WHO ACTS ON WHOM? PROBING THE BINDING OF OBJECTS TO ROLES IN VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans understand scenes not only by recognizing which objects are present but also by determining who acts on whom, that is, by assigning objects to event roles, such as agent and patient. Replicating this human capability could provide artificial intelligence with a foundation for representing and reasoning about relations between objects. Recent work shows that Vision Transformers (ViTs) learn to encode, in low-dimensional subspaces of their representations, whether two patches belong to the same object. However, it remains unclear how these models represent the binding of different objects to agent and patient roles. CLIP-style models with ViT image encoders are known to struggle to distinguish captions in which these roles are swapped, raising the question of whether role information is absent from their representations or present but insufficiently used. In this paper, we train a bilinear probe on ordered pairs of object patches and find that the binding of objects to roles is encoded in a *relation-binding subspace* of patch representations in CLIP-style models. Given the region of one object in a relation, the probe assigns high scores to the region of its partner (the agent given the patient, and vice versa). Ablating this subspace reduces image–text matching accuracy on captions with role swaps more than on any other type of caption modification, whereas ablating a random subspace of the same dimension has no such effect.

1 INTRODUCTION

When understanding scenes, humans not only recognize which objects are present but also determine who acts on whom (Griffin & Bock, 2000; Hafri et al., 2013). For example, "a person riding a horse" and "a horse riding a person" contain the same objects and event but describe different scenes. Understanding this difference requires assigning the person to the agent role (the one who rides) and the horse to the patient role (the one being ridden). This assignment of objects to event roles, known as role–filler binding (Hummel & Holyoak, 1997), provides a foundation for representing and reasoning about relations between objects. Together with object binding (Treisman, 1996), it helps ground symbols in perception and support compositionality (Greff et al., 2020). Vision–language models (VLMs) such as CLIP (Radford et al., 2021), whose image encoders are typically Vision Transformers (ViTs) (Dosovitskiy et al., 2021), are known to fall short of this ability. Given an image and two captions that differ only in which object plays which role, their accuracy is only moderately above chance (Yuksekgonul et al., 2023; Hsieh et al., 2023; Parcalabescu et al., 2022).

Such failures admit two readings. Either the role assignments are absent from the internal representations, or they are present but insufficiently used. Image–text matching accuracy alone cannot tell the two apart. For attribute–object binding, Koishigarina et al. (2026) show that the second reading holds; CLIP encodes the binding within each modality yet fails to align this information across modalities. Whether the same holds for agent–patient assignments is open, and answering it requires looking inside the image encoder. Li et al. (2025) show that whether two image patches belong to the same object can be decoded from a low-dimensional subspace of patch representations. This relation is symmetric, whereas role–filler binding is asymmetric. We hypothesize that ordered agent–patient associations can nonetheless be decoded from a low-dimensional subspace.

In this work, we extract such a subspace from patch representations of CLIP-style models, CLIP and SigLIP (Zhai et al., 2023), and call it the *relation-binding subspace*. As illustrated in Figure 1, we train a bilinear probe on ordered pairs of object patches to distinguish annotated relation pairs in the

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.

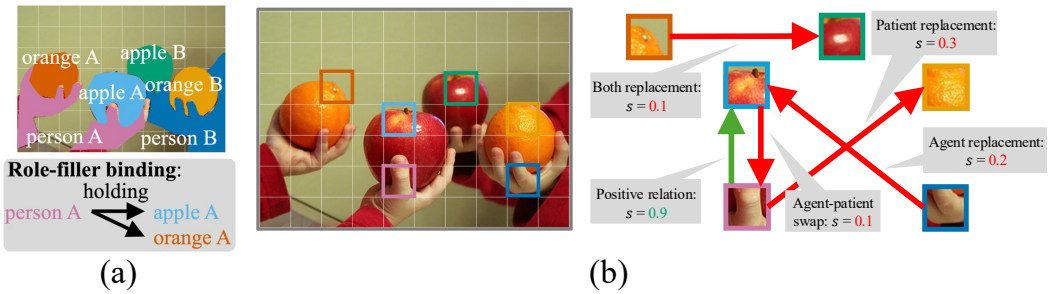


Figure 1: Overview of training the probe for role–filler binding. Given patch features from the agent and patient in an image, the probe distinguishes annotated relation pairs from negatives with altered order or object assignments. (a) Role–filler bindings for "holding," with person A as the agent and apple A and orange A as patients. (b) A positive ordered patch pair (green) and four negative types (red): agent–patient swap, agent replacement, patient replacement, and both replacement. Arrows point from the agent candidate to the patient candidate, and s denotes the probe score.

PSG dataset (Yang et al., 2022) from negatives with altered order or object assignments; the rows of its projection matrix span the subspace at each layer. Our contributions are threefold. (i) The probe reads out agent–patient pairs from held-out PSG images at 83–86% accuracy in all four ViT image encoders we examine, peaking in the middle layers. (ii) As Tables 1 and 2 show, mean-ablating the subspace at a layer selected on PSG lowers accuracy on SugarCrepe swap-object and VALSE actant-swap more than on any other task in all four models, whereas a random subspace of the same dimension has no such effect. (iii) Exchanging subspace components between either the agent or the patient and a third object shifts caption preference toward the caption in which that role is filled by the third object, more often than under the random-subspace control in all four CLIP-style models evaluated (shown in Figure 5). Together, (ii) and (iii) show that the subspace is not merely decodable from the representations but actually used in image–text matching.

2 Related Work

Vision–Language Models. VLMs have achieved strong results on various multimodal tasks. In particular, CLIP, which is pretrained with contrastive learning, is used in a wide range of applications, including zero-/few-shot image classification (Zhai et al., 2022b; Zhou et al., 2022b; Gao et al., 2024), object detection (Minderer et al., 2022; 2023), semantic segmentation (Rao et al., 2022; Zhou et al., 2022a), and text-to-image generation (Zhou et al., 2022c; Ramesh et al., 2022). CLIP-style image encoders also serve as backbones of generative VLMs (Li et al., 2023; Zhu et al., 2024; Deitke et al., 2025). However, CLIP still has difficulty with compositional reasoning, which requires fine-grained understanding of images and text (Yuksekgonul et al., 2023; Kwon et al., 2025), as shown by benchmarks evaluating object presence, attributes, and relations (Zhao et al., 2022; Parcalabescu et al., 2022; Yuksekgonul et al., 2023; Hsieh et al., 2023; Liu et al., 2023; Kamath et al., 2023). To address this, previous approaches train the model with hard negative captions (Yuksekgonul et al., 2023; Zhang et al., 2024; Patel et al., 2024; Oh et al., 2024) or add auxiliary training objectives (Kwon et al., 2025). These approaches modify the model through additional training, whereas we examine the role information that the pretrained image encoder already contains.

CLIP Interpretability. Many studies interpret the image representations of CLIP (Radford et al., 2021) by decomposing them into contributions from the tuples of layers, attention heads, and patches (Gandelsman et al., 2024), individual neurons (Gandelsman et al., 2025), or features learned by sparse autoencoders (Pach et al., 2025; Zaigrajew et al., 2025), identifying what each component encodes. The components identified in these studies mostly correspond to properties of individual objects or of the whole image, such as color, category, location, and count. Koishigarina et al. (2026) show that the unimodal image and text embeddings of CLIP retain attribute–object binding information, even for scenes containing two or more objects, although this information is not aligned

across modalities. However, to our knowledge, these studies do not identify internal representations of agent–patient assignments.

Binding in Transformers. Binding is the problem of associating separately processed features with the entities to which they belong, forming coherent representations (Treisman, 1996; Treisman & Gelade, 1980). Recent studies attribute the failures of VLMs on basic multi-object reasoning tasks, such as counting and visual search (Campbell et al., 2024), and the reversal curse in large language models (LLMs) (Wang & Sun, 2026) to binding failures, motivating studies of binding mechanisms in Transformers. Feng & Steinhardt (2024) show that LLMs bind attributes to entities through low-dimensional subspace of token representations called binding IDs. In generative VLMs, subspaces of token representations called position IDs and spatial IDs are found to associate visual objects with the text tokens that refer to them and explain some spatial relations (Assouel et al., 2026; Kang et al., 2026). In ViTs, a relation that two patches belong to the same object is encoded in a low-dimensional subspace of patch representations (Li et al., 2025). For spatial relations, Yamamoto et al. (2026) show that CLIP-style models can learn left–right relations when trained on synthetic spatial-relation data. Applying singular value decomposition to the query–key interaction matrix $\mathbf{W}_q^\top \mathbf{W}_k$ of ViTs, Pan et al. (2024) show that attention to dissimilar tokens becomes stronger in deeper layers and that some of these interactions can be interpreted as relations between objects. Lepori et al. (2024) show that fine-tuned ViTs represent same-different relations between objects in separable perceptual and relational stages. These studies examine attribute–object associations, position-based correspondences, spatial relations, or symmetric relations, whereas asymmetric role–filler binding, which assigns entities to the argument roles of a predicate (Hummel & Holyoak, 1997), has received little attention in ViTs. In particular, it has not been tested whether spatial and position IDs or the query–key interactions of Pan et al. (2024) account for agent–patient assignments in relations such as "holding," which cannot be determined from the object layout alone.

3 Extracting the Relation-Binding Subspace of ViT

3.1 Probing relation–binding representations

Probe. A ViT splits an image into fixed-size patches, tokenizes them, and lets the tokens interact through scaled dot-product attention (Dosovitskiy et al., 2021). If a ViT performs role–filler binding, its patch representations should encode whether two patches belong to the agent and the patient of a relation, which we call a relation-binding representation. To test this, we follow Li et al. (2025) and adopt probing, which measures the representations of a ViT with a lightweight classifier (Alain & Bengio, 2017). We show that the relation-binding representation can be decoded from patch representations, that is, role–filler binding can be read out from inside the model.

Let $h_\ell \in \mathbb{R}^d$ denote the residual-stream representation of a patch at the output of the ℓ-th Transformer block for $\ell = 1, \ldots, L$, where L is the number of blocks. As the input to the probe, we use h_ℓ normalized to zero mean and unit variance per dimension with fixed statistics. Let $x, y \in \mathbb{R}^d$ be the normalized representations of two patches, given as an ordered pair of an agent candidate and a patient candidate. We define the score of the bilinear probe as

$$s_\ell(x, y) = \mathrm{sigmoid}\big((W_\ell x)^\top V_\ell (W_\ell y) + b_\ell\big)\,, \tag{1}$$

where $b_\ell \in \mathbb{R}$ is a bias, and $s_\ell(x, y)$ evaluates whether the input pair is a positive agent–patient pair. $W_\ell \in \mathbb{R}^{k \times d}$ ($k \ll d$) is a linear map shared by both patches, on which we impose $W_\ell W_\ell^\top = I_k$, where I_k is the $k \times k$ identity matrix. We set $k = 16$ for all models and analyze other values of k in Appendix B.5. To ensure this constraint, W_ℓ is constructed with orthonormal rows from unconstrained parameters (Lezcano-Casado, 2019). Under this constraint, the rows of W_ℓ form an orthonormal basis of a k-dimensional subspace of the normalized representation space, and $W_\ell x$ and $W_\ell y$ are coordinates with respect to the same basis. $V_\ell \in \mathbb{R}^{k \times k}$ represents the interaction between coordinates within the subspace and is normalized to unit Frobenius norm as $V_\ell = \tilde{V}_\ell / \|\tilde{V}_\ell\|_F$ for a learnable matrix $\tilde{V}_\ell \in \mathbb{R}^{k \times k}$. No symmetry is imposed on $\tilde{V}_\ell$, and no additional scale factor is placed on the score. Since $s_\ell(x, y)$ and $s_\ell(y, x)$ can therefore take different values, the probe distinguishes the order of the agent and the patient.

