PACO-SAE: EVALUATING CONCEPT GROUNDING AND FEATURE SPLITTING IN VISION SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) are used to interpret vision models, but predicting a concept does not establish whether a feature activates on the object it represents. We introduce PACO-SAE, an object-level benchmark and evaluation protocol that evaluates concept grounding as semantic segmentation. Built from exhaustively annotated PACO-LVIS objects, it combines grounding F1 with separate feature selection and testing across 63 object concepts. Across four SAE families and four sparsity levels, image-level probing and peak localization select features with lower test grounding F1 than grounding-based selection, showing that neither substitutes for evaluating the entire activated region. PACO-SAE also enables object-level analysis of feature splitting: different features best ground different objects of one concept, substantially outperforming a single feature selected for the highest average grounding. To examine whether these splits reflect visual specialization or residual splitting, we introduce Pairwise Visual Separation (PVS), which compares visual differences between groups of objects with variation within each group, without additional subclass labels. A human study of 200 feature pairs supports PVS as an indicator of recognizable visual differences. Together, PACO-SAE establishes an object-level evaluation foundation, reveals the limits of prediction and localization as grounding criteria, and connects the extent of feature splitting to its visual meaning.

1 INTRODUCTION

Mechanistic interpretability aims to reverse-engineer neural networks by identifying their components and how they interact to produce outputs (Olah, 2022). This is difficult when individual neurons respond to unrelated patterns, as described by the superposition hypothesis: networks can encode more patterns than they have neurons (Elhage et al., 2022). Sparse autoencoders (SAEs) address this challenge by reconstructing a model's internal representations using only a small number of learned features at a time. This approach has revealed interpretable features in language, vision–language, and vision models (Huben et al., 2024; Pach et al., 2025; Stevens et al., 2025). However, SAEs are typically trained for accurate reconstruction and sparse activation, without supervision specifying what each feature should represent (Bussmann et al., 2024). Their usefulness for interpretability therefore requires evaluation beyond the reconstruction–sparsity trade-off, as emphasized by task-specific evaluations and SAEBench (Makelov et al., 2025; Karvonen et al., 2025).

In vision SAEs, *concept probing* tests whether feature activations predict a human-recognizable concept, such as 'dog' (Choi et al., 2025; Stevens et al., 2025). Image-level probing combines a feature's activations from different patches into a single score to predict whether the concept is present. However, a feature selected for accurately predicting 'dog' may activate only on the surrounding grass rather than on the dog itself. This motivates evaluating *concept grounding*: whether the feature activates on the object it is meant to represent. Figure 1(a) illustrates this grounding gap between predicting that a concept is present and spatially aligning activation with the corresponding object. ConceptScope evaluates SAE activation masks through binary semantic segmentation (Choi et al., 2025), while Stevens et al. (2025) use segmentation labels to evaluate patch-level concept predictions. VISTA's Spatial Localization Accuracy (SLA) instead checks whether the most strongly activating patch falls inside the object's bounding box (Kravets et al., 2026). These approaches dif-

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.

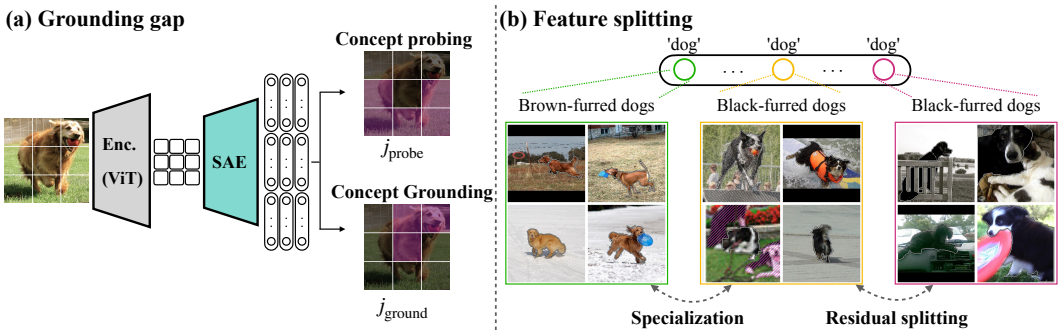


Figure 1: Overview of the grounding gap and feature splitting in vision SAEs. (a) Concept probing can select a feature that predicts concept presence without spatially aligning with the target object, creating a grounding gap. Concept grounding evaluates this spatial correspondence. (b) Feature splitting occurs when different features best ground different objects of the same concept. The resulting groups may reflect visual specialization, as when features separate brown- and black-furred dogs, or residual splitting, as when visually similar black-furred dogs are divided across features.

fer in how features are selected and scored. However, direct comparisons on the same objects are needed to establish which spatial errors these selection criteria overlook.

A further challenge is *feature splitting*, in which a concept is represented across multiple SAE features (Makelov et al., 2025; Chanin et al., 2025). Different features may respond to different examples of the same concept, so a single feature can ground some objects well while missing others. Other features may still ground the objects that it misses. Importantly, splitting need not be undesirable. Features responding separately to brown-furred dogs and black-furred dogs may capture meaningful *specialization*, whereas features dividing visually similar dogs without a consistent distinction may reflect *residual splitting*. Figure 1(b) illustrates these two cases by contrasting a separation based on fur color with a division between visually similar groups of black-furred dogs. Existing evaluations assess splitting through the benefit of using multiple features for concept prediction (Bussmann et al., 2025; Klotz et al., 2026). However, benchmarks that score several objects together can hide which objects each feature grounds, making both the extent of splitting and the distinction between specialization and residual splitting difficult to assess.

We introduce **PACO-SAE**, an object-level benchmark and evaluation protocol that formulates concept grounding in vision SAEs as a semantic segmentation task. We adapt existing object masks from PACO-LVIS (Ramanathan et al., 2023) through two complementary design choices. First, we use object-level crops with exhaustive annotations. The annotations define the target mask and the regions to evaluate, so we can score a feature's active patches as a segmentation prediction. Second, because SAE features do not come with concept labels, we use a selection set to select features for each concept and evaluate those same features on a separate test set (§ 3.2). We define a patch-level grounding F1 score that compares all patches activated by a feature with the target mask. We compare features selected by image-level probing, a mask-based version of SLA (mSLA), or grounding F1 on the same test images. This tests whether features selected by concept probing or mSLA ground objects as well as those selected by grounding F1 (§ 3.3).

We then investigate whether lower grounding scores for a single selected feature reflect feature splitting. We compare using one feature selected for a concept with using the best candidate for each object, keeping the candidates fixed on the selection set. We also measure how often different objects favor different features and how many features are involved (§ 3.4). To examine the visual differences underlying these splits, we group object examples by their best-grounding feature and introduce **Pairwise Visual Separation (PVS)**. PVS compares the visual differences between two groups with the variation within each group. It helps distinguish visual specialization from residual splitting without requiring additional subclass labels (§ 3.4).

Our empirical study spans four SAE families and four sparsity levels. Grounding-based selection consistently achieves higher mean grounding F1 on test images than selection based on image-level probing or mSLA (§ 4.1). We also observe extensive feature splitting across the evaluated vision SAEs: at $k = 128$, matching each object to its best candidate raises grounding F1 from about 0.52

for a single fixed feature to 0.73–0.74 (§ 4.2). In a human study of 200 feature pairs, the human-assessed specialization rate increases across the five PVS quintiles, supporting PVS as an indicator of visual specialization (§ 4.3). Across sparsity levels, reduced splitting does not necessarily coincide with clearer visual specialization. Our contributions are summarized as follows:

- **An object-level semantic segmentation benchmark.** PACO-SAE combines exhaustive annotations, grounding F1, and separate feature selection and testing.
- **Systematic evidence of grounding gaps.** Probing and mSLA select features with lower test grounding F1 than grounding-based selection.
- **Quantifying feature splitting.** We quantify how different features ground different objects of the same concept, revealing extensive splitting.
- **A human-supported measure of specialization.** PVS helps distinguish visual specialization from residual splitting without additional subclass labels.

