMs). Unlike purely self-supervised visual models, which are trained only on visual signals, VLMs align the visual representation with natural language through image-text supervision. This alignment can be the primary pretraining objective, as in CLIP (Radford et al., 2021), or be added on top of a visual encoder pretrained through other methods, as in `dino.txt` (Jose et al., 2025). Yet despite the richer semantic grounding provided by text alignment, VLMs fall short of human performance on surprisingly simple spatial tasks, as shown by the spatial reasoning benchmarks introduced by Kamath et al. (2023); Kargin et al. (2026).

3 METHODS

Relation hierarchy. To study relational information, we construct a hierarchy of properties that require different levels of contextual information. At the simplest level, we have context-free, single-entity attributes, such as an object's shape or size. Here, the relation is from the individual patches to the objects as a whole to which they belong. Note that, even at this level, the associations are not necessarily trivial, since self-supervised models can distinguish between "visual anagrams" (Doshi et al., 2025). At a higher level, we have relative relations, which require comparisons of absolute features across multiple entities in the same image. We probe whether individual patches encode that their object is larger than another, that it is the largest object in the scene, or that it is the odd one out in a set of otherwise identical entities. These properties are not well defined in isolation and require the local representations to encode some information about the global context. At the highest level of abstraction, we consider physical interactions in natural images. These relations involve semantic, highly asymmetric participation between different entities, such as a person riding a bicycle or holding an object.

Following this hierarchy allows us to i) verify that our methodology as a whole is effective and ii) study which level of relational structure, if any, is accessible from the frozen ViT representations.

Models. We study foundation ViT encoders with diverse training objectives, all known to produce dense feature maps with strong local semantics. From the DINO family we use DINOv2 (Oquab et al., 2024), DINOv3 (Siméoni et al., 2025) and `dino.txt` (Jose et al., 2025). DINOv2 is a self-supervised model that combines image-level DINO supervision with patch-level iBOT-style masked image modeling. DINOv3 adds Gram anchoring to the training schedule, which greatly improves the quality of the local maps. As a compromise between computational time and model capabilities, we use the distilled ViT-B variant for DINOv2 and DINOv3. `dino.txt` is obtained by aligning a frozen DINOv3 ViT-L backbone with a text encoder (Zhai et al., 2022; Jose et al., 2025). From the Joint-Embedding Predictive Architecture (JEPA) family, we consider the ViT-L version of LeV-JEPA (Kuhn et al., 2026) which relies on a recent regularizer, SIGReg, that prevents representation collapse by pushing embeddings toward an isotropic Gaussian distribution. Moreover, it is trained on video and uses block-causal attention across frames. Following the authors' protocol for static-image inference, we construct a static video by repeating each image across the temporal dimension and extract features from the last frame, which has the most context under block-causal attention. We include two models from the Perception Encoder family (Bolya et al., 2026). PE-Core is an image-video encoder trained with contrastive targets on image-text and video-text data pairs, whose embeddings carry strong local dense semantic information. The authors showed that its intermediate layers (around 0.6 depth) contain the most relevant features for the purpose of downstream spatial tasks: PE-Spatial is obtained by further alignment of these frozen features for dense prediction tasks.

We do not include reconstruction-based MAE models, whose dense features have been reported to be weakly structured (Song et al., 2024; Vanyan et al., 2023; Oquab et al., 2024; Lüddecke & Ecker, 2025).

Datasets. We evaluate relative relations on three controlled synthetic datasets: a 2D dataset inspired by the dSprites (Matthey et al., 2017) dataset, a 3D dataset based on the CLEVR (Johnson et al., 2017) dataset, and a Pixel Art dataset built from sprites selected from Elgazar (2026). All three datasets are designed to reduce shortcut solutions based on absolute attribute values. For interactions, we use HICO-DET (Chao et al., 2018), which contains triplets of real-world human-object interactions, along with bounding boxes associated with each entity. Because our analysis probes local patch- and entity-level representations, the original bounding boxes are not precise enough, so we augment the HICO-DET images with dense segmentation masks generated with SAM 3 (Carion et al., 2026). Details about the synthetic datasets and HICO-DET mask generation are given in Appendix A and Appendix B.

Single-patch and Pairwise probes. To decode relational information between objects in an image, we train linear probes on the frozen representations $z^l \in \mathbb{R}^d$ at layer l, of either the `[CLS]` token or a patch of a target object. We use logistic regression to probe the presence/absence of a particular relation. For the absolute scale, in Section 4, we use standard linear regression. We use $L2$ regularization to avoid overfitting, with regularization strength $\lambda = 100$. To probe the interaction between two entities with embeddings z_i^l, z_j^l, we train low-rank bilinear probes $s(z_i^l, z_j^l) = (z_i^l)^\top W z_j^l$, where $W = UV^\top$, with $U, V \in \mathbb{R}^{d\times r}$ and $r \ll d$. We use $r = 4$ for all experiments, as larger ranks do not improve performance (see Appendix D, Fig. 11). For directional relations, we use an antisymmetric parametrization $W = UV^\top - VU^\top$, so that exchanging the two entities reverses the sign of the score (details in Appendix D).

4 Relational Encoding in Synthetic Datasets

Our goal is to understand whether ViT embeddings go beyond encoding purely local properties, such as shape and absolute scale, to encode more complex contextual relations. We begin by studying simple contextual relations in controlled synthetic settings. These relations concern properties whose definition depends on the surrounding objects, such as identifying the largest object in an image or the odd one out in a set of otherwise identical elements. We investigate these questions using the methods and the 2D, 3D, and Pixel Art datasets introduced in Section 3.

Object attributes from local patches. Before asking whether dense ViT representations encode relations between objects, we first verify that local patch tokens contain reliable information about the object they belong to. In particular, we consider object shape and object scale on both the 2D and 3D controlled datasets. As expected (Amir et al., 2022; Jose et al., 2025; Siméoni et al., 2025; Balestriero & LeCun, 2025; Bolya et al., 2026), shape is almost perfectly linearly decodable across all models in both domains, reaching at least 98.5% peak accuracy, while scale reaches $R^2 > 92\%$ in 2D and, except for DINOv2, $R^2 > 94\%$ in 3D. Detailed results are available in Appendix C.1.

Relative relations from local patches. To test whether ViT patch embeddings capture simple relational information, we train probes on two relational tasks. First, we look at relative object size. For linear probes, we ask whether a target patch z belongs to the largest object in the image. We consider the same relation for bilinear probes, which compare a source and a target patch and predict the relative scale of the objects the patches belong to. We train and evaluate these probes on the 2D and 3D datasets described in Section 3, where the scale of the largest object is sampled uniformly across the available range. This reduces the risk that the probes rely on absolute object size as a shortcut.