Data preparation. We train a probe that classifies whether a pair of patches is in an agent–patient relation on the Panoptic Scene Graph (PSG) dataset (Yang et al., 2022). PSG provides pixel-level

masks of objects and background regions in COCO images, together with triplets of (subject, predicate, object), which describe relations between the regions; we treat the subject as the agent and the object as the patient. Among the predicates in PSG, we use interaction predicates (e.g., holding, riding, eating, and looking at), whereas we exclude spatial predicates that can be determined from patch positions alone (e.g., on, in, and beside) and symmetric predicates whose truth does not change when the agent and the patient are exchanged (e.g., kissing and talking to). We randomly hold out 10% of the PSG training images that contain such triplets as a split for layer selection (referred to as PSG val) and use the rest for training. The mean and standard deviation used to normalize the patch representations h_ℓ are estimated once from randomly chosen training images before training.

From the object masks and relation annotations, we construct ordered patch pairs (a, p) of an agent and a patient as positives (as shown in Figure 1(a)). For each positive, we construct one negative of each of four types, as shown in Figure 1(b). (i) Agent–patient swap (Swap): the pair (p, a) in which the order of the agent a and the patient p is exchanged. (ii) Agent replacement (Agent-Rep): the pair (c, p) in which the agent is replaced with an object c that has no relation with the patient p. (iii) Patient replacement (Patient-Rep): the pair (a, c) in which the patient is replaced with an object c that has no relation with the agent a. (iv) Both replacement (Both-Rep): the pair (c, c') of two objects other than a and p that have no relation with each other. c and c' are chosen at random among the eligible objects in the image; when no eligible object exists, the negative of that type is not constructed. The probe is trained on these positives and negatives with the binary cross-entropy loss. See Appendix A for detailed setup.

Models. We employ four CLIP-style models, namely CLIP ViT-B/16-224px (CLIP ViT-B; $d = 768$, $L = 12$), CLIP ViT-L/14-224px (CLIP ViT-L; $d = 1024$, $L = 24$) (Radford et al., 2021), OpenCLIP ViT-L/14-224px (OpenCLIP ViT-L; $d = 1024$, $L = 24$) (Ilharco et al., 2021), and SigLIP SoViT-400m/14-384px (SigLIP; $d = 1152$, $L = 27$) (Zhai et al., 2023).

Results: Probe Accuracy. We trained a probe independently at each layer of the four CLIP-style models and evaluated it on PSG val. Figure 2 shows the probe accuracy over five random seeds and over the four types of negatives. In all four models, probe accuracy increases almost monotonically in the first half of the layers, reaches its maximum at a relative depth of 0.50–0.70, and decreases slightly toward the last layer. The curves are almost the same for CLIP ViT-L and OpenCLIP ViT-L, which differ in pretraining data. The maximum accuracy falls within a narrow range of 83–86% across the four models, whereas the chance level is 50%, indicating that the relation is decodable to a similar extent in all of them. Among them, SigLIP reaches the highest maximum (85.9%) and attains high accuracy from early layers. Appendix B.1 reports probe accuracy for each negative type.

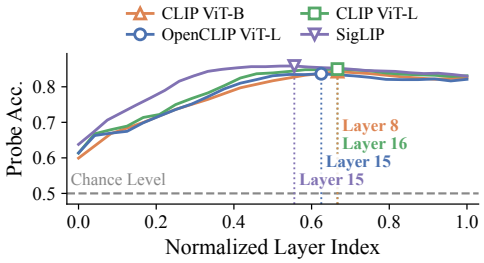


Figure 2: **Probe accuracy across layers.**

Results: Visualization of Probe Scores. Next, we visualize the probe scores, that is, the response of each patch given its partner object in Figure 3. We focus on the layer ℓ at which the probe accuracy is highest, shown in Figure. 2. Let $\bar{x}$ and $\bar{y}$ be the mean normalized representations within the masks of the agent and the patient, respectively, and let z_i be the normalized representation of patch i; the probe score of patch i as an agent candidate is $s_\ell(z_i, \bar{y})$, and that as a patient candidate is $s_\ell(\bar{x}, z_i)$. The former scans candidate positions of the agent given the patient, and the latter scans candidate positions of the patient given the agent. Figure 3(a) shows relation pairs with different kinds of agents and patients: a person acting on a non-human object, a person acting on another person, an animal as the agent, and an inanimate object as the agent. In each case, the probe assigns high scores to the region of the partner object, indicating that the ordered agent–patient association is present in the patch representations. Figure 3(b) shows images that contain two relation pairs. When the given object is switched from one pair to the other, the high-scoring region moves to the partner of the newly given object, indicating that the probe reads out not a property of an individual object but its pairing with a partner in a relation.

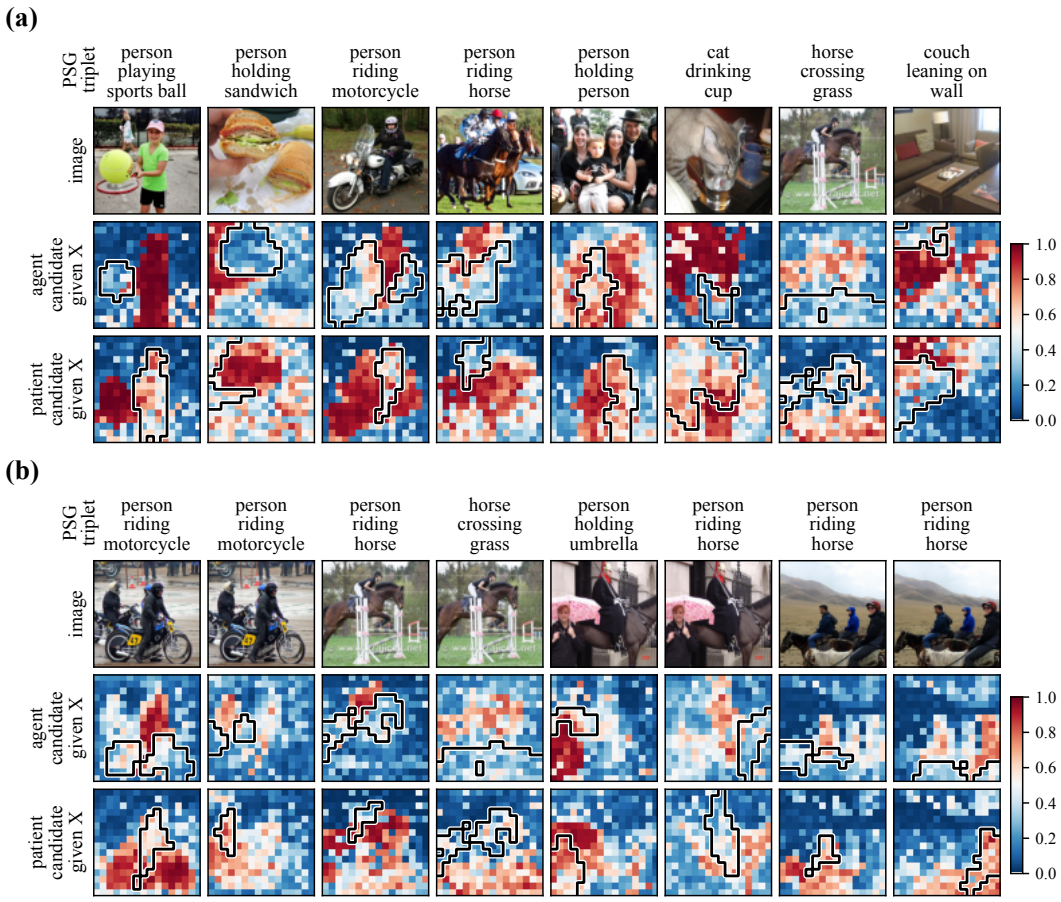


Figure 3: **Visualization of probe scores on PSG val images.** (a) Examples of relation pairs. (b) Examples of different relation pairs in the same image, where each pair of adjacent columns forms one set. The outlined mask in each heatmap indicates the given agent or patient. Each column shows, from top to bottom, the input image, the probe score of each patch as an agent candidate given the mean representation of the patient, and the probe score of each patch as a patient candidate given the mean representation of the agent.

Figure 4 shows, for the first set in Figure 3(b), how the probe scores of the two riders and their motorcycles change across layers. In early layers ($\ell \leq 12$), given either motorcycle, both riders receive similar scores as agent candidates, and given either rider, both motorcycles receive similar scores as patient candidates; the probe does not tell which rider belongs to which motorcycle. In the middle layers ($\ell = 16$), the probe separates the two pairs: given a motorcycle, the region of its own rider receives a higher score than that of the other rider. In deep layers ($\ell \geq 20$), this separation becomes again weaker. Visualizations for the other models are provided in Appendix B.2.

3.2 Subspace Ablation

Intervention. We define the *relation-binding subspace* as the k-dimensional subspace $\mathcal{S}_\ell = \mathrm{range}(W_\ell^\top) \subset \mathbb{R}^d$ spanned by the rows of W_ℓ learned by the probe from the ordered patch pairs. To test whether this subspace is used in image–text matching, we replace the normalized representation $x_{\ell,i}$ of patch i at layer ℓ with

$$x'_{\ell,i} = x_{\ell,i} - W_\ell^\top \left(W_\ell x_{\ell,i} - \bar{c}_\ell\right), \tag{2}$$

where $\bar{c}_\ell$ is the mean of $W_\ell x_{\ell,i}$ over all patches of the image before intervention; that is, the subspace component of every patch is replaced with its mean over the image (referred to as mean-ablation).

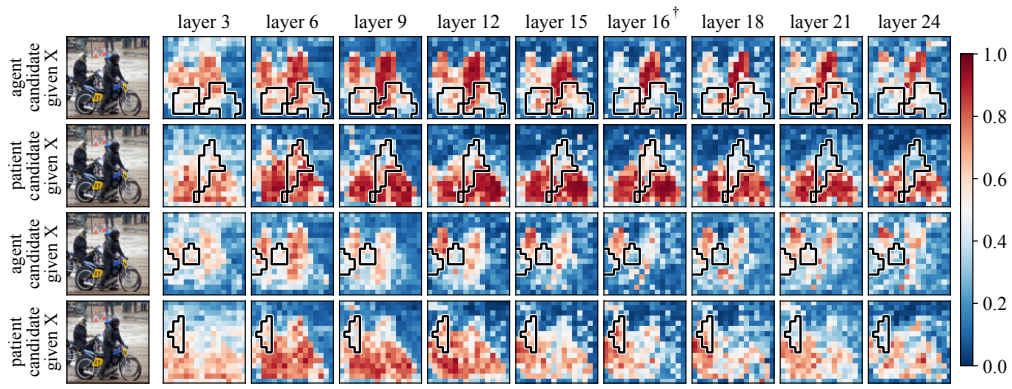


Figure 4: **Probe scores to relation pairs across layers.** The top two rows show a rider–motorcycle pair and the bottom two rows show the other, each in the order of agent and patient. Using the probe trained independently at each layer, the agent rows show the probe score of each patch as an agent candidate given the motorcycle, and the patient rows show that as a patient candidate given the rider. The columns show every third layer plus layer 16 at which the probe accuracy on PSG val is highest for CLIP ViT-L.