2 Related Work

SAE evaluation and concept grounding. Prior work evaluates SAEs along several dimensions, including reconstruction, sparsity, feature interpretability, and downstream behavior. Language-model studies use automated methods to evaluate individual feature interpretability (Huben et al., 2024), examine feature recovery and the sparsity of downstream effects (Gao et al., 2025), and assess task-specific disentanglement and control (Makelov et al., 2025). SAEBench brings multiple criteria into a common evaluation suite (Karvonen et al., 2025), while subsequent studies examine their reliability across related SAEs (Chanin, 2026) and whether common metrics distinguish SAEs fit to representations from trained and randomly initialized transformers (Heap et al., 2026). Vision SAE studies evaluate or interpret features through the semantic coherence of activating images (Pach et al., 2025), correspondence with annotated concepts (Klotz et al., 2026), and automatically generated feature descriptions (Zhang et al., 2025). Related spatial analyses compare unit activations with semantic masks (Bau et al., 2017), derive patch-wise spatial attributions (Lim et al., 2025), fit one-dimensional probes to patch-level segmentation labels (Stevens et al., 2025), evaluate activation masks against semantic segmentations (Choi et al., 2025), or test whether the strongest activation lies within an object bounding box (Kravets et al., 2026). PACO-SAE compares concept probing, peak localization, and grounding-based selection by scoring each selected feature's activation map against object masks from a separate test set with object-level F1.

Feature splitting and visual specialization. A separate line of work examines how semantic concepts are distributed across SAE features. Language-model studies document feature splitting, in which a concept is distributed across several features (Makelov et al., 2025), and feature absorption, in which a broader feature misses cases captured by more specific features (Chanin et al., 2025). Other work questions whether SAEs recover canonical units (Leask et al., 2025) and analyzes how they represent concept manifolds through groups of global or local features (Bhalla et al., 2026). In vision, concepts can be encoded across multiple CNN filters (Fong & Vedaldi, 2018), and finer visual sub-concepts can be discovered from concept embedding models without additional annotations (Hill et al., 2026). Training methods address feature granularity or consistency through nested dictionaries (Bussmann et al., 2025), structured sparsity (Liao et al., 2026), and cross-sample regularization (Jin et al., 2026). Existing evaluations characterize concept representations with sparse probes that may combine multiple SAE features (Bussmann et al., 2025) or explicitly permit many-to-one mappings from features to annotated concepts (Klotz et al., 2026). PACO-SAE studies how objects of the same category are assigned across features and whether the resulting groups are visually separated, without assuming predefined subclass labels.

3 Methods

3.1 Preliminaries

ViT-based vision encoders such as CLIP (Radford et al., 2021) and DINO (Siméoni et al., 2025) represent an image as a grid of patch-level features. Let $\mathbf{x}_{ip} \in \mathbb{R}^d$ denote the d-dimensional representation of patch p in image i. An SAE maps this representation to a sparse activation over a

dictionary of m features and reconstructs the input as

$$\mathbf{h}_{ip} = W_e \mathbf{x}_{ip} + \mathbf{b}_e, \qquad \mathbf{z}_{ip} = \sigma(\mathbf{h}_{ip}), \qquad \hat{\mathbf{x}}_{ip} = W_d \mathbf{z}_{ip} + \mathbf{b}_d, \tag{1}$$

where $W_e \in \mathbb{R}^{m \times d}$ and $\mathbf{b}_e \in \mathbb{R}^m$ are the encoder parameters, $W_d \in \mathbb{R}^{d \times m}$ and $\mathbf{b}_d \in \mathbb{R}^d$ are the decoder parameters, and σ is the activation function. SAE training encourages the reconstruction $\hat{\mathbf{x}}_{ip}$ to preserve the input representation $\mathbf{x}_{ip}$, with sparsity regularization on $\mathbf{z}_{ip}$. Each coordinate z_{ipj}, for $j \in \{1, \ldots, m\}$, denotes the activation of feature j at patch p. Consequently, only a small subset of features is active for each patch.

However, SAE training does not specify which dictionary feature is associated with a human-defined concept. Associating features with such concepts therefore requires a separate feature selection step. Conventional evaluation uses a selection split to identify the feature that best matches each concept under a given metric (Choi et al., 2025; Stevens et al., 2025). The selected feature–concept matches are then evaluated on a separate test split. For a metric $\mathcal{M}$ that quantifies the correspondence between feature j and concept c, let $\mathcal{M}_{\text{sel}}(j, c)$ and $\mathcal{M}_{\text{test}}(j, c)$ denote its values measured on the selection and test splits, respectively. The feature selected for concept c is

$$j_{\mathcal{M}}(c) = \arg\max_{j \in [m]} \mathcal{M}_{\text{sel}}(j, c). \tag{2}$$

The selected feature for each concept is then fixed and evaluated on the test split. The performance of the resulting feature–concept associations is summarized by the concept-macro average

$$\mathcal{M}_{\text{test}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathcal{M}_{\text{test}}\big(j_{\mathcal{M}}(c), c\big). \tag{3}$$

We consider two existing feature selection criteria. *Concept probing* uses image-level AUROC to identify features that predict concept presence (Lim et al., 2025). *Spatial Localization Accuracy* (SLA) checks whether the most strongly activated patch falls inside an object's bounding box (Kravets et al., 2026); we replace the box with the target mask to obtain mSLA. We test whether features selected by either criterion ground objects as well as those selected by the grounding score G (§ 3.3). We write j_{probe}, j_{mSLA}, and j_{ground} for the features selected with $\mathcal{M} = \text{AUROC}$, mSLA, and G, respectively; by Eq. equation 3, G_{test} is the test grounding score of j_{ground}. Implementation details are provided in Appendix A.

3.2 PACO-SAE: OBJECT-LEVEL EVALUATION WITH EXHAUSTIVE ANNOTATIONS

Semantic segmentation provides a natural basis for evaluating concept grounding, but an SAE outputs feature activations rather than class-labeled masks or object bounding boxes. The selection step in § 3.1 associates a feature with a concept, allowing its active patches to be treated as a prediction of that concept. To score this prediction reliably, all objects of the target concept must be annotated: otherwise, activation outside a labeled object may belong to an unlabeled object of the same concept rather than be an error. We also evaluate objects individually, since image-level scores can hide differences between objects of the same concept. For example, a feature may ground a black-furred dog but miss a brown-furred dog in the same image. Scoring the dogs together records only partial coverage of 'dog', obscuring the difference between accurate grounding of one object and failure on the other. We therefore construct **PACO-SAE** from PACO-LVIS (Ramanathan et al., 2023), using only objects whose concept is exhaustively annotated in the source image and evaluating each object in a crop that includes its surrounding context. This supports semantic segmentation-based grounding evaluation and analysis of feature splitting within the same concept. Although our experiments focus on objects, PACO's object–part hierarchy and attributes for both objects and parts support future extensions of this evaluation.

Each object is cropped with context in a square whose side is up to twice the longer bounding-box side, then resized to 224×224 pixels (Figure 2). The object mask defines the target; padding and other objects of the same concept are excluded from scoring. PACO-LVIS train and test yield 127,974 selection crops and 27,110 test crops across 63 object concepts, with no shared source images. Crop construction and eligibility criteria are detailed in Appendix B.