We also test whether ViT embeddings capture a variation of the concept of “sameness". Unlike Li et al. (2025), who used the term to identify patches belonging to the same “macroscopic" object. Here we ask to recognize the notion of “odd-one-out", i.e., the identity of an object relative to other objects in the same scene. Using the Pixel Art dataset introduced in Section 3, we train probes to predict whether a target patch belongs to the “odd-one-out" object in the image. At test time, we evaluate on entirely unseen objects. For each experiment, a train-test split is performed at the image

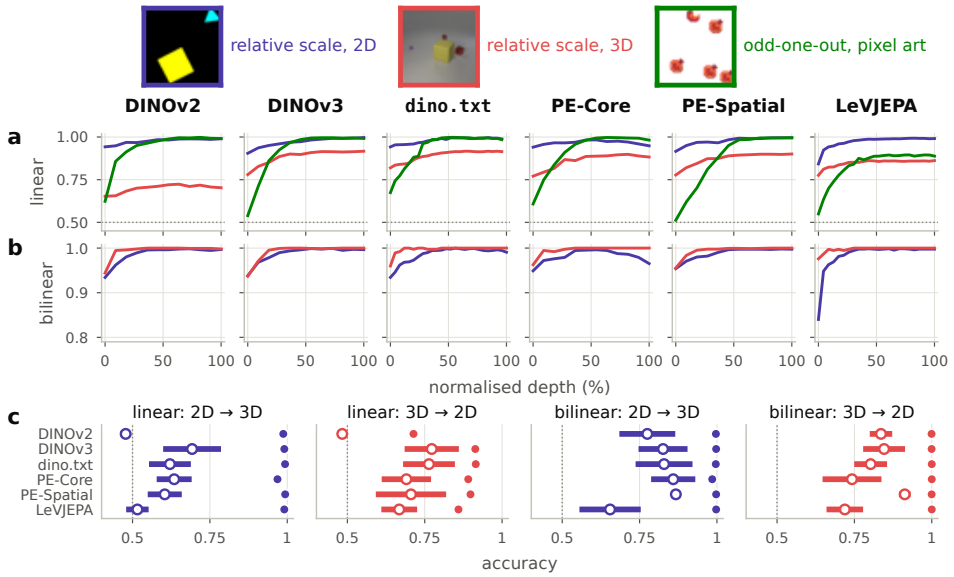


Figure 2: Relational probes on the synthetic datasets. Colors denote the task and dataset: relative scale for the 2D (blue) and 3D (red) datasets and odd-one-out in the Pixel Art dataset (green). **(a)** In-distribution accuracies of linear probes as a function of normalized model depth. **(b)** In-distribution accuracies of bilinear probes. **(c)** Cross-domain performance for the relative scale task (2D→3D, 3D→2D). Filled markers denote in-distribution accuracy. Hollow markers denote cross-domain accuracy, averaged over the last half of each model's layers (error bars span ±1 SD). Dotted lines mark chance level.

level to avoid correlated samples between the different splits, and patches are sampled randomly among the ones falling within the corresponding object mask.

Linear encoding. We first ask whether relative size is linearly accessible from a single local representation. Specifically, we train the linear probe defined in Section 3 on patches z^l to predict whether *patch z^l belongs to the largest object in the scene*. Relative size is highly linearly decodable in the 2D setting, with peak test accuracies ranging from 97% to 99% across models. Performance generally improves with network depth, and most models peak in the final layers. The exception is PE-Core, which drops from 97% to 94% in the final layers. Performance is lower in the 3D setting, although most models still reach the $87\% - 92\%$ range, well above the 50% chance level. DINOv3 achieves the highest overall performance, reaching approximately 92% accuracy at layer 11 of 12. In contrast, DINOv2 performs worst, with a maximum accuracy of about 73% at layer 10 of 12.

These results can be compared with the cross-domain tests, see Fig. 2c). Accuracy fluctuates with model's depth, so we report its mean (± SD) over the last half of each model. Linear probes trained on 2D perform markedly worse on 3D. Despite a 30-point drop, DINOv3 performs best, at $69 \pm 9\%$, followed by PE-Core ($64 \pm 6\%$), dino.txt ($62 \pm 7\%$) and PE-Spatial ($60 \pm 6\%$). LeVJEPA and DINOv2 are at chance level ($52 \pm 4\%$ and $48 \pm 2\%$). Despite performing worse in-distribution, probes trained on the 3D dataset generalize better. DINOv3 and dino.txt reach $77 \pm 9\%$ and $76 \pm 8\%$, the two PE models 69–71%, and LeVJEPA $67 \pm 6\%$. DINOv2 again remains at chance level ($48 \pm 1\%$).

The "odd-one-out" task is also linearly decodable from individual patch representations. Because the test images contain only unseen objects, this task already measures generalization to novel object identities. Peak accuracies exceed 96% for all models except LeVJEPA, which reaches 89%.

Bilinear relational probing. We next ask whether relative size is accessible from the joint representation of two local patches. Specifically, we train the bilinear probe defined in Section 3 to predict whether a target patch z^l_{target} belongs to a larger object than the one containing the source patch z^l_{source}.

Table 1: Composition of datasets for a target verb v. h, o (h', o') indicate participating (non-participating) humans and objects.

Probe	Positive	Soft negative	Hard negative
[CLS]	image w/ v	image w/o v	—
Human	h	h', image w/o v	h', image w/ v
Object	o	o' in $\mathcal{O}_v$, image w/o v	o', image w/ v
Bilinear	(h, o)	(h', o'), image w/o v	h and o but $(h, o) \notin \mathcal{A}_v(I)$

The results in Fig. 2b) show that bilinear probes are highly effective in extracting relative size from the joint representation of two patches. This is particularly evident in the 3D setting, where all models reach near-perfect accuracy ($\geq 99\%$) already by the third layer, 8 points above the best single-patch accuracy. In the 2D setting, bilinear decoding improves progressively with network depth, with high accuracies reached from the middle layers onward. All models peak above 99%. PE-Core is again the main exception: its performance peaks at 99% at layer 6 and subsequently decreases to 96% at the final layer.

Again, we test cross-domain accuracy by evaluating the probes trained on 2D data on the 3D test set, and vice versa. Bilinear probes trained on 2D images reach 66–87% on 3D, with PE-Spatial ($87 \pm 2\%$) and PE-Core ($86 \pm 7\%$) obtaining the highest values; all models except LeVJEPA ($66 \pm 10\%$) exceed the best linear transfer, obtained by DINOv3 ($69 \pm 9\%$). In the reverse direction, probes trained on 3D reach 72–91% on 2D, again with PE-Spatial highest ($91 \pm 2\%$) and well above the best linear result (DINOv3, $77 \pm 9\%$).

For the bilinear case, we study the probes' scores across continuous object scales in a small 3D test set (see Appendix C.2 for details). The scores follow the expected diagonal behavior of relative size ordering, indicating that the probes do not use absolute scale of a single object as proxy.

5 Interaction Encoding in Natural Images

The final class of relations considered in our analysis consists of human-object interactions in natural images from HICO-DET. This setting is substantially more challenging, as recognizing such interactions requires a more complex understanding of the entities involved and their interplay. HICO-DET annotates images with triplets (h, o, v), indicating that a human h interacts with an object o through the verb v. We write $\mathcal{O}_v$ for the object categories that appear with v in at least one annotation, and $\mathcal{A}_v(I)$ for the annotated pairs (h, o) in image I that carry the verb v. We say that a human or an object *participates* in v when it belongs to some pair in $\mathcal{A}_v(I)$; a single image may contain several interactions, and a human or object may be present in more than one.