After denormalizing $x'_{\ell,i}$, we replace the residual-stream representation of patch i with the resulting vector and continue the forward pass. The intervention is applied to all patch tokens except for the CLS token of CLIP. As a control, we apply the same mean-ablation to a subspace spanned by a random orthonormal basis of the same dimension k in the normalized representation space. The reported results are averaged over five runs, each using the probe trained with a different seed and an independently sampled random basis.

We evaluate on SugarCrepe (Hsieh et al., 2023) and VALSE (Parcalabescu et al., 2022), measuring the accuracy of selecting the correct caption from two candidates for an image. SugarCrepe consists of three types of tasks: replacing objects, attributes, or relations in captions (replace), swapping objects or attributes (swap), and adding objects or attributes (add). VALSE evaluates the existence of objects, plurality, counting, spatial relations, actions, coreference, and noun replacement (Foil-it!). SugarCrepe swap-object exchanges the phrases of two objects, and VALSE actant-swap exchanges the objects assigned to the two roles of an action. In both tasks, the correct and incorrect captions consist of the same words and differ only in which object fills which role, as in "a person chasing a dog" and "a dog chasing a person," so that they cannot be distinguished by object or attribute information alone. For each model, we apply mean-ablation at the layer at which the probe accuracy on PSG val is highest (shown in Figure 2): $\ell = 8$ for CLIP ViT-B, $\ell = 16$ for CLIP ViT-L, $\ell = 15$ for OpenCLIP ViT-L, and $\ell = 15$ for SigLIP.

Results. Table 1 shows the results on SugarCrepe. In all four models, mean-ablation causes the largest accuracy drop on swap-object, exceeding the drop under the same intervention on a random subspace. Table 2 shows the results on VALSE. In all four models, actant-swap shows the largest accuracy drop among all VALSE tasks under mean-ablation, exceeding the drop under random-subspace ablation. Zero-ablation gives the same pattern (Tables 5 and 6 in Appendix B.4). This consistent pattern supports the view that the learned subspace contributes to identifying which object fills each role.

3.3 Subspace Exchange

The mean-ablation in Section 3.2 shows that the relation-binding subspace is needed to tell apart captions in which the roles are swapped, but not what information its components carry. If these components encode which object is bound to which role of a relation, exchanging them between two objects should move the role assignment, while the remaining components, which include the information on what each object is, stay in place. This follows the logic by which Feng & Steinhardt (2024) test binding IDs in LLMs: they switch the binding ID vectors of two entities and find that

Table 1: SugarCrepe accuracy under mean-ablation in four vision–language models.

		Replace			Swap		Add		
	Methods	Obj	Att	Rel	Obj	Att	Obj	Att	Avg
CLIP ViT-B	base	93.04	81.60	66.79	60.41	65.62	78.42	66.33	76.95
	subspace ablation	90.38	78.45	65.58	52.41	60.63	73.41	65.75	73.68
	gap	−2.66	−3.15	−1.21	**−8.00**	−4.98	−5.00	−0.58	−3.27
	gap (random)	−0.05	−0.10	−0.11	−0.24	−0.42	−0.36	+0.20	−0.17
CLIP ViT-L	base	94.25	78.30	65.15	60.41	62.31	78.18	71.53	76.69
	subspace ablation	92.76	79.62	64.42	55.67	60.48	77.15	70.58	75.68
	gap	−1.49	+1.32	−0.73	**−4.73**	−1.83	−1.03	−0.95	−1.01
	gap (random)	−0.18	+0.66	+0.17	+0.33	−0.36	+0.30	−0.17	+0.11
OpenCLIP ViT-L	base	96.55	84.52	70.48	61.22	69.07	90.01	79.77	83.48
	subspace ablation	95.70	82.66	68.63	58.86	68.32	89.64	79.45	82.48
	gap	−0.85	−1.85	−1.85	**−2.37**	−0.75	−0.37	−0.32	−1.00
	gap (random)	−0.29	−0.66	+0.03	−0.57	−0.03	+0.35	+0.03	−0.05
SigLIP	base	97.46	87.18	70.91	69.80	77.78	91.80	83.38	85.91
	subspace ablation	97.20	86.24	70.21	66.20	77.33	91.64	83.73	85.46
	gap	−0.25	−0.94	−0.70	**−3.59**	−0.45	−0.16	+0.35	−0.46
	gap (random)	−0.07	−0.15	−0.40	−0.82	−0.15	−0.08	−0.32	−0.20

"base" denotes no intervention. Both interventions replace subspace components with their mean, using either the learned relation-binding subspace or a subspace spanned by a random orthonormal basis of the same dimension (random). "gap" denotes the difference from base; bold and underline mark the largest and second-largest drops. "Avg" is the accuracy over all examples. All values are averaged over five runs; standard deviations over the runs are given in Table 5.

the model then retrieves the attribute of the other entity. Analogously, we exchange the subspace components between two objects and observe which object the model then describes in the relation.

Intervention. We exchange the subspace components between two objects in the same image. For each triplet of an agent a, a predicate r, and a patient p in PSG val, we choose an object c from the same image and exchange the subspace components between a and c in the agent-side exchange, or between p and c in the patient-side exchange. Let $\bar{c}_\ell(S) = \frac{1}{|S|}\sum_{i\in S} W_\ell x_{\ell,i}$ denote the mean of the subspace coordinates over the patches of an object S. The agent-side exchange applies Equation 2 with $\bar{c}_\ell$ replaced by $\bar{c}_\ell(c)$ for the patches of a and by $\bar{c}_\ell(a)$ for the patches of c; the patient-side exchange does the same with a replaced by p. The patches of the object that is not exchanged (p in the agent-side exchange and a in the patient-side exchange) and the components orthogonal to $\mathcal{S}_\ell$ are left unchanged. The subspace components of the patches outside $a \cup p \cup c$ are also left unchanged. We intervene independently at each layer, $\ell = 1, \ldots, L-1$ for the three CLIP models and $\ell = 1, \ldots, L$ for SigLIP. The last layer of the CLIP models is excluded because an intervention of the last layer does not reach the CLS token.

The object c is chosen at random among the objects in the image that satisfy all of the following conditions. (i) c belongs to a category different from that of the object it replaces, so that the two captions compared below are not identical. (ii) c has no annotated relation with the object that is not exchanged, so that the caption after the exchange does not correctly describe the original image. (iii) The category of c fills the same role in at least one triplet of PSG, which reduces the items in which the original caption can be chosen from linguistic plausibility alone, without referring to the roles in the image. These conditions yield 816 items in each exchange for which a, p, and c each cover at least one patch in the patch grids of all four models. The choice of c depends neither on the probe seed nor on the layer, and the same set of items is used in all runs.

We evaluate the preference between two captions by the cosine similarity between the image embedding and each caption embedding. Let $e_a =$ "a photo of a $\{a\}$ $\{r\}$ a $\{p\}$" be the caption describing the original relation. It is compared with $e_c =$ "a photo of a $\{c\}$ $\{r\}$ a $\{p\}$" in the agent-side exchange and with $e_c =$ "a photo of a $\{a\}$ $\{r\}$ a $\{c\}$" in the patient-side exchange. As a no-exchange control, the subspace components of the patches of a and c are replaced with their own within-object

Table 2: VALSE accuracy under mean-ablation in four vision–language models.

		Existence	Plurality	Counting	Sp.rel.	Action		Coreference		Foil-it!	Avg
	Methods					repl.	actant swap	standard	clean		
CLIP ViT-B	base	65.54	55.58	65.15	60.75	82.56	72.60	47.60	53.85	90.56	67.54
	subspace ablation	62.02	55.39	63.12	58.02	79.57	63.92	48.36	52.88	86.87	64.74
	gap	−3.52	−0.19	−2.03	−2.73	−2.99	**−8.68**	+0.76	−0.96	−3.69	−2.80
	gap (random)	+0.44	+0.40	+0.68	−0.49	−0.86	−1.77	−0.17	−1.92	−0.51	−0.14
CLIP ViT-L	base	67.13	58.75	66.49	57.94	83.49	68.81	53.67	56.73	92.26	68.64
	subspace ablation	64.16	57.41	64.92	58.36	82.07	65.61	54.35	54.42	90.92	67.18
	gap	−2.97	−1.34	−1.57	+0.41	−1.42	**−3.20**	+0.68	−2.31	−1.34	−1.46
	gap (random)	−0.08	−0.59	−0.24	+0.71	−0.03	+0.34	+0.34	+1.92	+0.15	+0.02
OpenCLIP ViT-L	base	75.05	68.51	62.87	62.06	84.72	61.43	51.55	46.15	92.15	68.22
	subspace ablation	74.22	66.82	62.03	60.26	84.26	56.61	50.96	45.00	92.05	66.86
	gap	−0.83	−1.69	−0.85	−1.79	−0.46	**−4.83**	−0.59	−1.15	−0.11	−1.35
	gap (random)	+0.08	−0.54	−0.37	+0.07	−0.59	−0.40	+0.08	−1.92	+0.23	−0.25
SigLIP	base	77.62	69.92	68.52	63.55	90.12	57.85	59.04	54.81	93.11	71.38
	subspace ablation	78.57	70.22	68.82	61.72	89.85	55.64	57.85	55.38	92.53	70.98
	gap	+0.95	+0.31	+0.29	−1.83	−0.28	**−2.21**	−1.19	+0.58	−0.57	−0.41
	gap (random)	−0.28	+0.00	−0.24	−0.30	+0.06	+0.46	−0.37	−0.38	−0.13	−0.11

"base" denotes no intervention. Both interventions replace subspace components with their mean, using either the learned relation-binding subspace or a subspace spanned by a random orthonormal basis of the same dimension (random). "gap" denotes the difference from base; bold and underline mark the largest and second-largest drops. "Avg" is the accuracy over all valid examples. Under "Action", "repl." denotes action replacement; "Sp.rel." denotes spatial relations. All values are averaged over five runs; standard deviations over the runs are given in Table 6.

means $\bar{c}_\ell(a)$ and $\bar{c}_\ell(c)$, with a read as p in the patient-side exchange. For layer ℓ, let n_ℓ be the number of items for which the no-exchange control prefers e_a, and m_ℓ the number of those items for which the model prefers e_c after the exchange; the preference-flip rate is defined as m_ℓ/n_ℓ. As a further control, we apply the same exchange to a subspace spanned by a random orthonormal basis of the same dimension k. The results are averaged over five runs, each using the probe trained with a different seed and an independently sampled random basis.