3.3 MEASURING CONCEPT GROUNDING

We evaluate concept grounding as semantic segmentation by comparing a feature's active patches with the target object's mask. For feature j, let $b_{ipj} = \mathbf{1}[z_{ipj} > 0]$ indicate whether patch p is

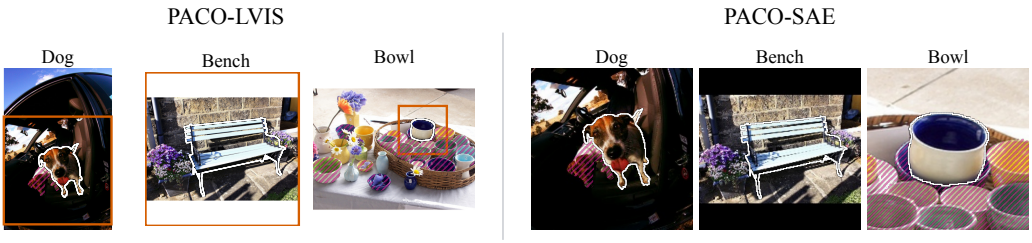


Figure 2: Constructing PACO-SAE. Source images (left) are converted into context-preserving anchor-object crops (right). Orange boxes show crop windows, white outlines mark target objects, and hatching marks other objects of the same category that are excluded from scoring. The examples illustrate a crop capped at the image width ('dog'), padding along one axis ('bench'), and multiple same-category objects in one source image ('bowl').

active. For a crop i of concept c, V_i contains the valid evaluation patches, excluding padding and other objects of the same concept, and $y_{ip}^c \in \{0,1\}$ indicates whether patch p belongs to the target object. The grounding F1 score is

$$\begin{aligned} \mathrm{TP}_i(j,c) &= \sum_{p\in V_i} b_{ipj} y_{ip}^c, & \mathrm{FP}_i(j,c) &= \sum_{p\in V_i} b_{ipj}(1-y_{ip}^c), \\ \mathrm{FN}_i(j,c) &= \sum_{p\in V_i} (1-b_{ipj}) y_{ip}^c, & \mathrm{F1}_i(j,c) &= \frac{2\mathrm{TP}_i(j,c)}{2\mathrm{TP}_i(j,c)+\mathrm{FP}_i(j,c)+\mathrm{FN}_i(j,c)}. \end{aligned} \tag{4}$$

For each split $s \in \{\mathrm{sel}, \mathrm{test}\}$, we average these F1 scores over objects of concept c to obtain the grounding score $G_s(j,c)$, which we also refer to as grounding F1.

3.4 From feature splitting to visual specialization

Oracle-based grounding analysis. A single feature $j_{\mathrm{ground}}(c)$ selected for a concept c may not provide the best grounding for every object of that concept. To examine this possibility, we compare $\mathrm{G}_{\mathrm{test}}$ with $\mathrm{G}_{\mathrm{oracle}}$, which averages, over all test images, the highest per-image grounding F1 among candidate features. An improvement of $\mathrm{G}_{\mathrm{oracle}}$ over $\mathrm{G}_{\mathrm{test}}$ indicates that the representation of a concept is split across multiple SAE features.

We characterize this feature splitting by defining the empirical winner distribution $p(j \mid c)$, for $j \in \mathcal{J}_c$, on the test set. We restrict candidate features to those that show a strong correspondence to concept c, rather than allowing arbitrary dictionary features to win on the test set. Specifically, $\mathcal{J}_c$ contains the 32 features with the highest grounding F1 scores for concept c on the selection set.

We estimate this distribution by measuring how often each candidate feature is the best match for a given concept across its test images. Concretely, for each test image i of concept c, let $w_i^c = \arg\max_{j\in\mathcal{J}_c} \mathrm{F1}_i(j,c)$ denote the candidate feature with the highest per-image grounding F1. Although an argmax always exists, all candidates may ground the target poorly. We therefore retain only *resolved* images, $\mathcal{R}_c = \{i : \mathrm{F1}_i(w_i^c, c) > \tau\}$, where τ is the 95th percentile of best-candidate grounding scores after randomly permuting target-patch labels on the selection set. For each candidate feature $j \in \mathcal{J}_c$, the empirical winner distribution is

$$p(j \mid c) = \frac{|A_{c,j}|}{|\mathcal{R}_c|}$$

where $A_{c,j} = \{i \in \mathcal{R}_c : w_i^c = j\}$ denotes the set of resolved test images for which feature j is the winner. Drawing on diversity measures from ecology, we quantify feature splitting using Simpson diversity and the effective feature count:

$$D_{\mathrm{oracle}}(c) = 1 - \sum_{j\in\mathcal{J}_c} p(j \mid c)^2, \qquad N_{\mathrm{eff}}(c) = \frac{1}{\sum_{j\in\mathcal{J}_c} p(j \mid c)^2} \tag{5}$$

$D_{\mathrm{oracle}}(c)$ is the probability that two objects sampled independently from $\mathcal{R}_c$ have different best-grounding features. $N_{\mathrm{eff}}(c)$ is the reciprocal of the probability of drawing the same best-grounding feature twice: a probability of $1/10$ corresponds to ten effective features.

Table 1: Held-out comparison of feature-selection criteria at $k = 128$. Each criterion selects one feature per concept on the selection set, which is then evaluated on the test set using concept-prediction AUROC and grounding F1. Values are concept-macro means over 63 concepts; bold marks the largest value within each SAE and metric. Random BatchTopK is untrained.

	TopK			BatchTopK			BatchTopK + Matryoshka			BatchTopK + C^2R			Random BatchTopK		
Metric	j_{probe}	j_{mSLA}	j_{ground}	j_{probe}	j_{mSLA}	j_{ground}	j_{probe}	j_{mSLA}	j_{ground}	j_{probe}	j_{mSLA}	j_{ground}	j_{probe}	j_{mSLA}	j_{ground}
AUROC	**0.959**	0.929	0.938	**0.960**	0.922	0.941	**0.965**	0.926	0.930	**0.960**	0.923	0.935	**0.861**	0.811	0.810
F1	0.480	0.490	**0.519**	0.483	0.499	**0.519**	0.489	0.490	**0.524**	0.482	0.493	**0.524**	0.268	0.319	**0.348**

Assessing visual specialization. Feature splitting may reflect visual specialization, such as differences in pose or appearance, rather than residual splitting. However, exhaustively annotating the many possible visual distinctions is impractical. We therefore compare visual differences within and between groups of objects matched to different features, without requiring additional subclass labels. For two groups $A_{c,j}$ and $A_{c,k}$, we represent each object by the normalized sum of its pre-SAE patch vectors within the target mask:

$$\mathbf{u}_i^c = \text{normalize}\left(\sum_{p \in V_i} y_{ip}^c \mathbf{x}_{ip}\right). \tag{6}$$

We compute the mean between-group distance g_{jk} and the mean within-group distances g_{jj} and g_{kk}:

$$g_{jk} = \frac{\sum_{i \in A_{c,j}} \sum_{\ell \in A_{c,k}} \|\mathbf{u}_i^c - \mathbf{u}_\ell^c\|_2}{|A_{c,j}||A_{c,k}|}, \qquad g_{jj} = \frac{\sum_{i,\ell \in A_{c,j}: i \neq \ell} \|\mathbf{u}_i^c - \mathbf{u}_\ell^c\|_2}{|A_{c,j}|(|A_{c,j}| - 1)}. \tag{7}$$

Inspired by the energy-distance comparison underlying DISCO analysis (Rizzo & Székely, 2010), we define **Pairwise Visual Separation (PVS)** between features j and k as

$$\text{PVS}_{jk} = \frac{2g_{jk} - g_{jj} - g_{kk}}{2g_{jk}}, \qquad g_{jk} > 0. \tag{8}$$

Higher PVS indicates greater visual differences between the two groups relative to the variation within each group. Excluding self-pairs from the within-group means mitigates the group-size dependence induced by their zero distances. Because PVS measures separation in the backbone encoder's representation space, we conduct human evaluation to assess whether it corresponds to perceptually meaningful differences.