We build a separate probing dataset for each verb and probe type, ensuring that the task cannot be solved through trivial cues. Object co-occurrence is the first of these tasks: a probe can recognize riding merely by detecting a bicycle. Therefore, we require every negative image to contain at least one object of a category in $\mathcal{O}_v$. We further distinguish soft from hard negatives according to whether the verb occurs anywhere in the image. For example, a soft negative for `riding` is a person standing beside a bicycle in an image where nobody rides; a hard negative is a bystander in an image where somebody else does. Table 1 summarizes the construction of the datasets. For the probes, we use the [CLS] token and, for each instance of h or o, the average of the patch embeddings within its segmentation mask. We train linear and bilinear probes on six of the most frequent verbs: *sitting_on, holding, carrying, straddling, riding, standing_on*.

Linear Interaction Encoding. We first study linear probes trained on the [CLS] token and on the pooled human (h) and object (o) patches. We report results only at the best-performing layer (selected from an initial run), averaged over three runs. Figure 3 shows mean accuracies over verbs for all models and probe types. On positive and soft-negative examples, the h and o probes perform moderately well, reaching 77–81% for all models except LeVJEPA, which remains around 70%. The two linear entity probes are close to each other: o matches or slightly exceeds h (by up to 3 points) for almost all models. The [CLS] probe is competitive for the DINO family (76.5–78.5%),

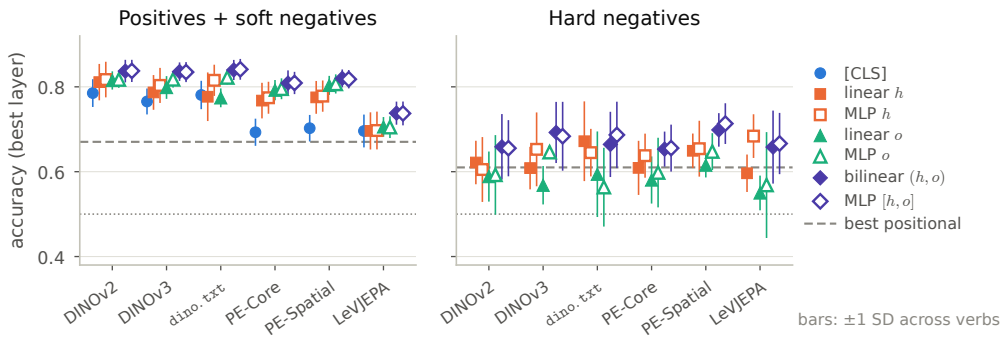


Figure 3: Average accuracy for all models on the HICO-DET dataset. Blue circles, orange squares, and green triangles show the linear probes on `[CLS]`, human, and object patches, respectively, while purple diamonds represent the bilinear probe. Hollow markers show the corresponding MLP probes. The dashed grey line marks the strongest positional baseline across models. Error bars: ± 1 SD across verbs. **(Left)** Average accuracy on positive + soft negative images. **(Right)** Average accuracy on hard negatives.

but considerably weaker for the two PE models (69–70%), where it trails the entity-based probes by 7–10 points. The picture changes for hard negatives, which are not defined for the [CLS] probes. Here, accuracy drops to 60–67% for h and to 55–62% for o, i.e., 10–23 points below the soft-negative setting, and the object probe is now consistently the weaker of the two. Hard negatives show that the probes can detect whether an interaction occurs in an image, but fail to distinguish participating from non-participating entities.

Bilinear Interaction Encoding. Next, we consider the bilinear probes. On positives and soft negatives, the bilinear probes improve consistently over the linear probes. In particular, models from the DINO family improve by $5-7$ points over the `[CLS]` probe, those from the PE family gain an even larger 12 points, and LeVJEPA gains 4. On hard negatives, performance drops again, and the differences between models largely vanish: all bilinear probes lie between 65% and 70%, with PE-spatial (69.8%) and DINOv3 (69.2%) performing best. For most models, this is a gain of 4–8 points over the linear h probe, with the exception of `dino.txt`, whose bilinear probe (66.4%) does not improve over its linear h probe (67.2%). Nevertheless, accuracy remains $8-18$ points below the soft-negative setting.

Positional probe. Spatial proximity and positional cues can act as proxies for true interactions. To estimate this effect, we train the bilinear probe on the same datasets, replacing each image with pure Gaussian noise before encoding it (Cuttano et al., 2026). Pooling the resulting patches over the same masks gives the tokens h_{noise} and o_{noise}, which carry only the position and extent of the two entities. Because the masks (the only available source of information) are identical for all backbones, we treat the strongest positional probe across models as the highest capability observed on position alone, and use it as a common reference for all models in Fig. 3. The positional probe reaches 63–67% for all models on positive and soft negatives. On hard negatives, it peaks at 61% for `dino.txt` and 60% for DINOv3, with models from the PE family and LeVJEPA performing 10 points worse. Against this reference, single-entity probes (55–68%) perform at a comparable level, whereas pair probes (65–71%) exceed it by 4–10 points for every model.

MLP probe. To investigate whether adding non-linearity to the probes improves decodability, we train MLP probes on h, on o, and on their concatenation $[h, o]$. On positive and soft-negative examples, the MLP on the joint $[h, o]$ matches the bilinear probe in performance, and remains comparable even for the hard negatives. For positive and soft negatives, the single-entity MLPs perform on average better than their linear counterparts, but by a small margin (at most $4-5$ points for `dino.txt`), while on hard negatives the gain is inconsistent across models (-3 to $+9$ points).

Generalization to OOD objects. To test the generalization capabilities of bilinear probes, we perform a leave-one-out analysis on the objects. For each verb v and each compatible object category

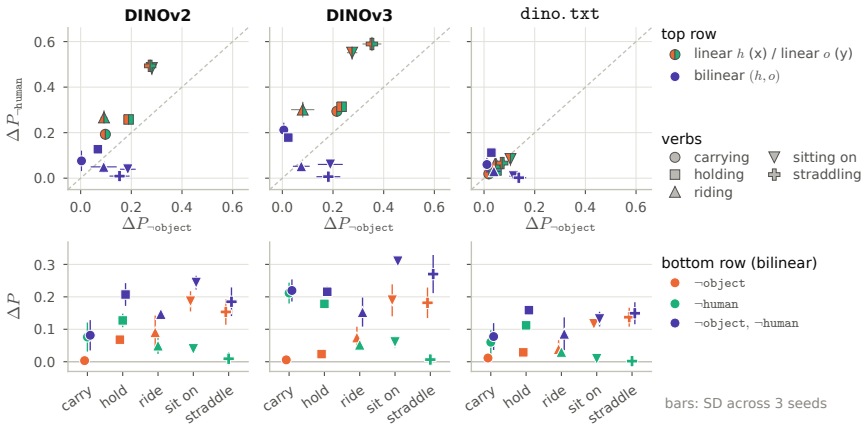


Figure 4: Inpainting results for the DINO family. **(Top)** $\Delta P_{\neg\texttt{object}}$ vs. $\Delta P_{\neg\texttt{human}}$ for linear and bilinear probes. **(Bottom)** bilinear probe, all three combinations of inputs. Shape denotes the verb. Error bars are reported as SD for 3 seeds.