Results. Figure 5 shows preference-flip rates across layers, and Table 3 reports the item counts at the peak layer for each model and exchange type. For both exchange types, the learned subspace yields higher flip rates than the random-subspace control at every layer in all four models. The effect also differs between the two roles: agent-side exchanges yield higher flip rates than patient-side exchanges at almost every layer in each model. This asymmetry may reflect the influence of object associations acquired during pretraining, which may remain in the unchanged components and support the original caption during patient-side exchanges.

In the three CLIP models, flip rates peak in the middle layers. For example, both exchange types in CLIP ViT-L peak at layer $\ell = 14$, with rates of 49.33% for agent-side exchanges and 14.58% for patient-side exchanges, while the random-subspace control remains at or below 1.1% across all layers. SigLIP, in contrast, shows low flip rates with little variation across layers. These rates nevertheless remain well above those of the random-subspace control.

The layer-wise effects may partly reflect how patch representations reach the image embedding. In the three CLIP models, patch information is aggregated into the CLS token for the final readout. If role information has already been incorporated into CLS, interventions on patch representations in later layers may have less effect. SigLIP instead pools the final patch tokens with a multihead attention pooling (MAP) head (Zhai et al., 2022a; 2023), without requiring this transfer to a CLS token. This helps explain why its preference-flip rates vary relatively little across layers.

The exchange experiment complements ablation by testing whether the subspace affects which object the model identifies as filling a relational role. Mean-ablation reduces accuracy in distinguish-

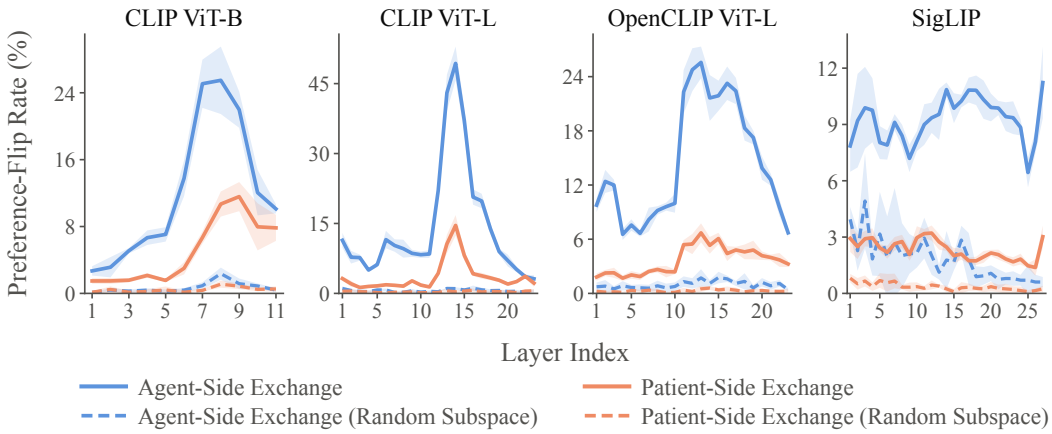


Figure 5: **Preference-flip rate under the exchange of subspace components.** Among items for which the no-exchange control prefers the original caption, the Preference-flip rate is the fraction that switch to the caption naming the third object. Lines show means over five runs. Shaded bands indicate ± 1 standard deviation.

Table 3: **Caption preferences before and after subspace exchange.** For each model and exchange, we report the layer with the highest flip rate under subspace exchange. Parentheses show percentages of the 816 items. m counts switches to e_c among the n items preferring e_a under the no-exchange control. "Random" uses the same layer and its own no-exchange control. Intervention results are means over five runs, with standard deviations in subscripts.

			Prefer e_a			Preference-flip rate (%)	
	Model	Layer	No intervention	No-exchange control (n)	Flipped (m)	subspace exchange	Random
Agent-side exchange	CLIP ViT-B	8	657 (80.51%)	$622.2_{\pm 4.8}$ (76.25%)	$158.6_{\pm 25.1}$	$25.49_{\pm 4.03}$	$2.36_{\pm 0.73}$
	CLIP ViT-L	14	567 (69.49%)	$536.2_{\pm 8.4}$ (65.71%)	$264.4_{\pm 17.7}$	$49.33_{\pm 3.61}$	$1.03_{\pm 0.29}$
	OpenCLIP ViT-L	13	524 (64.22%)	$499.6_{\pm 2.1}$ (61.23%)	$127.8_{\pm 9.0}$	$25.58_{\pm 1.75}$	$1.72_{\pm 0.96}$
	SigLIP	27	430 (52.70%)	$414.8_{\pm 1.5}$ (50.83%)	$46.6_{\pm 8.0}$	$11.23_{\pm 1.91}$	$0.60_{\pm 0.27}$
Patient-side exchange	CLIP ViT-B	9	672 (82.35%)	$669.8_{\pm 4.5}$ (82.08%)	$77.6_{\pm 12.0}$	$11.58_{\pm 1.74}$	$0.86_{\pm 0.35}$
	CLIP ViT-L	14	680 (83.33%)	$679.2_{\pm 1.9}$ (83.24%)	$99.0_{\pm 14.7}$	$14.58_{\pm 2.21}$	$0.35_{\pm 0.20}$
	OpenCLIP ViT-L	13	676 (82.84%)	$668.4_{\pm 1.8}$ (81.91%)	$44.8_{\pm 7.0}$	$6.70_{\pm 1.05}$	$0.50_{\pm 0.34}$
	SigLIP	12	667 (81.74%)	$666.0_{\pm 0.7}$ (81.62%)	$21.4_{\pm 2.7}$	$3.21_{\pm 0.40}$	$0.45_{\pm 0.24}$

ing captions with swapped roles, whereas exchanging components can shift preference toward the caption naming the third object in that role. Together, these results support a contribution of the subspace to binding objects to roles in image–text matching, beyond what a probe can read out.

4 Conclusion

By probing patch representations in the CLIP-style models studied here, we identified a low-dimensional *relation-binding subspace* that encodes ordered agent–patient relations. Ablating this subspace produces the largest mean accuracy drops on role-swap tasks within each evaluated benchmark. Exchanging its components between either the agent or the patient and a third object can shift caption preference toward the caption assigning that role to the third object. Together, these findings support a contribution of the subspace to binding objects to roles in image–text matching. Role-binding information can therefore be present and used even when a model does not reliably distinguish captions with swapped roles.

AI USE STATEMENT

We used OpenAI Codex to draft and edit the manuscript, translate text, search for related work, and assist with interpreting experimental results. We also used Codex to write, modify, and debug experimental code and code for data preprocessing and formatting, and to generate plots from experimental results supplied by the authors. We did not use generative AI to design methods or experiments, generate synthetic training or evaluation data, derive formulas, formulate mathematical claims, or write proofs. We read and reviewed the AI-assisted text and code, checked descriptions and citations of related work against the original papers, and assessed interpretations against the experimental results and prior work. We checked plotted values, axes, and legends against the underlying experimental results.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=3RQ863cRbx.

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 Webb. Understanding the limits of vision language models through the lens of the binding problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=Q5RYn6jagC.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Deitke_Molmo_and_PixMo_Open_Weights_and_Open_Data_for_State-of-the-Art_CVPR_2025_paper.html.

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.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=zb3b6oKO77.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=5Ca9sSzuDp.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting the second-order effects of neurons in CLIP. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=GPDcvoFGOL.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x. URL https://link.springer.com/article/10.1007/s11263-023-01891-x.

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

Zenzi M. Griffin and Kathryn Bock. What the eyes say about speaking. *Psychological Science*, 11 (4):274–279, 07 2000. doi: 10.1111/1467-9280.00255.

Alon Hafri, Anna Papafragou, and John C Trueswell. Getting the gist of events: recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, 142 (3):880–905, 2013. doi: 10.1037/a0030045.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/63461de0b4cb760fc498e85b18a7fe81-Abstract-Datasets_and_Benchmarks.html.

John E. Hummel and Keith J. Holyoak. Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3):427–466, 1997. URL https://reasoninglab.psych.ucla.edu/wp-content/uploads/sites/273/2021/04/Hummel-Holyoak.1997.pdf.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=RN5KLywTll.

Raphi Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=2zXRGiorSu.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=DldwXCCP25.

Jihoon Kwon, Kyle Min, and Jy yong Sohn. Enhancing compositional reasoning in CLIP via reconstruction and alignment of text descriptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=6uKIm4bfEe.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=8puv3c9CPg.

Mario Lezcano-Casado. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1b33d16fc562464579b7199ca3114982-Abstract.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp.

19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

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. URL https://openreview.net/forum?id=5BS6gBb4yP.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37/.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *European Conference on Computer Vision (ECCV)*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/7529_ECCV_2022_paper.php.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=mQPNcBWjGc.

Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. Preserving multimodal capabilities of pre-trained VLMs for improving vision-linguistic compositionality. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 19060–19076, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1062. URL https://aclanthology.org/2024.emnlp-main.1062/.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=DaNnkQJSQf.

Xu Pan, Aaron Philip, Ziqian Xie, and Odelia Schwartz. Dissecting query-key interaction in vision transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=dIktpSgK4F.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 8253–8280. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.567. URL http://dx.doi.org/10.18653/v1/2022.acl-long.567.

Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=ZfRGRK5Kxl.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *ArXiv*, abs/2204.06125, 2022. URL https://arxiv.org/abs/2204.06125.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18061–18070, 2022. doi: 10.1109/CVPR52688.2022.01755. URL https://doi.org/10.1109/CVPR52688.2022.01755.

Anne M. Treisman. The binding problem. *Current opinion in neurobiology*, 6(2):171–178, 1996. doi: 10.1016/S0959-4388(96)80070-5. URL https://doi.org/10.1016/S0959-4388(96)80070-5.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. ISSN 0010-0285. doi: 10.1016/0010-0285(80)90005-5. URL https://www.sciencedirect.com/science/article/pii/0010028580900055.

Boshi Wang and Huan Sun. Is the reversal curse a binding problem? uncovering limitations of transformers from a basic generalization failure. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=iiLLQhhrtb.

Takaki Yamamoto, Chihiro Noguchi, and Toshihiro Tanizawa. Left–right symmetry breaking in CLIP-style vision-language models trained on synthetic spatial-relation data. In *International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=9dKTHof9rW.

Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision (ECCV)*, 2022.

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? In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2025. URL https://openreview.net/forum?id=5MQQsenQBm.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling Vision Transformers . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1204–1213, Los Alamitos, CA, USA, June 2022a. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01179. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01179.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18102–18112, 2022b. doi: 10.1109/CVPR52688.2022.01759. URL https://doi.org/10.1109/CVPR52688.2022.01759.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13774–13784, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_Contrasting_Intra-Modal_and_Ranking_Cross-Modal_Hard_Negatives_to_Enhance_Visio-Linguistic_CVPR_2024_paper.html.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. VL-CheckList: Evaluating pre-trained vision-language models with objects, attributes and relations. *CoRR*, abs/2207.00221, 2022. URL https://doi.org/10.48550/arXiv.2207.00221.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision (ECCV)*, pp. 696–712, Berlin, Heidelberg, 2022a. Springer-Verlag. ISBN 978-3-031-19814-4. doi: 10.1007/978-3-031-19815-1_40. URL https://doi.org/10.1007/978-3-031-19815-1_40.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b. doi: 10.1007/s11263-022-01653-1. URL https://link.springer.com/article/10.1007/s11263-022-01653-1.