4 Experimental Results

Experimental settings. We evaluate TopK (Gao et al., 2025), BatchTopK (Bussmann et al., 2024), and BatchTopK with either Matryoshka training (Bussmann et al., 2025) or C^2R regularization (Jin et al., 2026), each at $k \in \{32, 64, 128, 256\}$. All 16 SAEs are trained on ImageNet-1K for 10^7 steps using the 11th (second-to-last) block of DINOv3 ViT-B/16 (Siméoni et al., 2025), with 49,152 features and an expansion factor of 64. We include an untrained, randomly initialized BatchTopK baseline. The reconstruction–sparsity frontier is reported in Appendix D (Figure A2).

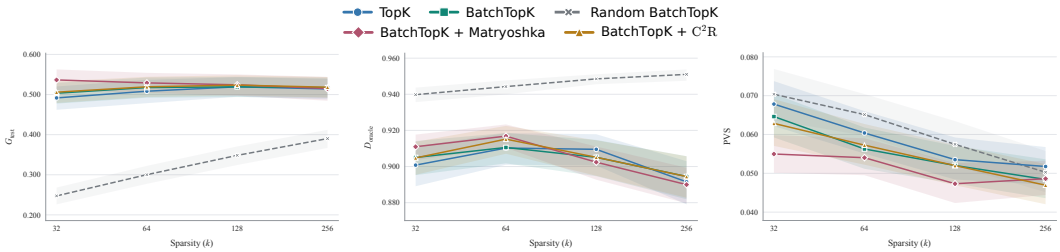
4.1 Concept prediction and mSLA do not ensure accurate grounding

Grounding-based selection identifies more spatially accurate features. PACO-SAE compares all selection criteria by grounding F1 on the same test objects (Table 1). Probing-selected features achieve AUROC of 0.959–0.965 but grounding F1 of 0.480–0.489. mSLA selection raises F1 to 0.490–0.499, below the 0.519–0.524 achieved by grounding-based selection. Grounding-selected features also have higher AUROC than mSLA-selected features in all four trained families. Thus, locating a strong response on an object does not establish that the feature grounds the whole object while avoiding the background. Paired concept-level grounding-F1 gaps remain positive on average at every sparsity level (Appendix E, Figure A3).



(a) Car: probing versus grounding. **(b)** Dog: mSLA versus grounding.

Figure 3: Grounding gaps under concept probing and mSLA selection. Activation maps compare features selected by concept probing or mSLA with grounding-selected features for 'car' in (a) and 'dog' in (b), using a trained BatchTopK SAE at $k = 128$. White outlines mark targets, hatching marks ignored same-category objects, and dots in (b) mark activation maxima. Each feature uses a fixed display scale based on its selection-set 99th percentile.



(a) Grounding (G_{test}). **(b)** Splitting (D_{oracle}). **(c)** Visual specialization (PVS).

Figure 4: Grounding, feature splitting, and visual specialization across activation budgets. Points are concept-macro means over 63 concepts in (a,b) and the common 58 concepts with valid PVS in (c). Shaded bands are pointwise 95% concept-bootstrap intervals from 2,000 resamples, not variation across training runs. Larger k permits denser activation, and dashed gray curves denote untrained Random BatchTopK.

Grounding distinguishes trained from random features. Under probing-based selection, untrained BatchTopK achieves AUROC of 0.861 but grounding F1 of 0.268; training increases these to 0.960 and 0.483. Strong image-level prediction can therefore arise before SAE training, while object grounding improves substantially with dictionary learning. Matryoshka and C^2R aim to preserve concept representations through hierarchical or cross-sample constraints (Bussmann et al., 2025; Jin et al., 2026), but each improves grounding-selected F1 by only 0.005 over BatchTopK, without closing the selector gap.

Probing and mSLA miss different spatial errors. The qualitative analysis in Figure 3 illustrates the spatial errors these criteria overlook. For 'car', the probing-selected feature activates mainly on the roadway, a region strongly associated with cars rather than the car itself. The grounding-selected feature instead responds to the car, including the partially visible example. For 'dog', the mSLA-selected feature places its strongest activation on the dog but also activates across the background. The grounding-selected feature is more focused on the dog. These examples distinguish predicting an object, locating one strong response, and grounding its full extent.

Better reconstruction does not guarantee better grounding. Although reconstruction improves as k increases (Figure A2), grounding F1 does not improve consistently (Figure 4a). TopK and BatchTopK improve at intermediate k before declining at 256, while Matryoshka decreases overall from 32. This supports evaluating object grounding directly, rather than inferring it from reconstruction quality (Makelov et al., 2025; Karvonen et al., 2025; Klotz et al., 2026).

Table 2: Grounding, feature splitting, and visual specialization at $k = 128$. G_{test} uses the grounding-selected feature, whereas G_{oracle} is a mask-informed analysis upper bound using the best of 32 selection-set candidates for each object. D_{oracle}, N_{eff}, and $p_{j_{\text{ground}}}$ summarize oracle-winner assignments among resolved objects, while PVS measures visual separation between winner groups. Values are concept-macro means over 63 concepts, except PVS over the common 58 valid concepts; Random BatchTopK is untrained.

SAE	G_{test}	G_{oracle}	D_{oracle}	N_{eff}	$p_{j_{\text{ground}}}$	PVS
TopK	0.519	0.740	0.910	12.44	0.158	0.054
BatchTopK	0.519	0.739	0.905	12.09	0.161	0.052
+Matryoshka	0.524	0.730	0.902	11.46	0.153	0.047
+C^2R	0.524	0.739	0.905	12.14	0.169	0.052
RandomBatchTopK	0.348	0.674	0.949	20.21	0.089	0.057



(a) Low PVS: 0.003 (percentile 7.7%) **(b)** High PVS: 0.327 (percentile 93.8%)

Figure 5: Low- and high-PVS feature pairs for 'fan'. For a trained BatchTopK SAE at $k = 128$, feature #28463 is shared by the low-PVS pair on the left and the high-PVS pair on the right. Each feature is represented by four examples from its retained oracle-winner group, with target objects outlined in white. PVS uses all retained group members, and percentiles are computed among the 66 valid feature pairs for 'fan'.

4.2 Object concepts are split across multiple features

Different features ground different objects of the same concept. We examine whether a single selected feature misses objects that other candidates ground well. Using the 32 candidates fixed on the selection set (§ 3.4), oracle matching raises F1 from 0.519–0.524 for G_{test} to 0.730–0.740 for G_{oracle}, a relative improvement of 39–43% (Table 2). This mask-informed comparison is an analysis upper bound, not a test-time selection method. The gap persists at every sparsity level (Appendix F, Figure A4), showing that different candidates ground objects better than one fixed feature.

Feature splitting persists across models and sparsity levels. Among objects exceeding the grounding threshold, $D_{\text{oracle}} = 0.902$–$0.910$ at $k = 128$: two objects of the same concept have different best-grounding features about 90% of the time, on average across concepts. The effective feature count is 11.46–12.44, and j_{ground} is best for only 15.3–16.9% of objects. Together, these results quantify extensive splitting rather than a few isolated mismatches. Splitting remains high throughout the sparsity sweep, rising slightly from $k = 32$ to 64 and decreasing toward 256 (Figure 4b). Denser activation neither consistently increases splitting nor eliminates it.

