RareCount: Interpretable Out-of-Distribution Detection in Sparse Concept Space

Antonella Coviello
Northeastern University, United States

Gabriele Onorato*
Northeastern University, United States

Nathaniel D. Bastian
United States Army

Francesco Restuccia
Northeastern University, United States

Abstract

Vision Foundation Models (VFMs) are increasingly deployed in critical applications where confident predictions on Out-of-Distribution (OOD) inputs can pose safety risks. Existing OOD detectors can flag unreliable inputs but do not explain what drives their decisions, thus limiting adoption in high-stakes scenarios. We introduce RareCount*, an interpretable, label- and head-free OOD detector that uses Sparse autoencoders (SAEs) to extract concepts. RareCount identifies OOD samples through unusually high activation of rare concepts relative to similar samples. Operating in concept space enables score decomposition and spatial grounding of the concepts responsible for detection. On OpenOOD, RareCount is competitive with post-hoc detectors, reduces the mean far-OOD FPR95 of the only other SAE-based detector by up to* 14.8 *points, and generalizes across DINOv2, ViT-MAE, DINOv3, and different sparsity levels. Local conditioning improves AUROC over global rarity by* 26.06 *points on near-OOD and* 34.75 *on far-OOD, while agreement, occlusion, and patch-atypicality analyses validate concept grounding, further supported by a controlled object-level demo in mixed In-Distribution (ID)/OOD scenes.*

1. Introduction

VFMs are increasingly used in critical applications such as medical imaging and autonomous driving [19, 25, 48]. In these settings, over-confident predictions can lead to unsafe decisions and wasted resources [16, 27]. This has motivated OOD detectors that identify potentially unreliable inputs and enable downstream systems to abstain or defer [37]. However, existing OOD detectors can be viewed as *opaque*, returning only anomaly scores or binary decisions without explaining which input properties drive them [14, 23, 42]. This makes it difficult to determine which input regions and features drive an OOD decision, limiting their use in high-stakes settings [7, 36].

*Corresponding author: onorato.g@northeastern.edu

Key Issue. An interpretable OOD detector requires both a decomposable score and interpretable underlying representations composed of individually meaningful features [2, 18]. This is challenging to satisfy with dense VFM activations, where *superposition* can induce *polysemanticity*, i.e., an effect where neurons respond to unrelated patterns [10, 32]. SAEs address this by decomposing dense activations into sparse learned features [2, 18, 43]. Following prior work [1, 22], we refer to the learned sparse features as *concepts*. Thus, deriving the OOD score directly from these concepts enables attribution to the specific concepts responsible for detection.

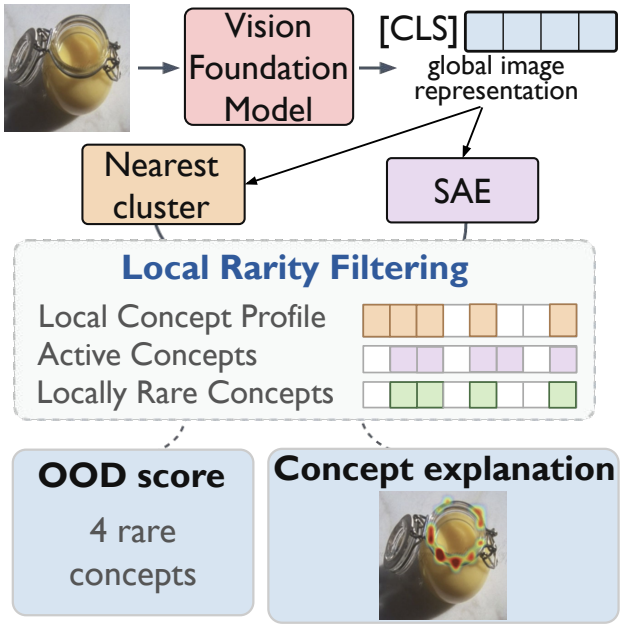


Figure 1. A test image is encoded by a frozen VFM, and its global [CLS] representation is used both to select the nearest cluster. Local rarity filtering compares the active concepts with the concept profile of the selected cluster and identifies the concepts that are active in the image but rare in its local neighborhood. The number of these locally rare concepts defines the OOD score, while their spatial grounding provides a concept-level explanation.