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17886–17896, 2022c. doi: 10.1109/CVPR52688.2022.01738. URL https://doi.org/10.1109/CVPR52688.2022.01738.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1tZbq88f27.

A Experimental Setup

A.1 Dataset Details

We use relation annotations from PSG (Yang et al., 2022), which annotates 48,749 COCO images with panoptic masks of objects and background regions and with (subject, predicate, object) triplets over 56 predicate classes. Among the 56 predicates, we use the 30 interaction predicates listed in Table 4 and exclude the other 26: spatial and positional predicates that can be determined from patch positions alone (e.g., on, in, and beside) and symmetric predicates whose truth does not change when the agent and the patient are exchanged (kissing, playing with, talking to, and touching). Of the PSG training images, 16,690 contain at least one triplet with these predicates. Each patch of the model input is assigned to the segment that covers the largest part of it, and we retain only triplets whose agent and patient are both present in the patches; images without such a triplet are discarded. We apply this filtering using the 14×14 patch grid of CLIP ViT-B, which has the fewest patches among the evaluated models. We randomly hold out 10% of the retained PSG training images as PSG val. The resulting training set contains 8,876 images and 12,622 interaction triplets, and PSG val contains 986 images and 1,406 triplets.

Table 4: **The interaction predicates of PSG used in this work and their training triplet counts.**

Interaction predicate	Training triplets	Interaction predicate	Training triplets
holding	3,980	drinking	65
looking at	2,836	jumping over	60
riding	1,170	feeding	50
wearing	776	pushing	41
carrying	673	guiding	51
eating	722	catching	36
playing	618	throwing	22
leaning on	562	kicking	21
swinging	216	about to hit	30
pulling	161	chasing	12
driving	123	cooking	13
jumping from	114	cleaning	10
crossing	77	climbing	7
biting	87	opening	6
slicing	78	picking	5

A.2 Training Setup

The image encoders are frozen and run in half precision. A separate probe is trained on the patch representations at every layer, including the patch embeddings ($\ell = 0$). Each mini-batch contains 24 training images. For each image, we sample 16 retained triplets with replacement and randomly select one ordered agent–patient patch pair from each. For each positive, one negative of each type is constructed from the same image. Swap uses the same two patches in the reversed order. For Agent-Rep, the candidates for c are the objects other than a and p that have no annotated relation with p. Patient-Rep is constructed in the same way with the roles of a and p exchanged. For Both-Rep, (c, c') is chosen at random among all ordered pairs of objects other than a and p in the image that have no annotated relation with each other. For all four models, the training split gives 142,016 positive pairs and 142,016 Swap, 132,714 Agent-Rep, 128,186 Patient-Rep, and 136,736 Both-Rep negatives, and PSG val gives 15,776 positive pairs and 15,776, 14,923, 14,235, and 15,312 negatives of the four types, respectively. The loss is the binary cross-entropy, weighted so that the positives receive half of the total weight and each of the four negative types one eighth, divided equally among the valid pairs of each type in the mini-batch. We use Adam with a learning rate of 10^{-3}, which is multiplied by 0.2 after 8 epochs, and train for 16 epochs, that is, 5,904, 6,352, and 8,000 iterations for the 14×14, 16×16, and 27×27 patch grids. The mean and standard deviation for normalization are computed per dimension over the patch tokens of the first 40 mini-batches (960 images) before training. The five runs of each model use different random seeds, which change the initialization of the probes, the order of the training images, and the images used for the normalization statistics.

B Additional Experiments

B.1 Probe accuracy

Figure 6 shows the probe accuracy at each layer for each type of negative.

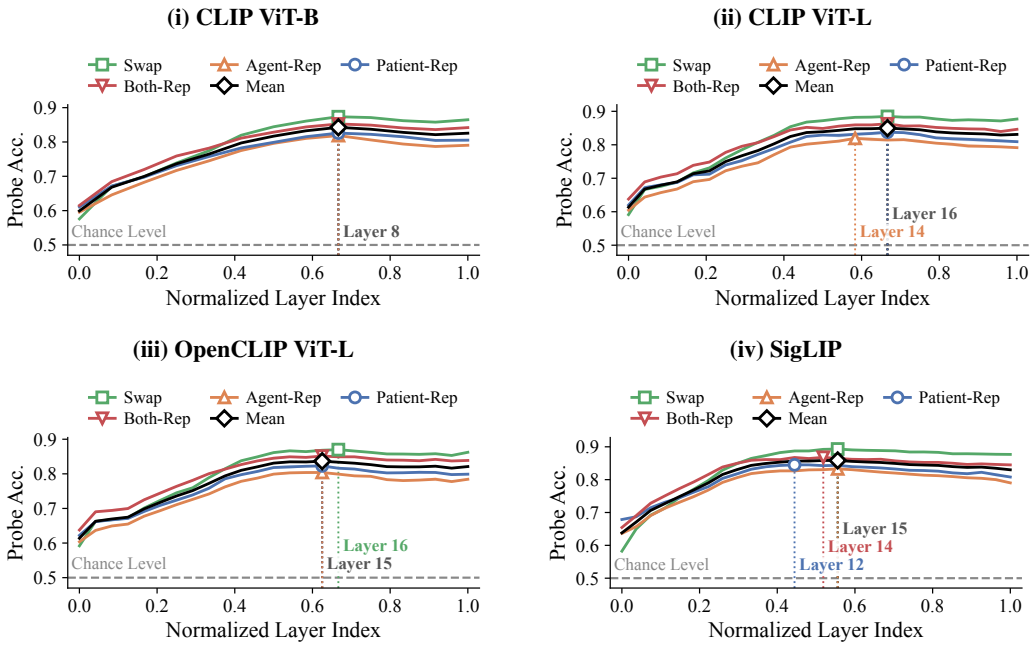


Figure 6: **Probe accuracy on PSG val across layers for each negative type.** Lines show the accuracy for each negative type and their mean, averaged over five seeds; bands show ± one standard deviation.

B.2 VISUALIZATION OF PROBE PREDICTIONS

(a)

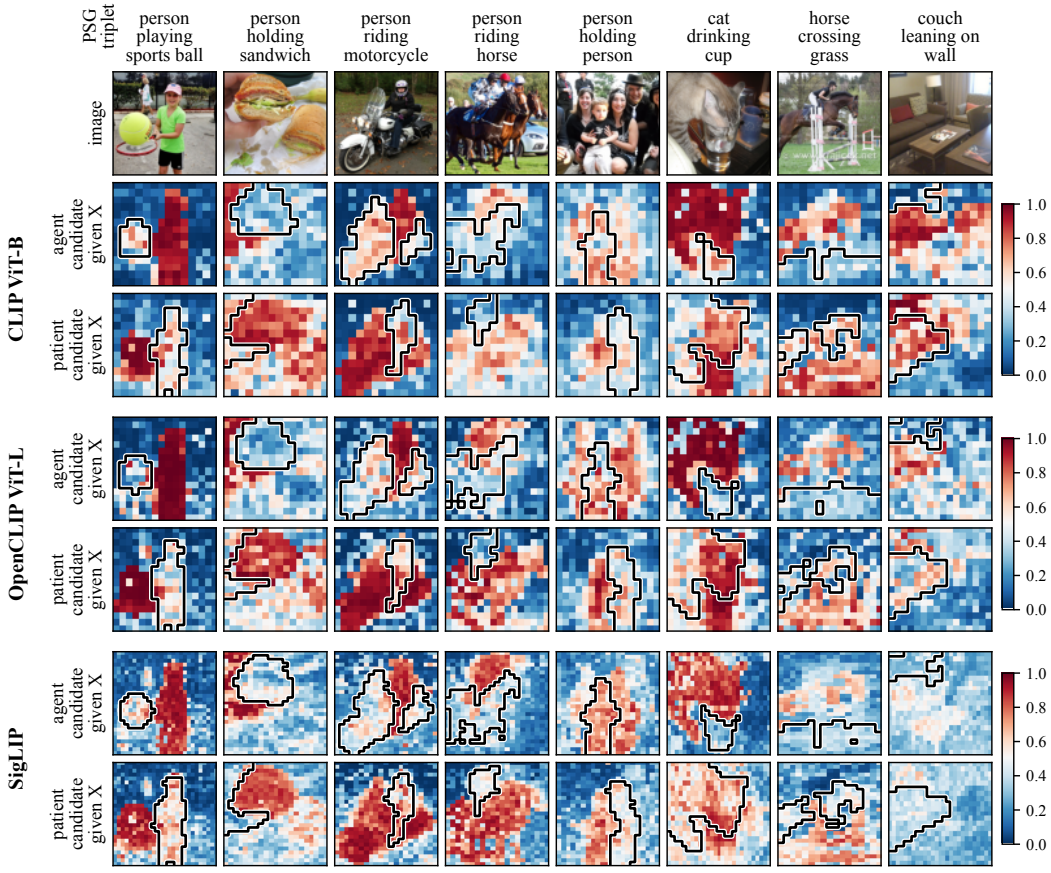


Figure 7: **Visualization of probe predictions on PSG val images for CLIP ViT-B, OpenCLIP ViT-L, and SigLIP.** (a) Examples of relation pairs. Panel (b) and the full description continue on the next page.