Training reduces but does not eliminate feature splitting. Training BatchTopK raises G_{test} from 0.348 to 0.519 and G_{oracle} from 0.674 to 0.739, while reducing N_{eff} from 20.21 to 12.09 and D_{oracle} from 0.949 to 0.905. Training therefore improves grounding and concentrates the best matches on fewer features, but many features still divide the objects of each concept. Whether these splits reflect specialization or residual splitting requires examining the objects they separate.

4.3 PVS helps distinguish specialization from residual splitting

Feature splitting can reflect specialization or residual splitting. Figure 5 contrasts two feature pairs for 'fan'. Features #28463 and #28602 both match portable fans and have PVS 0.003. Pairing #28463 with #34374 instead gives PVS 0.327 and separates portable from ceiling fans. The

latter captures recognizable specialization, whereas the former has little visual distinction between groups. PVS thus helps assess the visual differences behind a split without assuming that every split represents a distinct subclass.

Human judgments support PVS as a specialization indicator. We evaluate 200 feature pairs from BatchTopK at $k = 128$, showing four randomly sampled images per feature with the target objects outlined. Annotators judge whether the groups show a consistent visual difference; 195 pairs remain after the NA exclusions in Appendix C. The majority-Yes specialization rate increases steadily across five PVS quintiles (Figure 6). Higher PVS therefore identifies pairs more likely to show human-recognizable differences, but does not cleanly separate specialization from residual splitting.

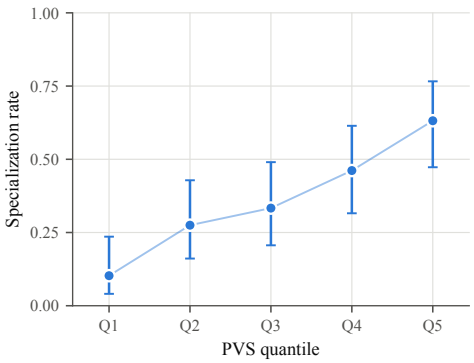


Figure 6: Human-assessed specialization across PVS quintiles. The study samples 200 feature pairs, with 20 per global PVS decile, and summarizes specialization rates in five PVS quintiles. Specialization is determined by majority-Yes judgments of a recurring visual difference between the two groups.

Matryoshka and C^{2}R have limited effects on feature splitting. Matryoshka reduces absorption (Bussmann et al., 2025), and C^{2}R regularizes cross-sample feature use to mitigate splitting (Jin et al., 2026) in language models. Here, at $k = 128$, C^{2}R matches BatchTopK in D_{oracle} (0.905) and PVS (0.052) at the reported precision, despite its small grounding gain. Matryoshka reduces D_{oracle} by 0.003 but also lowers PVS from 0.052 to 0.047. Neither variant produces a clear improvement in both splitting and visual specialization in this evaluation. Random BatchTopK has higher PVS (0.057) than trained BatchTopK (0.052), despite much poorer grounding: a higher PVS alone does not imply better object grounding.

Less feature splitting does not imply clearer specialization. PVS generally decreases as k increases (Figure 4c). Denser activation may allow redundant responses across features, leaving their object groups less visually distinct. From $k = 64$ to 256, splitting also decreases (Figure 4b), while grounding F1 changes little (Figure 4a). Less splitting therefore need not yield more visually distinct groups. This comparison shows why measuring the extent of splitting should be accompanied by examining what visual differences the splits represent.

5 Conclusion

PACO-SAE provides an object-level benchmark and protocol for evaluating concept grounding in vision SAEs as semantic segmentation. Our comparisons show that image-level probing and mSLA do not replace whole-object grounding evaluation. The benchmark also enables object-level analysis of feature splitting, while PVS and human judgments help assess the visual meaning of these splits without additional subclass labels. Grounding, the extent of splitting, and visual specialization should therefore be evaluated together. We hope PACO-SAE serves as a systematic benchmark for vision SAE evaluation and supports the mechanistic interpretability community in developing more reliable interpretations of vision models.

Limitations. Our experiments use one backbone, one layer, and one training seed per configuration, and do not evaluate causal effects on model behavior. PVS does not provide a threshold that reliably classifies individual splits as specialization or residual splitting.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback during research ideation, specifically on the framing of the evaluation methodology. We have not used generative AI tools for other tasks requiring disclosure, and theory- and proof-related disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to identify and summarize potentially relevant literature, draft and revise portions of the manuscript, and improve the clarity and readability of the writing. We have reviewed all AI-assisted work. Suggestions concerning the research framing were assessed by the authors, who made all final methodological decisions and conducted the experiments independently. Literature suggested by generative AI tools was checked against the original sources, and factual and numerical statements in AI-assisted text were verified against the authors' experimental records before inclusion. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our human evaluation asks annotators to compare groups of object images and judge whether the groups show a recurring visual difference. Appendix C describes the interface, response options, exclusion criteria, and aggregation rule used in the analysis.

REPRODUCIBILITY STATEMENT

Section 3 describes PACO-SAE, the held-out feature-selection protocol, concept grounding, feature splitting, and PVS, while Section 4 reports the evaluated models and experimental settings. Appendix B provides the dataset-construction rules, Appendix A specifies the concept-probing and mSLA implementations, and Appendix C documents the human-evaluation protocol and response aggregation.

REFERENCES

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.354. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Bau_Network_Dissection_Quantifying_CVPR_2017_paper.html.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Jack Merullo, Atticus Geiger, and Ekdeep Singh Lubana. Do sparse autoencoders capture concept manifolds? In *The 3rd Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=euoAoOflaN.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders, 2024. URL https://arxiv.org/abs/2412.06410.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6077–6101. PMLR, 2025. URL https://arxiv.org/abs/2503.17547.

David Chanin. Are sparse autoencoder benchmarks reliable?, 2026. URL https://arxiv.org/abs/2605.18229.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2755. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/764ff7477b8e24dbe01531f6791e8bdf-Abstract-Conference.html.

Jinho Choi, Hyesu Lim, Steffen Schneider, and Jaegul Choo. ConceptScope: Characterizing dataset bias via disentangled visual concepts. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0325. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0ddf14f20994636eeecc3d96fa8545cf-Abstract-Conference.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. doi: 10.48550/arXiv.2209.10652. URL https://arxiv.org/abs/2209.10652.

Ruth Fong and Andrea Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8730–8738, 2018. doi: 10.1109/CVPR.2018.00910. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Fong_Net2Vec_Quantifying_and_CVPR_2018_paper.html.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7ab63a5314680e2f083cb288abeaeb8e-Abstract-Conference.html.

Oscar Hill, Mateo Espinosa Zarlenga, and Mateja Jamnik. Hierarchical concept-based interpretable models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1395b425d06a50e42fafe91cf04f3a98-Abstract-Conference.html.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Haoran Jin, Xiting Wang, Shijie Ren, Hong Xie, and Defu Lian. C^2R: Cross-sample consistency regularization mitigates feature splitting and absorption in sparse autoencoders. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=VKFhtdShU0.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart Mcdougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264. PMLR, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html.

Jonas Klotz, Cassio F. Dantas, Pallavi Jain, Diego Marcos, and Begüm Demir. Evaluating the interpretability of sparse autoencoders with concept annotations, 2026. URL https://arxiv.org/abs/2606.24716. Accepted at ECCV 2026.