$o \in O_v$, we train a probe on all images except those containing (v, o), and evaluate it on the held-out images. To limit imbalance across object categories, the positives of each (v, o') pair are drawn from at most 50 images. The "held-out performance" drop Δ_{OOD} is computed in Appendix F, comparing the probes trained with and without the held-out class. We observe some degree of generalization: for 50% of the held-out instances, performance remains close to that of the original probe ($|\Delta_{OOD}| < 5\%$) (with a spread of $30 - 65\%$ depending on the model). To exclude that this is a purely spatial effect, we compare the drops Δ_{OOD} with those of the positional probe trained with the same protocol: the drops are uncorrelated ($\rho \in [-0.09, 0.23]$, all confidence intervals including zero).

We then ask whether a held-out category is harder when it is more separated from the other categories. We define the distance of a class as the mean cosine distance to their three nearest samples of the *other* categories, averaged over all its instances. We compute it either for the real tokens $[h, o]$ ("semantic" distance) or on $[h_{\texttt{noise}}, o_{\texttt{noise}}]$ ("positional" distance). We find that the semantic distance predicts drops better than the positional one (ρ in $[-0.30, -0.19]$ against $[-0.25, -0.03]$), suggesting that the OOD generalization is not explained by geometry alone. Details and per-verb results are in Appendix F.

Inpainting analysis. To see what a probe relies on, we perform a counterfactual intervention on a subset of *positive* test examples for *carrying*, *holding*, *sitting_on*, *riding*, and *straddling*. From each image, we inpaint either the human or the object with LaMa (Suvorov et al., 2021), obtaining a $\neg\texttt{human}$ and a $\neg\texttt{object}$ image. Using the masks of the original image, we pool the human patches of the $\neg\texttt{object}$ image to obtain $h_{\neg\texttt{object}}$, and the object patches of the $\neg\texttt{human}$ image to obtain $o_{\neg\texttt{human}}$. We report the change ΔP in the probe score with respect to the original probes for the DINO family of models, averaged over three training seeds (Fig. 4).

For linear probes Fig. 4, (Top), the object probe is consistently more sensitive to the removal of the human, with $\Delta P_{\neg\texttt{human}} = 19 - 59\%$ range while $\Delta P_{\neg\texttt{object}} = 8 - 35\%$. The probes, however, remain close to the best positional baseline on hard negatives (Fig. 3). This suggests that both probes rely on proxies: the human probe on pose and the object probe on proximity to a person. Probes for `dino.txt` are much less sensitive ($\Delta P < 10\%$). Results for the bilinear probes, shown in Fig. 4, (Bottom), have the largest drop when both entities are inpainted. However, verbs seem to have different behaviors: for *carrying* and *holding* the evidence is carried by the object entity, while for the remaining verbs the pattern is reversed.

6 Discussion and Conclusion

In this work, we investigate which forms of relational information are accessible from the frozen representations of visual foundation models. We progress from object attributes to relative com-

parisons and finally human-object interactions. Our results demonstrate that, even for those models that are shown to encode rich semantic information within local patches, extracting information about relational and physical interactions between objects remains challenging. The experiments on synthetic datasets show that the proposed methodology can extract contextual information from the dense feature map, showing some degree of generalization of abstract concepts such as identity and scale across different data domains or unseen entities.

When studying natural human-object interactions (HICO-DET), however, we find clear limitations. Linear, bilinear and MLP probes detect the presence or absence of complex physical interactions with reasonably high accuracy, with the best performing pairwise probes reaching 74–84% on average. However, when tasked to distinguish between entities that actively participate in an interaction from mere bystanders, accuracy drops significantly. On these hard negatives, single-entity probes (55–68%) are comparable to the best purely positional probe (61%). Object probes at most exceed it by 4 points, and human probes by 7. Pair probes exceed it by 4–10 points. Out-of-distribution testing shows some degree of generalization for bilinear probes, which stay within 5 points of their in-distribution accuracy for about half of the held-out object categories.

Increasing probe capacity does not significantly change this result. Single-entity MLP probes provide only modest average improvements over their linear counterparts, and this improvement is inconsistent on hard negatives. Moreover, the MLP operating on concatenated human and object representations performs similarly to the bilinear probe. Therefore, the main limitation appears to be having access to single-entity local representations and not the linearity of the probes.

Counterfactual inpainting and positional probing reveal that part of the recoverable interaction signal can be explained by spatial proxies and pose information. On the other hand, out-of-distribution testing indicates that the pairwise signal cannot be explained by spatial geometry alone. Counterfactual inpainting interventions are particularly useful for identifying which probing strategies are active for different interactions: replacing the human or the object of an interaction with an inpainted background affects the probes differently. Human probes are only weakly affected by object removal, consistent with a strong dependency on human pose. The object probe on the contrary is strongly affected by inpainting the human; however, on hard negatives it does not exceed the best positional baseline, suggesting potential shortcuts related to spatial proximity. For the bilinear probes, variations are visible across verbs. Taking for example the DINOv3 model, removing the human strongly affects results for “carrying" but not at all for “straddling". These results indicate that the probe can rely on 2D human poses and spatial proximity rather than on a deeper understanding of the interaction encoded in the underlying embedding.

Finally, we note that the performance is not substantially changed when different training modes are included. Representations from the DINOv2 and DINOv3 models, which are self-supervised on static images, encode relational information to a limited extent. `dino.txt` augment the SSL paradigm with a fine-tuning step of text-alignment, which in principle could enrich the representation with the semantic understanding contained in the captions. This effect does not stands out in our experiments, nor it does for the PE models which are natively trained with a contrastive target of image-text and video-captions pairs. The PE-spatial model is fine-tuned on spatial understanding, and it does improve on its PE-Core base model by 4–6 points on hard negatives, where it matches the static-image DINOv3. Lastly, the LeVJEPA is an architecture which is natively trained on video, where the attention blocks are exposed to the full temporal causality of videos. When analyzing static images, however, this model appears to rely on the same shortcuts as image-trained models. This does not exclude that its video-based training may provide an advantage when genuine temporal information is available.

Probes extract a strong signal of the presence or absence of complex human-object physical interactions from the dense feature maps of foundation models. Single patch probes, however, fail to discriminate which local entities actively participate in the interactions better than a purely positional baseline, indicating that this information is not decodable from local patches only. Pair probes fare better, but still lose 8–18 points relative to the simpler detection task. We conclude that the reliance on proxies hinders the utility of ViTs encodings beyond object-centric tasks. Although extremely effective for segmentation purposes, retrieval, depth-estimation, etc., ViTs' feature maps do not immediately translate into the true scene understanding needed to transform image encodings into queryable objects. Our results suggest that, for physical interactions, a natural direction for scene understanding remains that of joint read-out of entities.

Declaration of AI use. In this work, we used generative AI tools at several stages of the research process, primarily for software implementation, literature summarization, and the initial drafting of the research paper. During the initial literature review, we used generative AI tools to summarize the methodologies of papers that we had manually identified as relevant to our work.