(b)

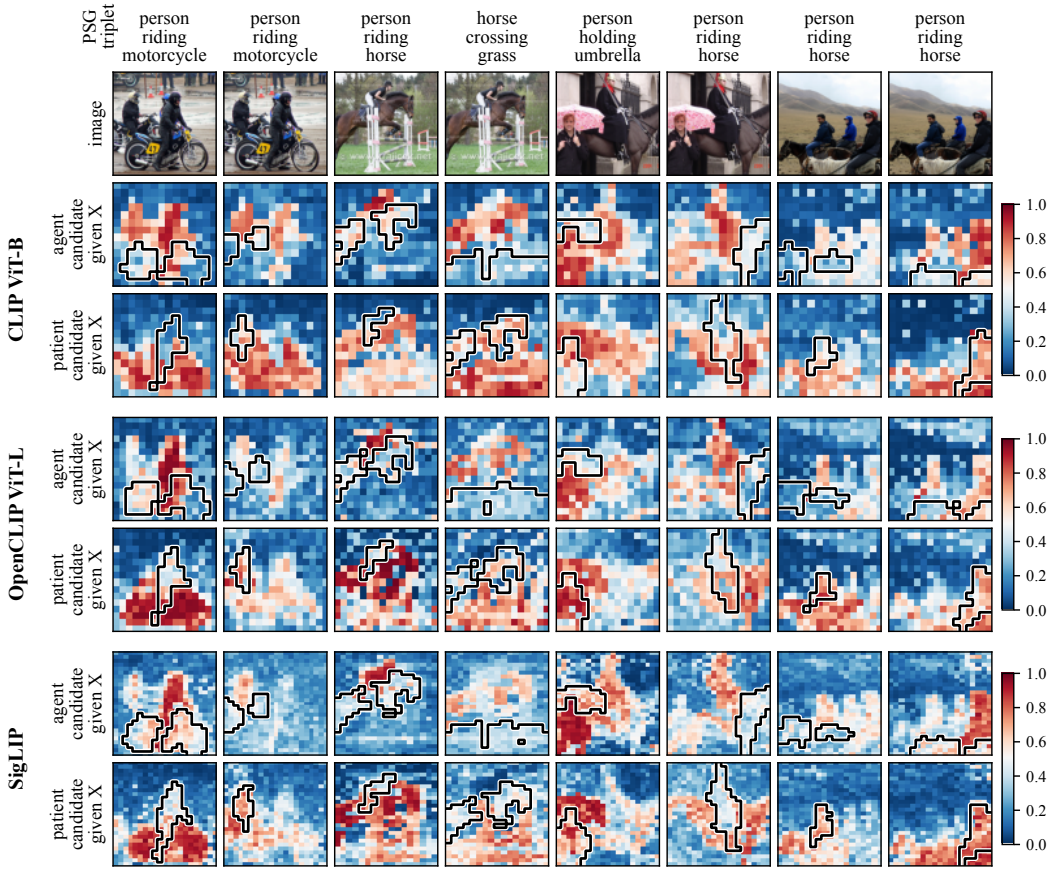
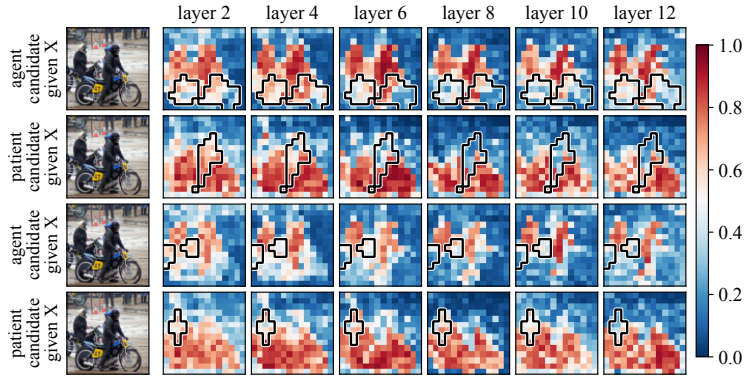


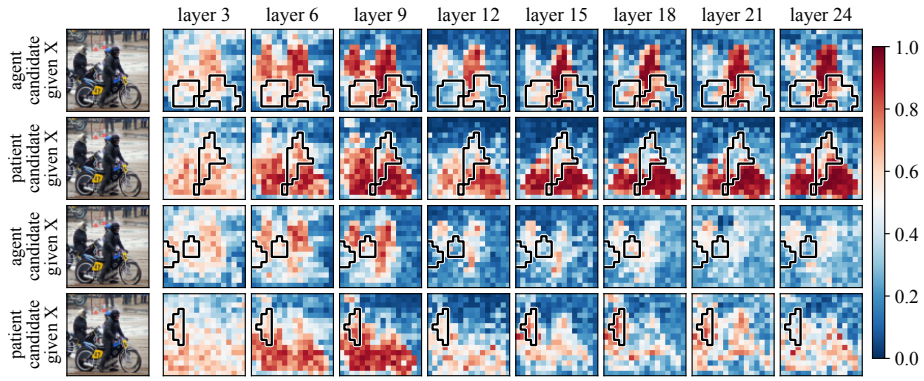
Figure 7: **Visualization of probe predictions on PSG val images for CLIP ViT-B, OpenCLIP ViT-L, and SigLIP (continued).** (a) Examples of relation pairs. (b) Examples of different relation pairs in the same image, where each pair of adjacent columns forms one set.

B.3 PROBE RESPONSES ACROSS LAYERS

(i) CLIP ViT-B



(ii) OpenCLIP ViT-L



(iii) SigLIP

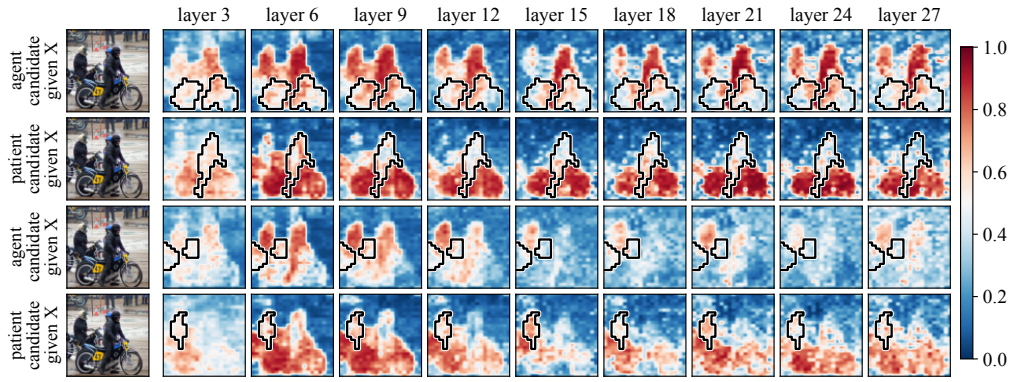


Figure 8: **Probe responses to relation pairs across layers for CLIP ViT-B, OpenCLIP ViT-L, and SigLIP.**

B.4 MEAN- AND ZERO-ABLATION WITH STANDARD DEVIATIONS OVER RUNS

Tables 5 and 6 list the gaps of Tables 1 and 2 together with their standard deviations over the five runs, and the corresponding gaps under zero-ablation, which sets $\bar{c}_\ell = 0$ in Equation 2 so that the subspace component of every patch is removed instead of being replaced with its mean. Zero-ablation follows the same pattern as mean-ablation in all four models.

Table 5: SugarCrepe accuracy gaps under mean- and zero-ablation, with standard deviations over five runs.

		Replace			Swap		Add		
Model	Methods	Obj	Att	Rel	Obj	Att	Obj	Att	Avg
CLIP ViT-B	base	93.04	81.60	66.79	60.41	65.62	78.42	66.33	76.95
	mean-ablation	−2.66	−3.15	−1.21	**−8.00**	−4.98	−5.00	−0.58	−3.27
		±0.92	±1.04	±0.89	±2.11	±1.31	±1.42	±0.80	±0.71
	mean-ablation (random)	−0.05	−0.10	−0.11	−0.24	−0.42	−0.36	+0.20	−0.17
		±0.22	±0.83	±0.36	±0.89	±0.78	±0.96	±0.73	±0.39
	zero-ablation	−2.37	−2.99	−1.00	**−7.84**	−5.59	−6.65	−0.09	−3.61
		±0.80	±1.40	±0.79	±2.52	±1.61	±1.26	±0.56	±0.77
	zero-ablation (random)	−0.06	−0.10	−0.10	+0.08	−0.36	−0.27	+0.06	−0.14
		±0.18	±0.82	±0.36	±1.02	±0.88	±0.89	±0.73	±0.35
CLIP ViT-L	base	94.25	78.30	65.15	60.41	62.31	78.18	71.53	76.69
	mean-ablation	−1.49	+1.32	−0.73	**−4.73**	−1.83	−1.03	−0.95	−1.01
		±0.55	±0.53	±0.86	±1.18	±1.12	±0.35	±1.06	±0.18
	mean-ablation (random)	−0.18	+0.66	+0.17	+0.33	−0.36	+0.30	−0.17	+0.11
		±0.15	±0.59	±0.56	±1.17	±0.57	±0.26	±0.70	±0.17
	zero-ablation	−1.43	+1.32	+0.03	**−4.90**	−1.83	−1.45	−1.01	−0.99
		±0.78	±0.62	±0.53	±0.65	±1.26	±0.30	±1.22	±0.33
	zero-ablation (random)	−0.17	+0.66	+0.24	+0.41	−0.27	+0.30	−0.17	+0.13
		±0.16	±0.66	±0.50	±1.19	±0.56	±0.19	±0.67	±0.14
OpenCLIP ViT-L	base	96.55	84.52	70.48	61.22	69.07	90.01	79.77	83.48
	mean-ablation	−0.85	−1.85	−1.85	**−2.37**	−0.75	−0.37	−0.32	−1.00
		±0.33	±0.47	±0.36	±1.24	±0.24	±0.49	±0.85	±0.30
	mean-ablation (random)	−0.29	−0.66	+0.03	−0.57	−0.03	+0.35	+0.03	−0.05
		±0.25	±0.44	±0.49	±0.68	±0.20	±0.19	±0.21	±0.17
	zero-ablation	−0.87	−2.18	−2.01	**−2.53**	−1.05	−0.36	−0.23	−1.09
		±0.35	±0.52	±0.47	±1.06	±0.35	±0.31	±0.60	±0.29
	zero-ablation (random)	−0.24	−0.66	+0.01	−0.65	+0.00	+0.36	+0.03	−0.04
		±0.27	±0.44	±0.47	±0.55	±0.24	±0.16	±0.19	±0.14
SigLIP	base	97.46	87.18	70.91	69.80	77.78	91.80	83.38	85.91
	mean-ablation	−0.25	−0.94	−0.70	**−3.59**	−0.45	−0.16	+0.35	−0.46
		±0.10	±0.21	±0.18	±0.34	±0.38	±0.10	±0.19	±0.10
	mean-ablation (random)	−0.07	−0.15	−0.40	−0.82	−0.15	−0.08	−0.32	−0.20
		±0.12	±0.21	±0.28	±0.41	±0.55	±0.16	±0.66	±0.11
	zero-ablation	−0.23	−0.99	−0.68	**−3.43**	−0.51	−0.16	+0.26	−0.46
		±0.11	±0.26	±0.13	±0.47	±0.50	±0.12	±0.24	±0.11
	zero-ablation (random)	−0.07	−0.13	−0.38	−0.73	−0.15	−0.06	−0.32	−0.18
		±0.12	±0.20	±0.23	±0.34	±0.55	±0.14	±0.53	±0.09

"base" denotes accuracy without intervention. For each ablation, the upper row reports the mean gap (ablation - base) in percentage points and the row below it reports the standard deviation of the gap over five runs; "(random)" denotes the gap under the same ablation applied to a subspace spanned by a random orthonormal basis of the same dimension. Bold and underline mark the largest and second-largest drops. "Avg" is the accuracy over all examples.

Table 6: VALSE accuracy gaps under mean- and zero-ablation, with standard deviations over five runs.