Alexey Kravets, Da Li, Chuan Li, Da Chen, and Vinay P. Namboodiri. Interpretability transfer from language to vision via sparse autoencoders. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2605.24946. URL https://arxiv.org/abs/2605.24946.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/84ca3f2d9d9bfca13f69b48ea63eb4a5-Abstract-Conference.html.

Weiduo Liao, Yunqiao Yang, and Ying Wei. When structured sparse autoencoders learn consistent concepts across modalities, 2026. URL https://arxiv.org/abs/2607.08605.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3d5b603d631d595f56bc36b373458b27-Abstract-Conference.html.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/53356aebeea8ffd40a8ac3bb66243162-Abstract-Conference.html.

Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. Transformer Circuits Thread, June 2022. URL https://www.transformer-circuits.pub/2022/mech-interp-essay.

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*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/89e83382abeee53b932a6df62edbf9cc-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7141–7151, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ramanathan_PACO_Parts_and_Attributes_of_Common_Objects_CVPR_2023_paper.html.

Maria L. Rizzo and Gábor J. Székely. DISCO analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055, 2010. doi: 10.1214/09-AOAS245.

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. *arXiv preprint arXiv:2508.10104*, 2025.

Samuel Stevens, Jacob Beattie, Tanya Berger-Wolf, and Yu Su. Towards open-ended visual scientific discovery with sparse autoencoders, 2025. URL https://arxiv.org/abs/2511.17735.

Kaichen Zhang, Yifei Shen, Bo Li, and Ziwei Liu. Large multi-modal models can interpret features in large multi-modal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3650–3661, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Large_Multi-modal_Models_Can_Interpret_Features_in_Large_Multi-modal_Models_ICCV_2025_paper.html.

A CONCEPT PROBING & MSLA

This section gives the implementation of the two baseline criteria $\mathcal{M}$ introduced in § 3.1.

Concept probing. For crop i, let P_i denote its non-padding patches. We pool the feature's activation into a crop-level score

$$s_{ij} = \max_{p \in P_i} \max(z_{ipj}, 0). \tag{9}$$

Max pooling lets a feature that fires on a small object still yield a high score. Positive crops are those whose anchor object belongs to c. Negative crops are those whose source image lists c among PACO-LVIS's verified negative categories (`neg_category_ids`). Crops with neither label are left out, because an unannotated category is not evidence that the concept is absent.

Several crops can share one source image, so an unweighted score would count crowded scenes more than once. We therefore give each crop the weight $w_i = 1/n_i$, where n_i is the number of crops in its class (positive or negative) that come from the same source image. Each source image then carries equal total weight within each class. Writing $\mathcal{P}_c$ and $\mathcal{N}_c$ for the positive and negative crops, the probing score is the weighted AUROC

$$\mathrm{AUROC}(j, c) = \frac{\sum_{i \in \mathcal{P}_c} \sum_{k \in \mathcal{N}_c} w_i w_k \left(\mathbf{1}[s_{ij} > s_{kj}] + \frac{1}{2} \mathbf{1}[s_{ij} = s_{kj}] \right)}{\sum_{i \in \mathcal{P}_c} w_i \sum_{k \in \mathcal{N}_c} w_k}. \tag{10}$$

Tied scores contribute half a pair. Ties are common because a sparse feature scores zero on most crops.

Mask-based Spatial Localization Accuracy (mSLA). We replace the bounding box in SLA (Kravets et al., 2026) with the target mask, since a bounding box includes background pixels around non-rectangular objects. For a positive crop i of concept c, let

$$p^*_{ij} = \arg\max_{p \in V_i} z_{ipj} \tag{11}$$

be the peak patch within the valid evaluation region V_i of Eq. 4. The crop counts as a hit when the feature fires at that patch and the patch is a target patch:

$$\mathrm{mSLA}(j, c) = \frac{1}{|\mathcal{P}_c|} \sum_{i \in \mathcal{P}_c} \mathbf{1}\left[z_{ip^*_{ij}j} > 0 \right] y^c_{ip^*_{ij}}. \tag{12}$$

A feature that does not fire anywhere in V_i counts as a miss rather than being dropped. A rarely active feature therefore cannot inflate its score by being evaluated on few crops. Restricting the argmax to V_i means that a peak on another instance of the same category is neither rewarded nor penalized.

Unlike VISTA, we do not match features to concepts through text descriptions. Each criterion selects its own feature directly from the annotations, so mSLA inherits SLA's localization test but not its description-based matching.

B PACO-SAE CONSTRUCTION

This section gives the rules used to derive PACO-SAE from PACO-LVIS (Ramanathan et al., 2023), in the order they are applied. Table A1 reports the resulting statistics.

Object retention. We rasterize each object's ground-truth mask at the source resolution and retain the object only when the mask covers at least 100 pixels. The threshold is applied to the rasterized mask rather than to the annotated area or the bounding-box area. Objects below it occupy too few pixels for the crop to remain legible after resizing. This step removes 35,956 of 217,117 train objects and 7,441 of 45,861 test objects.

Table A1: PACO-SAE statistics. PACO-LVIS columns count original images and object annotations; PACO-SAE columns count source images and retained anchor crops. Selection and test use the corresponding original splits. Parts are not evaluation targets.

	PACO-LVIS		PACO-SAE	
	Train	Test	Selection	Test
Source images	45,790	9,443	41,208	8,606
# Objects	217,117	45,861	127,974	27,110
# Concepts	75	75	63	63

Crop geometry. Let (w, h) be the object's bounding-box size and (W, H) the source image size. We set the crop side to $L = \min(2\max(w, h),\, W)$ when $w \geq h$ and $L = \min(2\max(w, h),\, H)$ otherwise, so the cap follows the object's longer axis and square boxes are capped by width. The crop is a square of side L centered on the bounding box. When the centered window falls outside the image, we first shift it back inside along that axis. Padding is applied only when L exceeds the image extent on an axis, in which case the window is centered and padded on that axis alone. Capping L at the image extent keeps the object from being scaled down relative to a smaller context window. Each crop is then resized to 224×224 pixels, with bilinear interpolation for the image and nearest-neighbor for the masks. Bounding boxes and areas are recomputed from the transformed masks rather than carried over from the source annotation.

Evaluation and target regions. Each crop stores two auxiliary masks. The *invalid* region records padded pixels, which carry no image content. The *ignore* region records the masks of other instances of the anchor category that fall inside the crop. Activation on those pixels is neither correct nor incorrect for the anchor object, since they depict the same concept on a different object. The valid evaluation region is $E = \neg(\text{invalid} \vee \text{ignore})$, and the target region Y is the anchor object's mask. These are the regions that define V_i and y^c_{ip} in Eq. 4.

Patch eligibility. Eligibility is evaluated on the 14×14 grid of 16×16 patches that tiles a 224×224 crop, matching the patch resolution of the evaluated ViT backbone. We discretize E and Y to this grid separately, labeling a patch positive when the region covers at least half its area ($\rho_E = \rho_Y = 0.5$). The number of target patches is then $n_Y = |Y \cap E|$. A crop is retained when $|E| > n_Y \geq 2$. The lower bound excludes objects too small to score at patch resolution, and the strict inequality guarantees at least one valid non-target patch so that activation outside the target can be evaluated.

Concept selection. We first drop crops whose anchor category is not exhaustively annotated in the source image, identified through PACO-LVIS's `not_exhaustive_category_ids`. Of 181,161 train crops, 136,128 are exhaustive and 128,787 also pass the patch criterion. The corresponding test counts are 28,804 and 27,325 of 38,420. We then retain a concept only when it has at least 100 such occurrences in the selection split and at least 30 in the test split. This leaves 63 of the 75 PACO-LVIS object concepts. Removing crops of the excluded concepts yields the final 127,974 selection and 27,110 test crops. Both thresholds are applied before any model is run, so concept selection does not depend on SAE activations. Object–part annotations are carried through the same crop transform but are not subject to eligibility filtering, since parts are not evaluation targets in our experiments.