For software development, we used AI-assisted development tools, such as Cursor and Claude Code. These tools assisted us in implementing the code used to generate the synthetic datasets and to conduct the experiments presented in this work. AI-generated code was reviewed by the authors to verify its consistency with the implementation plans, its correctness, and its computational efficiency. AI tools were also used to develop part of the data analysis code and to generate part of the scientific figures.

We also made limited use of AI tools during the final revision of the manuscript to improve readability, correct language, and assist with translation. We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine research hypotheses, design or provide feedback on the research methodology or experiments, or interpret the experimental results. We did not leverage generative AI for qualitative or thematic analysis, dataset cleaning, or reformatting. All experiments and data analysis were designed and carried out by the authors.

We take full responsibility for the final content of the paper, including all artifacts produced or assisted by generative AI.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. In *European Conference on Computer Vision Workshops (ECCVW)*, 2022. What is Motion For? Workshop.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics, 2025. URL https://arxiv.org/abs/2511.08544. arXiv preprint arXiv:2511.08544.

Anand Bhattad, Konpat Preechakul, and Alexei Efros. Visual jenga: Discovering object dependencies via counterfactual inpainting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2026. URL https://arxiv.org/abs/2511.16719. arXiv preprint arXiv:2511.16719.

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)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV.2021.00951.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 381–389, 2018. doi: 10.1109/WACV.2018.00048.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Fenil R. Doshi, Thomas Fel, Talia Konkle, and George Alvarez. From local to contextually-enriched local representations: A mechanism for holistic processing in dinov2 vits. In *Mechanistic Interpretability Workshop at NeurIPS*, 2025. URL https://openreview.net/forum?id=2XA7zVjoej.

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. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806, 2024.

Ebrahim Elgazar. Pixel art. Kaggle Dataset, 2026. URL https://www.kaggle.com/datasets/ebrahimelgazar/pixel-art. Accessed: 2026-09-09.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997, 2017. doi: 10.1109/CVPR.2017.215.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, and Maxime Oquab. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24905–24916, 2025.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568.

Turhan Can Kargin, Wojciech Jasiński, Adam Pardyl, Bartosz Zieliński, and Marcin Przewiélikowski. Sparrta: A synthetic benchmark for evaluating spatial intelligence in visual foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Lukas Kuhn, Lucas Maes, Giuseppe Serra, Quentin Le, Yann LeCun, Randall Balestriero, and Florian Buettner. Levjepa: Efficient and scalable video pretraining without the heuristics. *arXiv preprint arXiv:2608.27395*, 2026.

Michael A. Lepori, Alexa R. Tartaglini, Wai Keen Vong, Thomas Serre, Brenden M. Lake, and Ellie Pavlick. Beyond the doors of perception: Vision transformers represent relations between objects, 2024. URL https://arxiv.org/abs/2406.15955. arXiv preprint arXiv:2406.15955.

Yihao Li, Saeed Salehi, Lyle Ungar, and Konrad P. Kording. Does object binding naturally emerge in large pretrained vision transformers? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Timo Lüddecke and Alexander S. Ecker. Characterizing vision backbones for dense prediction with dense attentive probing. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=neMAx4uBlh.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

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. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104. arXiv preprint arXiv:2508.10104.

Xinrui Song, Xuanang Xu, and Pingkun Yan. DINO-Reg: General purpose image encoder for training-free multi-modal deformable medical image registration. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2024)*, volume 15002 of *Lecture Notes in Computer Science*, pp. 608–617, 2024.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions, 2021. URL https://arxiv.org/abs/2109.07161. arXiv preprint arXiv:2109.07161.

Ani Vanyan, Alvard Barseghyan, Hakob Tamazyan, Vahan Huroyan, Hrant Khachatrian, and Martin Danelljan. Analyzing local representations of self-supervised vision transformers, 2023. URL https://arxiv.org/abs/2401.00463. arXiv preprint arXiv:2401.00463.

Chuan Wen, Dinesh Jayaraman, and Yang Gao. Can transformers capture spatial relations between objects? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/9973907a2acfd04f030cd244e6a0fe35-Paper-Conference.pdf.

Sicheng Yang, Hongqiu Wang, Zhaohu Xing, Sixiang Chen, and Lei Zhu. Segdino: An efficient design for medical and natural image segmentation with dino-v3, 2025. URL https://arxiv.org/abs/2509.00833. arXiv preprint arXiv:2509.00833.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning, 2022. URL https://arxiv.org/abs/2111.07991. arXiv preprint arXiv:2111.07991.

Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3d physical understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 43468–43498, 2024. doi: 10.52202/079017-1377.

A Details on Datasets

Since multiple patches can be extracted from the same image, for all experiments and all datasets, train-test split is performed at image level prior to any computation. In particular, we performed 80-10-10 train-validation-test and ensured that all extracted patches belonging to the same images belong to the same split, to avoid correlated samples.

A.1 2D DATASET

Inspired by the dSprites dataset (Matthey et al., 2017), we identified color, shape, scale, and rotation as the main object features and built a pool of possible values for each variable. In particular, we considered 6 colors, 6 shapes, 11 scale values distributed in $[0, 1]$, and 40 possible rotation values, evenly distributed in $[0, 2\pi]$.

To generate each sample, we randomly place two or three objects on a 224×224 canvas, ensuring that their borders do not overlap. Object features are all randomly sampled, except for scale. The scale value is first assigned to the largest object and is uniformly sampled among all possible values in $[0.3, 1.0]$. At this point, the scales of the other objects are randomly sampled among the remaining possible values, ensuring a difference of at least 0.15 from the closest object in scale.

This construction allows us to obtain a uniform distribution of all features across the objects in the dataset, as well as a uniform distribution of scales for the largest object in each image, which helps prevent shortcuts in prediction, i.e., learning that if the scale is higher than a certain threshold, then the object is the largest in the image.

We generated a dataset of 3000 images, half of which contain two objects and the other half three. An array of binary masks, taking value 1 for the pixels belonging to a specific object and 0 otherwise, is associated with each image to later locate the precise patches to be extracted.

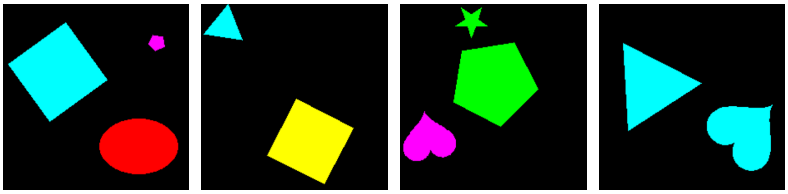


Figure 5: Samples from the generated 2D dataset

A.2 3D DATASET

The 3D dataset follows the same general construction logic as the previously described 2D dataset. We used the code provided with the CLEVR dataset to render a new set of CLEVR-like images in Blender. We retained the same object types and corresponding values as in the original dataset. We introduce an additional size category which we name “tiny”, with a smaller scale than the existing “small” and “large”.

We then rendered scenes containing a random number of objects (between 2 and 6), with all object features randomly sampled except scale, which we explicitly control each scene to ensure that the relatively largest object cannot be identified from its absolute scale alone. Moreover, each image contains a unique largest object, to avoid ambiguity.