Model	Methods	Existence	Plurality	Counting	Sp.rel.	Action		Coreference		Foil-it!	Avg
						repl.	actant swap	standard	clean		
CLIP ViT-B	base	65.54	55.58	65.15	60.75	82.56	72.60	47.60	53.85	90.56	67.54
	mean-ablation	−3.52	−0.19	−2.03	−2.73	−2.99	**−8.68**	+0.76	−0.96	−3.69	−2.80
		±2.29	±0.83	±0.64	±1.05	±1.28	±1.80	±1.28	±2.04	±0.61	±0.82
	mean-ablation (random)	+0.44	+0.40	+0.68	−0.49	−0.86	−1.77	−0.17	−1.92	−0.51	−0.14
		±1.17	±0.46	±0.57	±1.33	±0.40	±0.76	±0.51	±1.80	±0.42	±0.42
	zero-ablation	−5.90	−0.94	−2.85	−2.92	−2.62	**−8.75**	+0.59	−2.12	−3.69	−3.33
		±2.11	±0.43	±0.85	±1.24	±1.50	±1.81	±0.62	±0.80	±0.79	±0.65
	zero-ablation (random)	+0.71	+0.54	+0.63	−0.64	−0.86	−1.88	−0.23	−2.12	−0.36	−0.13
		±0.96	±0.52	±0.58	±1.43	±0.50	±1.00	±0.41	±2.19	±0.50	±0.41
CLIP ViT-L	base	67.13	58.75	66.49	57.94	83.49	68.81	53.67	56.73	92.26	68.64
	mean-ablation	−2.97	−1.34	−1.57	+0.41	−1.42	**−3.20**	+0.68	−2.31	−1.34	−1.46
		±0.75	±0.34	±0.72	±0.92	±0.64	±0.49	±0.91	±2.11	±0.82	±0.38
	mean-ablation (random)	−0.08	−0.59	−0.24	+0.71	−0.03	+0.34	+0.34	+1.92	+0.15	+0.02
		±0.68	±0.45	±0.25	±0.79	±0.40	±0.39	±0.80	±1.36	±0.24	±0.12
	zero-ablation	−2.85	−1.41	−1.71	+0.52	−1.70	**−3.44**	+0.76	−2.12	−1.44	−1.55
		±0.36	±0.54	±1.12	±0.91	±0.67	±0.14	±0.64	±1.85	±1.02	±0.45
	zero-ablation (random)	−0.12	−0.68	−0.24	+0.71	−0.15	+0.34	+0.37	+1.92	+0.11	−0.01
		±0.77	±0.42	±0.18	±0.74	±0.52	±0.39	±0.80	±1.36	±0.26	±0.13
OpenCLIP ViT-L	base	75.05	68.51	62.87	62.06	84.72	61.43	51.55	46.15	92.15	68.22
	mean-ablation	−0.83	−1.69	−0.85	−1.79	−0.46	**−4.83**	−0.59	−1.15	−0.11	−1.35
		±0.45	±0.26	±0.24	±0.57	±0.41	±0.35	±0.71	±1.25	±0.51	±0.14
	mean-ablation (random)	+0.08	−0.54	−0.37	+0.07	−0.59	−0.40	+0.08	−1.92	+0.23	−0.25
		±0.27	±0.40	±0.42	±0.54	±0.13	±0.70	±0.71	±1.36	±0.20	±0.12
	zero-ablation	−0.83	−1.62	−0.94	−1.50	−0.43	**−4.95**	−0.68	−0.96	−0.11	−1.37
		±0.51	±0.19	±0.27	±0.91	±0.46	±0.65	±0.52	±1.80	±0.40	±0.10
	zero-ablation (random)	+0.04	−0.49	−0.38	+0.15	−0.62	−0.40	+0.11	−1.92	+0.19	−0.26
		±0.33	±0.43	±0.40	±0.45	±0.15	±0.73	±0.80	±1.36	±0.19	±0.14
SigLIP	base	77.62	69.92	68.52	63.55	90.12	57.85	59.04	54.81	93.11	71.38
	mean-ablation	+0.95	+0.31	+0.29	−1.83	−0.28	**−2.21**	−1.19	+0.58	−0.57	−0.41
		±0.09	±0.46	±0.28	±0.41	±0.30	±0.35	±0.60	±1.29	±0.26	±0.18
	mean-ablation (random)	−0.28	+0.00	−0.24	−0.30	+0.06	+0.46	−0.37	−0.38	−0.13	−0.11
		±0.36	±0.51	±0.20	±0.34	±0.23	±0.46	±0.53	±1.10	±0.12	±0.11
	zero-ablation	+0.91	+0.24	+0.24	−1.83	−0.15	**−2.26**	−1.19	+0.58	−0.55	−0.42
		±0.11	±0.51	±0.27	±0.52	±0.19	±0.34	±0.47	±1.29	±0.19	±0.19
	zero-ablation (random)	−0.20	+0.05	−0.27	−0.22	−0.06	+0.34	−0.34	−0.38	−0.15	−0.13
		±0.37	±0.45	±0.20	±0.36	±0.18	±0.46	±0.55	±1.10	±0.12	±0.09

"base" denotes accuracy without intervention. For each ablation, the upper row reports the mean gap (ablation − base) in percentage points and the row below it reports the standard deviation of the gap over five runs; "(random) denotes the gap under the same ablation applied to a subspace spanned by a random orthonormal basis of the same dimension. Bold and underline mark the largest and second-largest drops. "Avg" is the accuracy over all valid examples. Under "Action", "repl." denotes action replacement; "Sp.rel." denotes spatial relations.

B.5 CHOICE OF THE SUBSPACE DIMENSION k

We use $k = 16$ in all main experiments, on the basis of the probe accuracy, the agreement of the subspaces across runs, and the effect of ablation for different values of k. Figure 9 shows the probe accuracy at each layer for each subspace dimension $k \in \{2, 4, 8, 16, 32\}$; the accuracy on PSG val increases up to $k = 16$ and changes by at most 0.21 points from $k = 16$ to $k = 32$ in all four models. To examine the agreement across runs, for each layer, we train the probe with 5 random seeds and compare the resulting subspaces for all $\binom{5}{2} = 10$ pairs of runs. Table 7 shows the results for $k \in \{2, 4, 8, 16\}$ at the layer with the highest probe accuracy on PSG val for each model and k: the overlap between the subspaces of two runs, the mean, maximum, and minimum of the cosines of the principal angles between them, and their chance levels (Random). Let $W_1, W_2 \in \mathbb{R}^{k \times d}$ denote the probe projection matrices learned at the same layer in two different runs. Overlap is $\frac{1}{k}\sum_i \cos^2\theta_i = \frac{\|W_1 W_2^\top\|_F^2}{k}$, where $\theta_1 \le \cdots \le \theta_k$ are the principal angles between the two subspaces (1 for identical and 0 for orthogonal subspaces); we report its mean $\pm$ standard deviation over the 10 pairs, and the cosines are averaged over the 10 pairs. Random is the chance level: two independent, uniformly random k-dimensional subspaces of the d-dimensional residual stream ($d = 1024$ for CLIP ViT-L and OpenCLIP ViT-L, 768 for CLIP ViT-B, and 1152 for SigLIP). Its overlap is the expectation k/d, and its cosines are averaged over 2,000 sampled pairs. In all four models and for all k, the overlap between runs is far above the chance level, so the probes trained with different seeds share a substantial part of the subspace. The standard deviation of the overlap over pairs of runs is the smallest at $k = 16$ in the three CLIP models, and in SigLIP it is small at both $k = 2$ and $k = 16$. The subspace obtained at $k = 16$ is thus stable across runs. Tables 8 and 9 show the SugarCrepe and VALSE accuracy of SigLIP under mean- and zero-ablation of the relation-binding subspace with smaller dimensions $k \in \{2, 4, 8\}$, applied at the layer at which the probe accuracy on PSG val is highest for each k. In SigLIP, swap-object shows the largest drop among the SugarCrepe tasks for every $k \in \{2, 4, 8\}$, as for $k = 16$ (Table 1), so even a two-dimensional subspace carries role information used for this task. In VALSE, the drop on actant-swap increases with k, and under mean-ablation actant-swap becomes the largest drop among the VALSE tasks only at $k = 16$. Thus, $k = 16$ is the smallest dimension at which the probe accuracy saturates and the ablation effect concentrates on the role-swap tasks in both benchmarks, and the subspace obtained at this dimension is stable across runs.

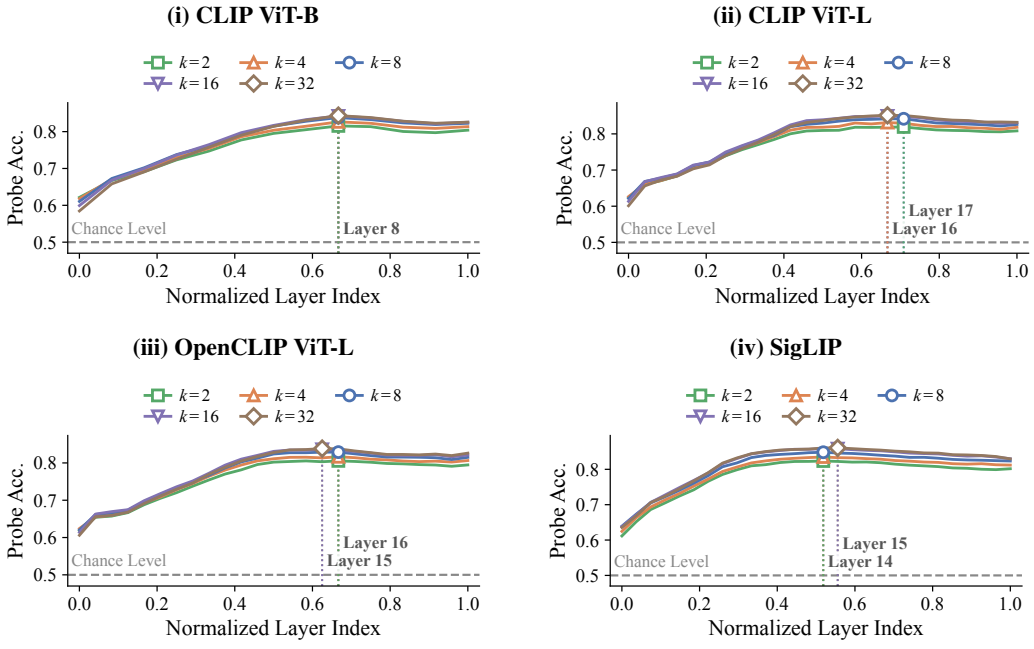


Figure 9: **Probe accuracy on PSG val across layers for each subspace dimension** $k \in \{2, 4, 8, 16, 32\}$**.** Lines show the accuracy averaged over the four negative types for each k, averaged over five seeds; bands show $\pm$ one standard deviation.

Table 7: Agreement between relation-Binding Subspaces learned in 5 runs for $k \in \{2, 4, 8, 16\}$, at the layer with the highest probe accuracy on the validation set for each model and k (marked with †).