C Human Evaluation

As shown in Figure A1, we present four randomly sampled images from each group side by side and ask annotators whether the two groups differ. Across the 200 pairs we collected 1,217 answers from 17 annotators, with at least 5 answers per pair. We first drop the 5 pairs that received 3 or more NA answers, leaving 195. Within each remaining pair we then discard the NA answers and count the pair as specialization when Yes outnumbers No, so the majority is taken over the decided answers only. Because a pair can receive an equal number of Yes and No answers, the 25 such ties are counted as non-specialization.

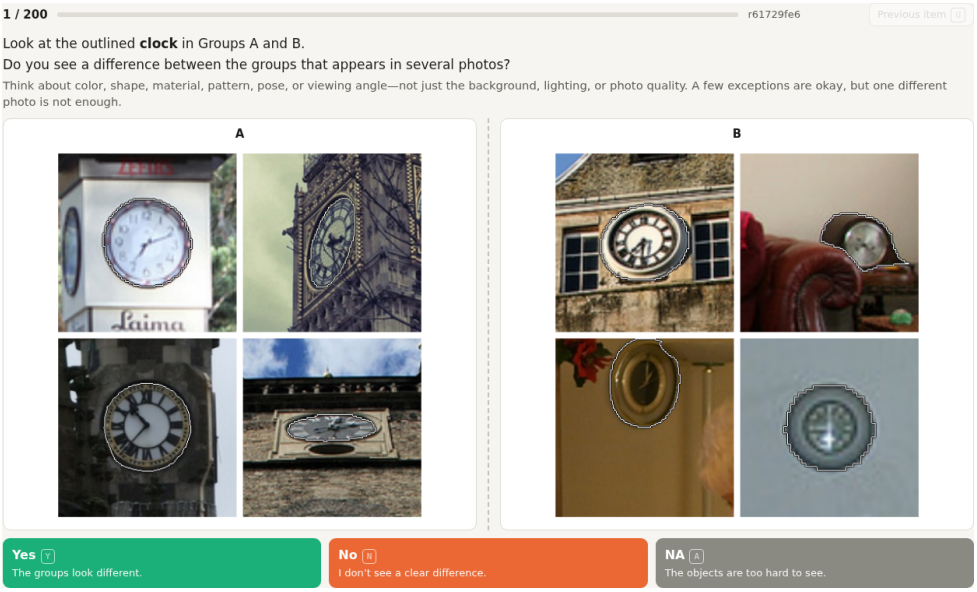


Figure A1: Group A and group B each show four images of one feature, with the target object outlined in white. The question and the three response options are identical across items.

D RECONSTRUCTION–SPARSITY FRONTIER

Figure A2 places the grounding analyses in the context of reconstruction quality. The horizontal axis reports the observed mean number of active features per patch on PACO-SAE test crops, rather than the nominal training value k shown next to each point. Larger values therefore indicate denser activation. Reconstruction error decreases within each SAE family as k increases, whereas the grounding results in Figure 4a show no corresponding consistent improvement.

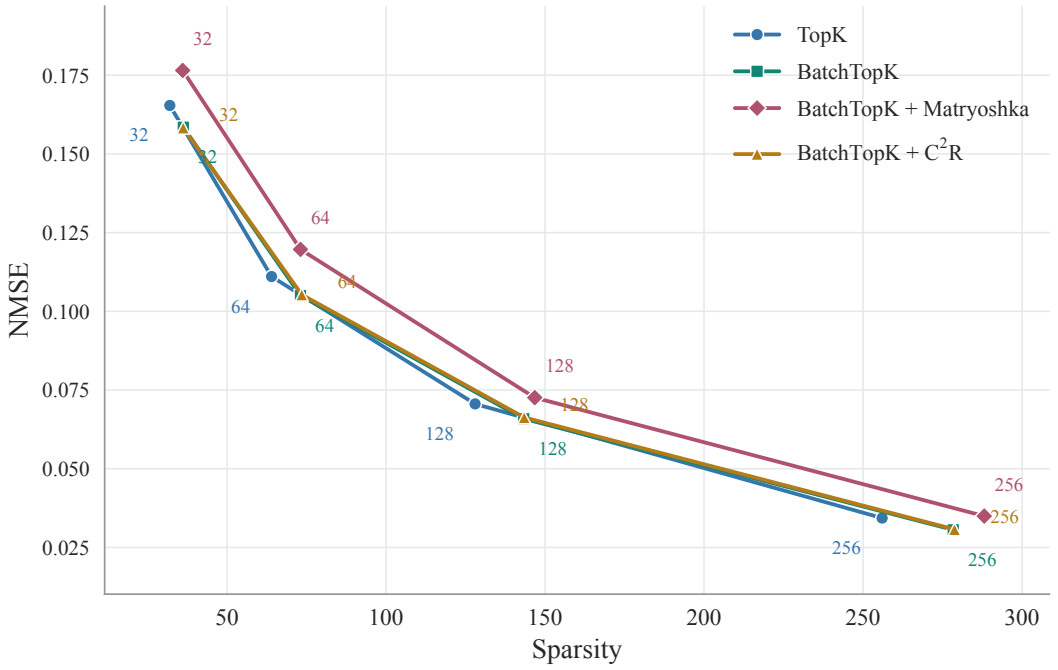


Figure A2: Reconstruction–sparsity frontier across the 16 trained SAEs. Normalized mean-squared reconstruction error (NMSE) is plotted against the observed mean number of active features per patch (the axis labeled "Sparsity"). Point labels indicate nominal training k; connected points belong to the same SAE family. Observed activation counts need not equal k.

E Paired Grounding Gaps Across Sparsity Levels

For each concept, we compare the test grounding F1 of the grounding-selected feature with that of the probing- or mSLA-selected feature:

$$\Delta_{\text{probe}}(c) = G_{\text{test}}(j_{\text{ground}}(c), c) - G_{\text{test}}(j_{\text{probe}}(c), c),$$
$$\Delta_{\text{mSLA}}(c) = G_{\text{test}}(j_{\text{ground}}(c), c) - G_{\text{test}}(j_{\text{mSLA}}(c), c). \tag{13}$$

Figure A3 averages these paired differences over the same 63 concepts. Both mean gaps are positive for every trained configuration. The intervals resample concepts with feature selections and model checkpoints fixed; they do not measure variation across training seeds or imply an improvement for every object.