Each image has pixel-level object masks for each of its objects. We generate up to 3000 images, split into training, validation, and test sets.

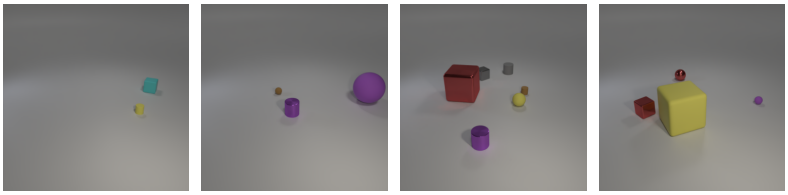


Figure 6: Samples of 3D dataset

A.3 PIXEL ART DATASET

The original Pixel Art dataset contains 89,000 pixel art images of different characters and objects extracted from an online game and divided into 5 classes (Elgazar, 2026).

To construct the training and validation sets, we manually selected 88 visually distinct elements from the first four classes, with at least 20 elements selected from each class. We have excluded the fifth class because visually it is too similar to class 1.

To construct the test set, we followed the same selection procedure using a *separate* pool of elements, selecting at least 10 additional elements per class. Thus, elements appearing in training and test are effectively all different.

For each image, we select two different elements from the corresponding pool. One of the two elements is designated as the odd object and appears once, while the other is repeated four times. The resulting five instances are placed at random valid positions on the canvas while ensuring that their bounding boxes do not overlap.

We generated 3,728 images for training and validation, equally divided between two types of examples: in half of the images, the odd and repeated elements belong to the same original class, while in the other half they belong to different classes. Element pairs are sampled, with the same pairs appearing only once in the dataset. We approximately balance how frequently each element appears as the odd-one and in the repeated role. The resulting images were subsequently divided into 90% training and 10% validation data. For the test set, we generated 1,500 images according to the same procedure, but exclusively using elements from the held-out pool.

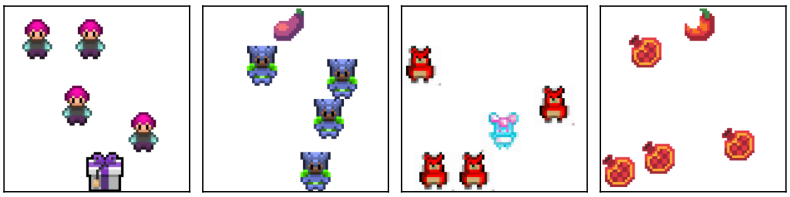


Figure 7: Samples from Pixel Art training set.

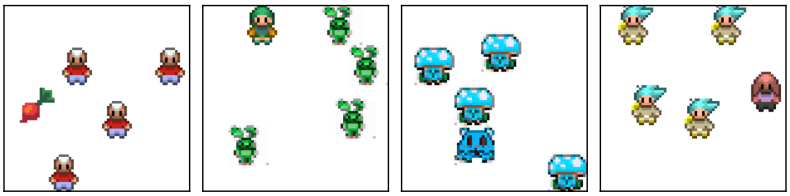


Figure 8: Samples from the Pixel Art test set containing completely different sprites with respect to the training set.

B SAM 3 DENSE MASKS FOR HICO-DET

Our experiments operate on local ViT patch tokens and require a high precision in assigning individual patches to the human and object entities. HICO-DET provides human-object interaction annotations as bounding boxes, which are too coarse for this purpose. The bounding box for an human will include almost surely a large portion of background, nearby instances. Bounding boxes often have large overlaps between pair of interacting entities. To solve this, we compute dense masks for each individual. This allows us to keep the original HICO-DET labels and splits unchanged, at the same time obtaining a cleaner entity-level support for patch pooling and patch-level analyses.

We process the original HICO-DET train and test splits independently. For each annotated HOI, we recover the human-object pairs from the HICO annotations. Each instance is segmented with SAM 3 using both a box prompt and a text prompt. Human boxes are prompted with "person", while object boxes are prompted with the corresponding HICO object category.

SAM 3 is run with confidence threshold 0.3. When multiple masks are returned, we select the candidate with the largest overlap with the HICO box B:

$$M^\star = \arg\max_M |M \cap B|.$$

In the case no candidate overlaps B, we use the largest predicted mask. The selected mask is then clipped to the original HICO box, i.e. we store $M^\star \cap B$, using the annotated instance box as a ground-truth for the real extension of the instance.

To avoid assigning differently the same entity across multiple HOI annotations, we reuse previously computed masks within each image when the current HICO box has IoU > 0.7 with an existing box.

If this fails, we run SAM 3 and compare the tight box of the predicted mask with previous tight boxes, using IoU > 0.6 for humans and IoU > 0.3 for objects. Object masks are merged only when the HICO object category is identical. As a final merge criterion, we reuse an existing mask when either

$$\frac{|M_{\text{new}} \cap M_{\text{old}}|}{|M_{\text{new}}|} \geq 0.85 \quad \text{or} \quad \frac{|M_{\text{new}} \cap M_{\text{old}}|}{|M_{\text{old}}|} \geq 0.85.$$

For each image, we save one binary PNG mask per distinct human or object instance.

HICO-DET does not provide explicit identities' annotations for the entities participating in each interaction. Therefore, when a human or object is involved in multiple interactions, the annotations do not allow to recover whether they refer to the same underlying entity. The mask merging procedure allow us to recover identities, which are important when constructing within-image hard negatives.

C Further Experiments

C.1 Decoding Absolute Attributes

We investigate the decodability of object attributes, i.e., properties that can be defined for a single object independently of the other entities in the scene. In particular, we consider object shape and object scale on both the 2D and 3D controlled datasets. These tasks test whether dense patch embeddings act as meaningful object descriptors before we move to relational labels that require comparison across entities.

For each image, we assign ViT patches to object instances using the ground-truth object masks. A patch is considered an object patch when at least a fraction t of its pixels overlap with the corresponding object mask. For each layer, we train a linear probe with L_2 regularization ($\lambda = 100$ inverse regularization parameter) on the patch embeddings.

For shape prediction, we train a multiclass logistic classifier on a balanced set of object patches and evaluate performance using accuracy on a held-out test set.

For scale prediction, we use the same patch-level representations but train a linear regressor to predict the numeric scale of the corresponding object and evaluate scale decodability by reporting the coefficient of determination R^2 on the held-out test set.

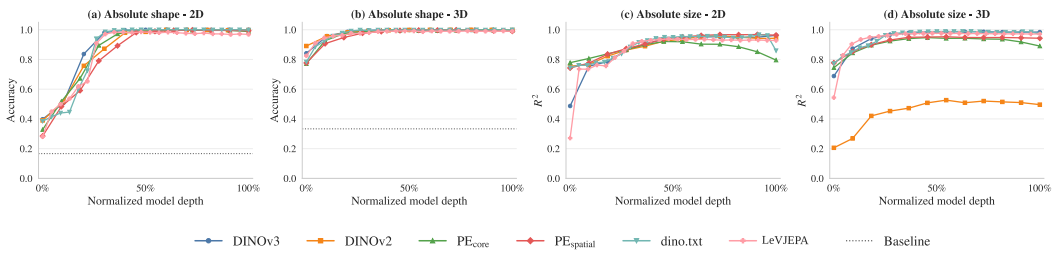


Figure 9: Absolute attributes in synthetic datasets. (**a**) Accuracy over layer of linear probe for absolute shape detection in 2D domain. (**b**) Accuracy over layer for absolute shape detection in 3D domain. (**c**) R^2 as function of model depth of linear regressor predicting absolute scale in 2D domain. (**d**) R^2 over layers for scale prediction in 3D.