$k = 2$

Model	Layer	Overlap	$\cos\theta_i$ mean	max	min
CLIP ViT-B ($d = 768$)	$8^\dagger$	$0.580_{\pm 0.081}$	0.742	0.878	0.606
	Random	0.003	0.041	0.064	0.019
CLIP ViT-L ($d = 1024$)	$17^\dagger$	$0.737_{\pm 0.085}$	0.854	0.915	0.793
	Random	0.002	0.036	0.055	0.016
OpenCLIP ViT-L ($d = 1024$)	$16^\dagger$	$0.521_{\pm 0.131}$	0.675	0.901	0.450
	Random	0.002	0.036	0.055	0.016
SigLIP ($d = 1152$)	$14^\dagger$	$0.962_{\pm 0.004}$	0.981	0.984	0.978
	Random	0.002	0.034	0.052	0.015

$k = 4$

Model	Layer	Overlap	$\cos\theta_i$ mean	max	min
CLIP ViT-B ($d = 768$)	$8^\dagger$	$0.597_{\pm 0.076}$	0.709	0.979	0.294
	Random	0.005	0.060	0.113	0.013
CLIP ViT-L ($d = 1024$)	$16^\dagger$	$0.577_{\pm 0.027}$	0.677	0.979	0.149
	Random	0.004	0.051	0.097	0.011
OpenCLIP ViT-L ($d = 1024$)	$16^\dagger$	$0.558_{\pm 0.069}$	0.662	0.975	0.180
	Random	0.004	0.051	0.097	0.011
SigLIP ($d = 1152$)	$14^\dagger$	$0.717_{\pm 0.091}$	0.815	0.982	0.535
	Random	0.003	0.049	0.092	0.010

$k = 8$

Model	Layer	Overlap	$\cos\theta_i$ mean	max	min
CLIP ViT-B ($d = 768$)	$8^\dagger$	$0.553_{\pm 0.051}$	0.675	0.985	0.113
	Random	0.010	0.086	0.178	0.009
CLIP ViT-L ($d = 1024$)	$17^\dagger$	$0.534_{\pm 0.055}$	0.664	0.983	0.108
	Random	0.008	0.074	0.153	0.007
OpenCLIP ViT-L ($d = 1024$)	$16^\dagger$	$0.486_{\pm 0.037}$	0.624	0.980	0.105
	Random	0.008	0.074	0.153	0.007
SigLIP ($d = 1152$)	$14^\dagger$	$0.629_{\pm 0.062}$	0.740	0.985	0.176
	Random	0.007	0.070	0.145	0.007

$k = 16$

Model	Layer	Overlap	$\cos\theta_i$ mean	max	min
CLIP ViT-B ($d = 768$)	$8^\dagger$	$0.596_{\pm 0.025}$	0.723	0.989	0.064
	Random	0.021	0.122	0.264	0.006
CLIP ViT-L ($d = 1024$)	$16^\dagger$	$0.569_{\pm 0.018}$	0.695	0.987	0.046
	Random	0.016	0.105	0.230	0.005
OpenCLIP ViT-L ($d = 1024$)	$15^\dagger$	$0.536_{\pm 0.016}$	0.676	0.983	0.066
	Random	0.016	0.105	0.230	0.005
SigLIP ($d = 1152$)	$15^\dagger$	$0.635_{\pm 0.015}$	0.751	0.989	0.091
	Random	0.014	0.099	0.216	0.005

Table 8: SugarCrepe accuracy gaps of SigLIP under mean- and zero-ablation of the relation-binding subspace with $k \in \{2, 4, 8\}$, with standard deviations over five runs.

		Replace			Swap		Add		
k	Methods	Obj	Att	Rel	Obj	Att	Obj	Att	Avg
	base	97.46	87.18	70.91	69.80	77.78	91.80	83.38	85.91
$k=2$ (layer 14)	mean-ablation	-0.02	-0.43	$\underline{-0.84}$	$\mathbf{-3.02}$	$+0.39$	-0.11	$+0.00$	-0.30
		± 0.03	± 0.14	± 0.06	± 0.22	± 0.08	± 0.12	± 0.10	± 0.02
	mean-ablation (random)	$+0.00$	$+0.13$	-0.11	-0.73	$+0.00$	-0.01	$+0.14$	-0.02
		± 0.11	± 0.25	± 0.13	± 0.34	± 0.18	± 0.13	± 0.27	± 0.04
	zero-ablation	-0.02	-0.41	$\underline{-0.83}$	$\mathbf{-3.18}$	$+0.21$	-0.12	-0.03	-0.32
		± 0.03	± 0.11	± 0.16	± 0.34	± 0.20	± 0.10	± 0.12	± 0.03
	zero-ablation (random)	$+0.00$	$+0.13$	-0.10	-0.57	$+0.00$	-0.01	$+0.14$	-0.01
		± 0.11	± 0.25	± 0.16	± 0.37	± 0.18	± 0.13	± 0.27	± 0.05
$k=4$ (layer 14)	mean-ablation	-0.05	-0.46	$\underline{-0.90}$	$\mathbf{-1.80}$	-0.06	-0.16	$+0.35$	-0.30
		± 0.10	± 0.25	± 0.15	± 0.22	± 0.25	± 0.07	± 0.33	± 0.07
	mean-ablation (random)	-0.04	$+0.23$	$+0.03$	-0.90	$+0.21$	-0.04	$+0.23$	$+0.02$
		± 0.10	± 0.23	± 0.20	± 0.18	± 0.13	± 0.13	± 0.28	± 0.08
	zero-ablation	-0.05	-0.41	$\underline{-0.84}$	$\mathbf{-2.12}$	-0.18	-0.17	$+0.38$	-0.31
		± 0.10	± 0.21	± 0.16	± 0.45	± 0.13	± 0.11	± 0.39	± 0.06
	zero-ablation (random)	-0.04	$+0.23$	$+0.00$	-0.90	$+0.21$	-0.05	$+0.20$	$+0.01$
		± 0.10	± 0.19	± 0.16	± 0.18	± 0.13	± 0.14	± 0.26	± 0.08
$k=8$ (layer 14)	mean-ablation	-0.22	-0.66	$\underline{-0.84}$	$\mathbf{-2.94}$	-0.33	-0.42	$+0.29$	-0.49
		± 0.10	± 0.17	± 0.15	± 0.61	± 0.57	± 0.19	± 0.43	± 0.12
	mean-ablation (random)	-0.05	-0.18	-0.21	-0.65	-0.12	$+0.01$	-0.03	-0.10
		± 0.03	± 0.19	± 0.21	± 0.22	± 0.13	± 0.28	± 0.28	± 0.09
	zero-ablation	-0.16	-0.61	$\underline{-0.74}$	$\mathbf{-2.86}$	-0.42	-0.42	$+0.32$	-0.45
		± 0.15	± 0.23	± 0.08	± 0.65	± 0.33	± 0.20	± 0.45	± 0.14
	zero-ablation (random)	-0.05	-0.18	-0.20	-0.65	-0.09	$+0.01$	-0.09	-0.10
		± 0.03	± 0.19	± 0.20	± 0.22	± 0.17	± 0.29	± 0.32	± 0.08

"base" denotes accuracy without intervention. For each ablation, the upper row reports the mean gap (ablation $-$ base) in percentage points and the row below it reports the standard deviation of the gap over five runs; "(random)" denotes the gap under the same ablation applied to a subspace spanned by a random orthonormal basis of the same dimension. Bold and underline mark the largest and second-largest drops. "Avg" is the accuracy over all examples. For each k, the intervened layer is the one at which probe accuracy on PSG val is highest for that k.

Table 9: VALSE accuracy gaps of SigLIP under mean- and zero-ablation of the relation-binding subspace with $k \in \{2, 4, 8\}$, with standard deviations over five runs.

k	Methods	Existence	Plurality	Counting	Sp.rel.	Action		Coreference		Foil-it!	Avg
						repl.	actant swap	standard	clean		
	base	77.62	69.92	68.52	63.55	90.12	57.85	59.04	54.81	93.11	71.38
$k=2$ (layer 14)	mean-ablation	+0.48	+0.63	−0.11	**−1.01**	−0.31	−0.32	−0.71	−0.38	−0.02	−0.14
		±0.23	±0.18	±0.11	±0.21	±0.22	±0.17	±0.14	±0.53	±0.05	±0.04
	mean-ablation (random)	+0.24	+0.21	+0.01	−0.34	+0.00	−0.04	+0.06	+0.19	+0.00	+0.02
		±0.17	±0.28	±0.20	±0.41	±0.22	±0.22	±0.19	±0.43	±0.07	±0.11
	zero-ablation	+0.48	+0.63	−0.15	**−1.05**	−0.22	−0.36	−0.59	−0.96	−0.04	−0.15
		±0.23	±0.13	±0.14	±0.17	±0.18	±0.22	±0.18	±0.96	±0.06	±0.04
	zero-ablation (random)	+0.24	+0.16	+0.02	−0.34	+0.00	−0.06	+0.06	+0.19	+0.00	+0.02
		±0.17	±0.27	±0.23	±0.31	±0.22	±0.24	±0.19	±0.43	±0.07	±0.11
$k=4$ (layer 14)	mean-ablation	+0.20	−0.05	−0.26	−0.82	−0.28	−0.59	−0.59	**−0.96**	−0.25	−0.33
		±0.24	±0.11	±0.14	±0.31	±0.13	±0.24	±0.37	±0.96	±0.21	±0.08
	mean-ablation (random)	−0.20	+0.12	−0.02	−0.15	+0.12	+0.02	+0.08	+0.58	+0.04	+0.02
		±0.28	±0.40	±0.12	±0.43	±0.13	±0.14	±0.31	±1.29	±0.12	±0.07
	zero-ablation	+0.28	+0.00	−0.36	**−0.93**	−0.28	−0.78	−0.59	−0.77	−0.19	−0.37
		±0.23	±0.14	±0.18	±0.30	±0.13	±0.27	±0.27	±0.80	±0.16	±0.04
	zero-ablation (random)	−0.20	+0.14	−0.09	−0.15	+0.09	+0.11	+0.11	+0.58	+0.02	+0.01
		±0.28	±0.37	±0.14	±0.43	±0.14	±0.07	±0.32	±1.29	±0.12	±0.06
$k=8$ (layer 14)	mean-ablation	+0.44	−0.33	−0.53	**−1.35**	−0.28	−1.10	−0.93	−1.15	−0.34	−0.57
		±0.26	±0.32	±0.07	±0.38	±0.17	±0.33	±0.26	±0.80	±0.09	±0.07
	mean-ablation (random)	+0.48	+0.12	−0.41	−0.71	+0.15	+0.34	−0.42	+0.19	−0.04	−0.12
		±0.41	±0.30	±0.28	±0.61	±0.38	±0.50	±0.91	±1.25	±0.24	±0.17
	zero-ablation	+0.48	−0.33	−0.50	−1.27	−0.28	**−1.31**	−0.99	−0.58	−0.34	−0.58
		±0.27	±0.29	±0.09	±0.41	±0.30	±0.27	±0.33	±0.86	±0.09	±0.08
	zero-ablation (random)	+0.51	+0.12	−0.40	−0.67	+0.12	+0.36	−0.40	+0.19	−0.04	−0.11
		±0.23	±0.37	±0.18	±0.57	±0.30	±0.22	±0.92	±1.25	±0.24	±0.08

"base" denotes accuracy without intervention. For each ablation, the upper row reports the mean gap (ablation − base) in percentage points and the row below it reports the standard deviation of the gap over five runs; "(random)" denotes the gap under the same ablation applied to a subspace spanned by a random orthonormal basis of the same dimension. Bold and underline mark the largest and second-largest drops. "Avg" is the accuracy over all valid examples. Under "Action", "repl." denotes action replacement; "Sp.rel." denotes spatial relations. For each k, the intervened layer is the one at which probe accuracy on PSG val is highest for that k.