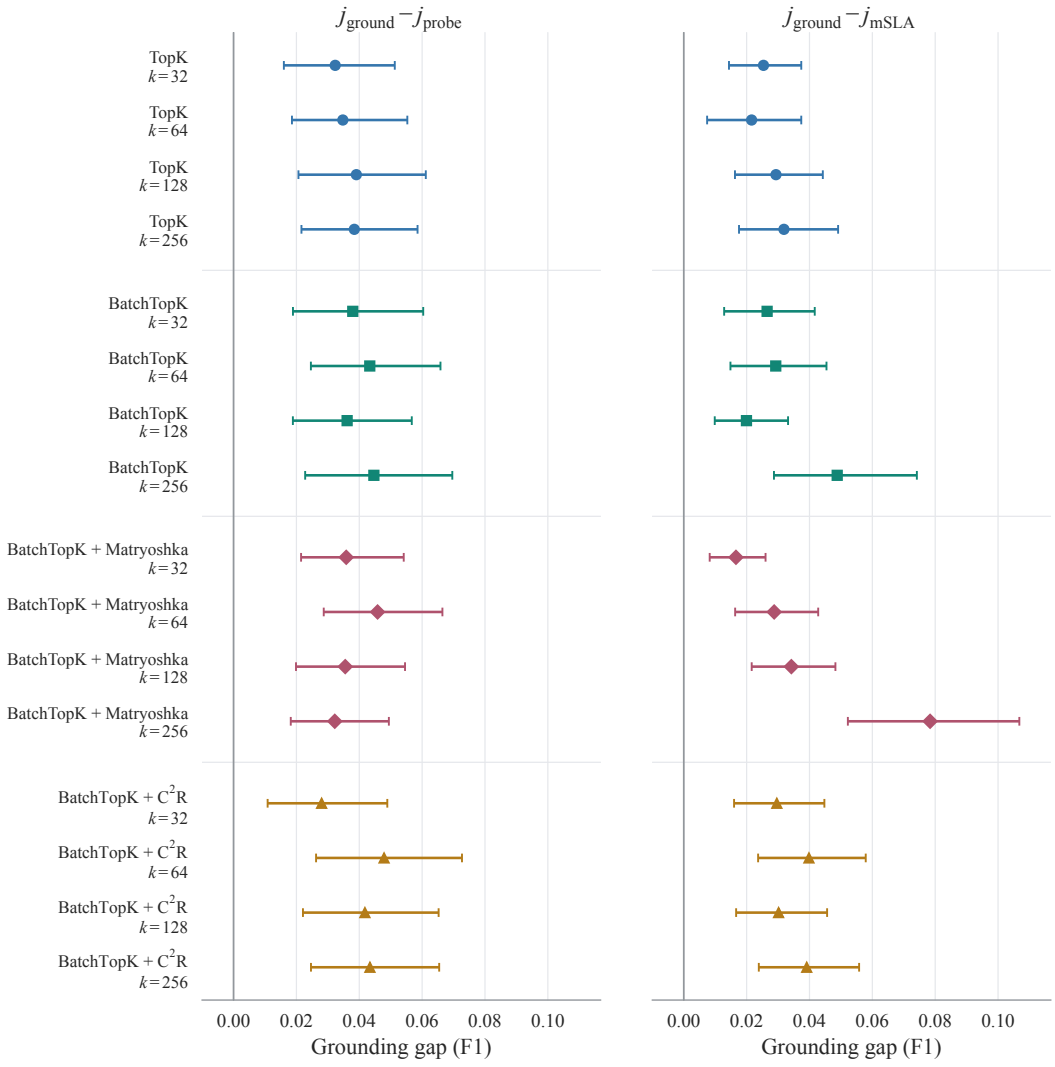


Figure A3: Paired test grounding-F1 gaps across all four SAE families and sparsity levels. Left: grounding-based versus probing-based selection. Right: grounding-based versus mSLA selection. Points are concept-macro differences over 63 concepts; bars show pointwise 95% concept-bootstrap intervals from 2,000 paired resamples. All comparisons use the same grounding F1 metric, not differences between AUROC, mSLA, and F1.

F Fixed-Feature and Oracle Grounding Across Sparsity Levels

Figure A4 extends the comparison in Table 2 to all 16 trained configurations. For each concept, the single feature and the 32 candidates are fixed using the selection set. G_{test} uses that single feature for every test object; G_{oracle} uses each object's best candidate as determined by its test mask. The latter remains an analysis upper bound, not a test-time selection method. The substantial gap persists across all four sparsity levels, showing that different candidates provide better grounding for different objects than one fixed feature provides.

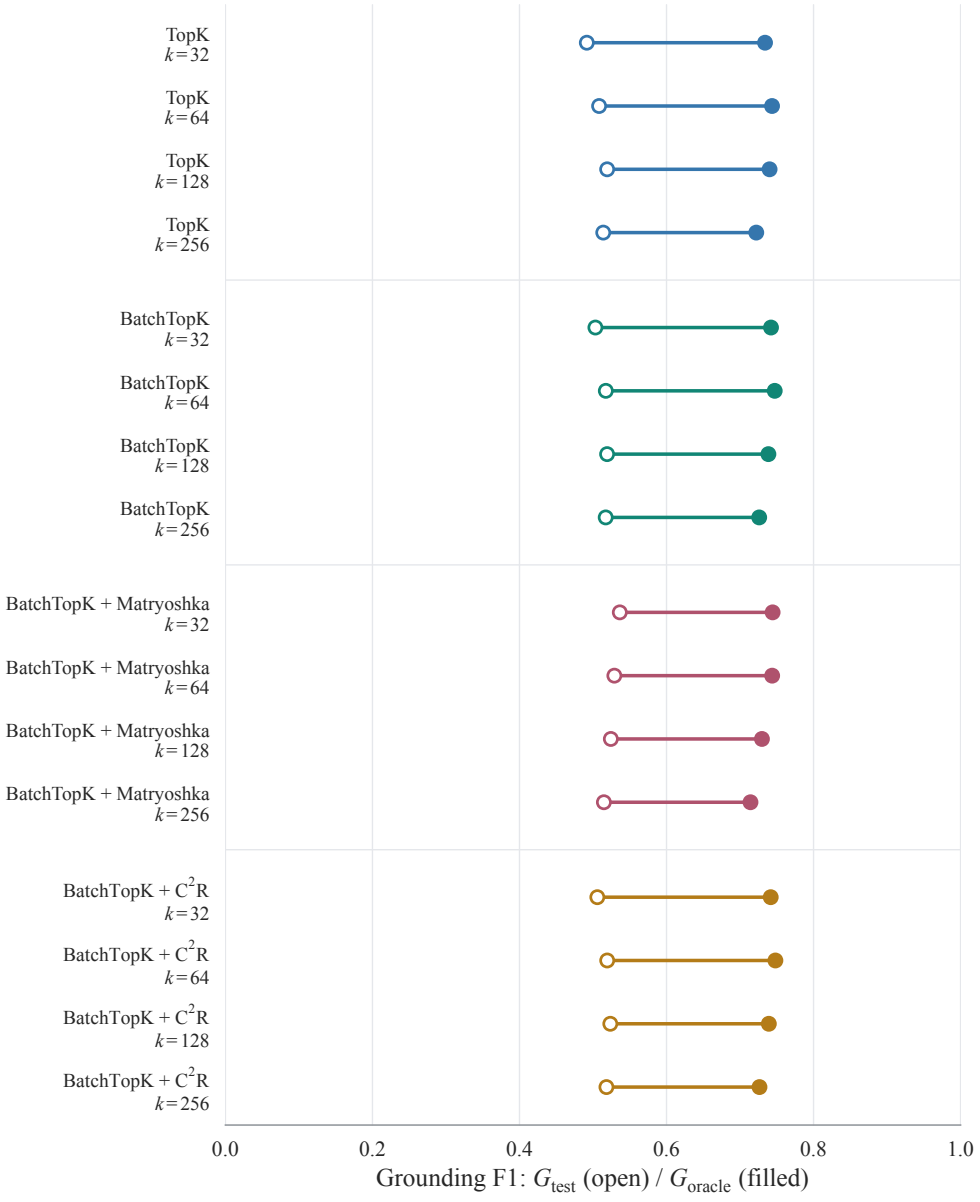


Figure A4: Fixed-feature and oracle grounding across all trained configurations. Open markers show G_{test} for the grounding-selected feature; filled markers show G_{oracle} using the best of 32 fixed candidates for each object. Lines connect the two scores for the same model and sparsity setting. Both scores are concept-macro means over the same 63 concepts and include all positive test objects.