In both 2D and 3D dataset, the results show that probes are able to access absolute object attributes from local patch embeddings with high accuracy. Shape is reliably encoded by local patch tokens: every model reaches at least 98% peak accuracy in both 2D and 3D. In the 3D setting, all models already exceed 90% accuracy within the initial fifth of their model depth. Scale is also strongly linearly decodable. All models obtain a peak R^2 above 92% in 2D, and, with the only exception of DINOv2, the models exceed 94% in 3D. DINOv2 reaches only a $R^2 = 53\%$ for 3D scale.

C.2 Beyond absolute scale

To check whether the probes can detect the proper relative scale without proxying it with the absolute scale, we can do a simple experiment where we average the probability that an object i is larger than an object j conditioned on their size. A probe proxying on absolute scale can predict only a vertical transition, while the bilinear probe does correctly predict a diagonal transition; see Fig. 10

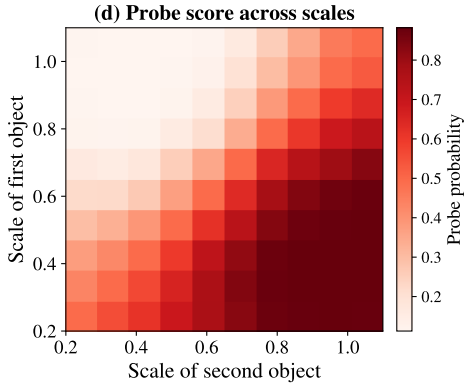


Figure 10: Results for the relative scale task by bilinear probe. Averaging the score probabilities conditioned on the size of both elements show that the probe is capable of going beyond an absolute scale proxy.

D Probing methods

Let $f(\cdot)$ denote a frozen visual encoder and let $z^l \in \mathbb{R}^d$ be the representation at a given layer l. In our work, this representation may be either of a single token, such as the CLS token or a single image patch, or an average from an object-level entity, obtained by averaging inside a segmentation mask.

A linear probe $g(\cdot)$ is trained on top of z^l while keeping the encoder parameters fixed:

$$\hat{y} = g(z^l) = Wz^l + b, \tag{1}$$

where $W \in \mathbb{R}^{k \times d}$ and $b \in \mathbb{R}^k$ are the trainable parameters of the probe, with k denoting the dimensionality of the target space.

For categorical targets (shape in 2D, scale in 3D, and odd-one-out in Pixel art), the probe is implemented either as a multinomial or binary logistic classifier, whereas for continuous targets (scale in 2D) it is implemented as a linear regressor. We apply L_2 regularization with $\lambda = 100$.

To go beyond single patch probing, we use methods that train on pairs. Let $z_i^l, z_j^l \in \mathbb{R}^d$ denote the representations of two objects. The general form of a bilinear probe reads

$$s(z_i^l, z_j^l) = (z_i^l)^\top W z_j^l,$$

where $W \in \mathbb{R}^{d \times d}$ is the learned parameter matrix.

To reduce the number of trainable parameters, we constrain the relation representation to a low-dimensional space by imposing a low-rank factorization $W = UV^\top$, with $U, V \in \mathbb{R}^{d \times r}$ and $r \ll d$. In all experiments, we use $r = 4$, since increasing bilinear rank does not improve results, as can be seen in Figure 11.

Defining

$$u_i = U^\top z_i^l, \qquad v_j = V^\top z_j^l,$$

we obtain

$$s(z_i^l, z_j^l) = (z_i^l)^\top U V^\top z_j^l = u_i^\top v_j.$$

In our implementation, we normalize the low-rank vectors; hence, the bilinear score becomes

$$s(z_i^l, z_j^l) = \frac{u_i}{\|u_i\|_2}^\top \frac{v_j}{\|v_j\|_2}.$$

For directional relations, we impose antisymmetry by constraining $W = -W^\top$. In the low-rank setting, this is achieved through

$$W = UV^\top - VU^\top,$$

which yields the relation score

$$s(z_i^l, z_j^l) = \left(U^\top z_i^l\right)^\top \left(V^\top z_j^l\right) - \left(V^\top z_i^l\right)^\top \left(U^\top z_j^l\right),$$

guaranteeing that $s(z_i^l, z_j^l) = -s(z_j^l, z_i^l)$.

The final bilinear score is mapped to the classifier logit by a learned scale and a bias,

$$\hat{s}(z_i^l, z_j^l) = \alpha\, s(z_i^l, z_j^l) + b,$$

where $\alpha, b \in \mathbb{R}$.

For the bilinear probes, the projected representations Uz_i^l and Vz_j^l are L_2-normalized before computing their dot product, followed by a learned scale and bias. We optimize the bilinear probes with binary cross-entropy on the logits using AdamW with a learning rate of 5×10^{-3} for 500 steps. For the MLP probes, we L_2-normalize z_i^l and z_j^l, concatenate them into a single vector, and pass the resulting $2d$-dimensional vector through an MLP with one hidden layer of dimension $d_{\text{MLP}} = 32$ and a ReLU activation, producing a scalar score. The MLP probes are optimized for 2,000 steps. Predictions are obtained by applying a sigmoid to the resulting logit.

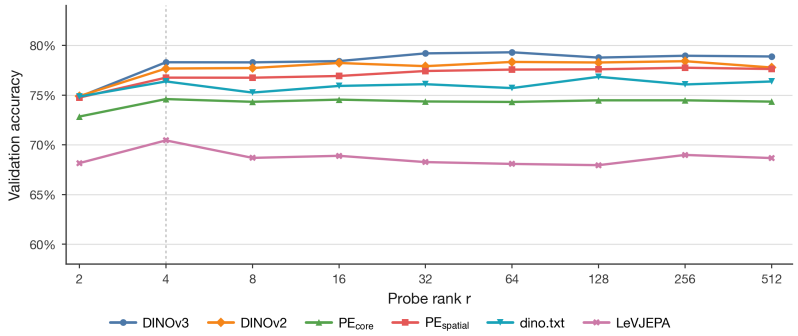


Figure 11: Rank ablation for the bilinear probe. Validation accuracy as a function of the bilinear probe rank r, averaged across all verbs. The rank sweep is taken at each model's final layer; the dotted line marks the chosen rank $r = 4$.

E Inpainting per-image results

Fig. 12 shows the results of inpainting at the image level. We report the probability drop ΔP for all images, calculated from the bilinear probes trained on the DINO family.

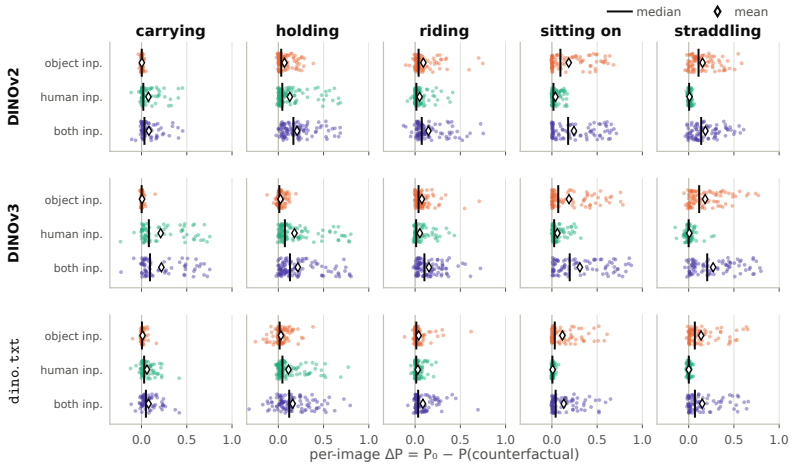


Figure 12: Per-image distribution of the counterfactual probability drop $\Delta P = P_0 - P(\text{counterfactual})$ of the bilinear probes, for the DINO family and the five verbs considered in inpainting. Rows: human input from modified image $(h_{\neg o}, o)$, object input from modified image $(h, o_{\neg h})$, and both inputs from modified images $(h_{\neg o}, o_{\neg h})$. The median is shown as a black bar, while the mean is shown as diamonds.

F OOD ANALYSIS

For each verb, we ask whether the drop in accuracy on a held-out object category can be explained by geometry alone (the category having positional relations different from all other categories) or by semantics alone (the category having embeddings different from all other categories). To do so, for each pair (h, o) we compute the mask-averaged embeddings of the original image ("semantic") or of a Gaussian-noise image with the same masks ("positional"). We then measure the *class distance* of a category by finding, for each of its instances, the three nearest neighbours among the pairs of *other* categories under cosine distance, and averaging these distances over the neighbours and over the instances of the category.

We compute Spearman's ρ between the class distance, positional and semantic, and the OOD drop of the category when it is held out from training, pooling the categories of the four verbs with a sufficiently large object vocabulary (riding, carrying, sitting on, holding; Fig. 14, Fig. 13 and Table 2). Confidence intervals are 95% percentile intervals from 4000 bootstrap resamples of the categories. Standing on and straddling, with five and three valid object categories, are excluded: holding out one category there removes a large fraction of the training set, so the drop also reflects the loss of training data rather than transfer to a novel object alone.

Semantic distance correlates negatively with the drop for every model, significantly so for DINOv2, DINOv3 and PE-Core: categories whose appearance is unlike the training categories lose more accuracy when held out. Positional distance shows no such relation for DINOv2, DINOv3 and PE-Spatial, and a weaker one for the remaining models. The effect is modest in either case ($|\rho| \leq 0.3$); most of the per-category variation in the drop is not explained by the distance of the category from the training pool.

		semantic (real images)		positional (noise images)	
model	n	ρ	95% CI	ρ	95% CI
DINOv2	76	$\mathbf{-0.30}$	$[-0.50, -0.07]$	-0.07	$[-0.29, +0.16]$
DINOv3	76	$\mathbf{-0.26}$	$[-0.46, -0.05]$	-0.03	$[-0.26, +0.22]$
`dino.txt`	75	-0.22	$[-0.44, +0.03]$	-0.17	$[-0.38, +0.05]$
PE-Core	76	$\mathbf{-0.23}$	$[-0.42, -0.02]$	$\mathbf{-0.25}$	$[-0.46, -0.02]$
PE-Spatial	76	-0.21	$[-0.41, +0.03]$	-0.11	$[-0.33, +0.13]$
LeVJEPA	72	-0.19	$[-0.41, +0.05]$	-0.16	$[-0.39, +0.07]$

Table 2: Pooled Spearman correlation between the OOD drop of the bilinear probe and the class distance of the held-out category, for real-image embeddings (semantic) and Gaussian-noise embeddings (positional), over riding, carrying, sitting on and holding. n is the number of held-out categories; intervals are 95% bootstrap confidence intervals over categories; bold marks intervals that exclude zero.

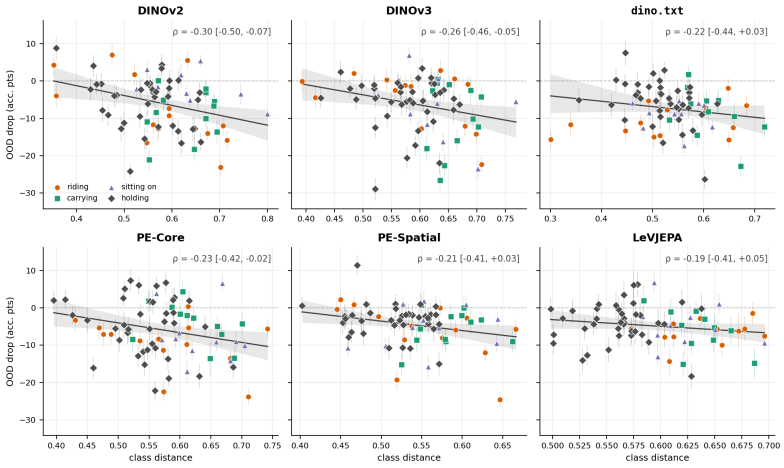


Figure 13: OOD drop for the bilinear probes against the class distance of the held-out object category, computed on the embedding of the original images. The solid line shows the least-squares fit with a 95% bootstrap error. ρ is the within-verb Spearman correlation with 95% confidence interval.

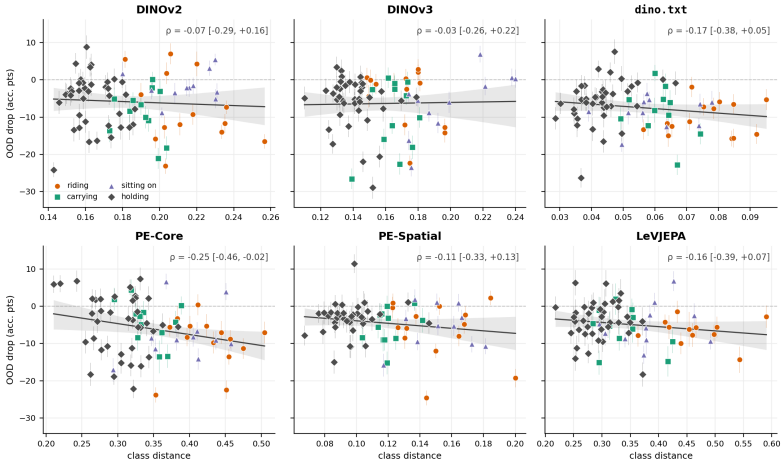


Figure 14: OOD drop for the bilinear probes against the class distance of the held-out object category, computed on the embedding of Gaussian noise images. The solid line shows the least-squares fit with a 95% bootstrap error computed by Bootstrap. ρ is the within-verb Spearman correlation with 95% confidence interval.