Proposed Approach. We propose *RareCount* (overviewed in Fig. 1), an OOD detector that scores an input by the fraction of its active concepts that are rare within its assigned ID region. By measuring rarity relative to semantically similar ID samples rather than globally, RareCount accounts for concepts that may be infrequent overall but common within a specific region of ID samples. Thus, removing this local conditioning substantially reduces detection performance (Sec. 4.4). Because the score is computed directly from per-concept rarity indicators, it also exposes the concepts responsible for each detection and allows them to be localized in the image (Fig. 3). RareCount turns an opaque decision, i.e., *"this image is anomalous"* into an interpretable one, i.e., *"this image is anomalous because it activates concepts x, y, and z, localized in these regions."*

Summary of Novel Contributions

❶ Label- and head-free OOD detection in concept space. We introduce RareCount, a label- and head-free OOD detector that operates directly in the sparse concept space obtained from frozen VFM activations using a SAE. The SAE and local reference statistics are learned exclusively from unlabeled ID data, without labels or class-conditioned statistics.

❷ Detection and explanation from the same representation. The RareCount score is directly decomposable into the concepts responsible for each decision, satisfying both the requirement of score decomposability and that of interpretability of the underlying representation, making the explanation faithful by design, rather than relying on a post-hoc attribution method. We analyze the concepts into which each decision decomposes at both the per-sample and dataset levels, together with their spatial grounding in the input image through patch-level activations, in Section 4.5.

❸ Competitive and generalizable post-hoc detection. RareCount is competitive with established post-hoc OOD detectors and reduces far-OOD FPR95 by up to 14.8 points relative to EPD [28], the only other SAE-based method, despite removing its class labels and classifier head. We further show that the detection signal also persists across TopK, BatchTopK, Gated, and ReLU SAEs and across DINOv2, ViT-MAE, and DINOv3, showing that local concept rarity does not depend on a particular sparsity level or on a specific backbone. We analyze detection performance in Section 4.1 and generalization in Section 4.2.

2. Related Work

Vision foundation models. Self-supervised approaches include masked reconstruction methods, e.g., ViT-MAE [12] and self-distillation methods such as DINOv1 [5], DINOv2 [30] and DINOv3 [39]. These pretrained representations can be reused by downstream methods while keeping the backbone frozen, which makes them particularly suitable for post-hoc analysis. RareCount builds on this by keeping the VFM frozen and performing OOD detection directly in a sparse decomposition of its representation space.

Sparse autoencoders. SAEs decompose model activations into sparse, approximately monosemantic features [18, 43], following the linear representation hypothesis [32], with major SAE variants differing mainly in how sparsity is enforced [3, 11, 33]. Recent work has extended SAEs to visual representations, showing behaviorally relevant and semantically meaningful features as well as improved robustness under distribution shift [29, 31, 40]. For OOD detection, EPD [28] applies a TopK SAE [3] to ViT CLS activations and scores class-conditional activation-profile divergence. In contrast, RareCount is label-/classifier-free and scores the local rarity of active concepts over the full sparse code, thus identifying the concepts responsible for each OOD decision.

Out-of-distribution detection. OOD methods are broadly divided into training-time and post-hoc approaches. The former modify the model or objective through auxiliary outliers, synthetic outliers, confidence regularization, or representation shaping [9, 15, 26, 44], making them unsuitable when the foundation model must remain frozen [20]. Post-hoc methods instead score outputs, activations, gradients, or representation geometry [6, 14, 17, 21, 23, 35, 41, 42]. Most require a classifier head or labeled ID data. While KNN avoids both, it returns only a scalar distance. Conversely, *RareCount* operates in the sparse concept space of a frozen backbone, requires neither labels nor a classifier, and decomposes each OOD score into the relevant concepts.

3. The RareCount OOD Detection Framework

Problem setting. We consider the problem of *semantic OOD detection*, i.e., identifying inputs whose semantic content differs from that of the reference data [45, 46]. In-distribution (ID) samples follow the distribution $\mathcal{P}_{\text{in}}$ of the reference dataset while OOD samples come from an unknown distribution $\mathcal{P}_{\text{out}}$ that contains semantic categories that are outside the ID label set. A detector assigns each input an anomaly score $s(x)$, with larger values indicating higher evidence of being OOD. RareCount operates in a *class-free* setting and as such, it does not require labels or a classifier head.

Representation. We use a frozen model f and, following standard VFM practice, extract the final block [CLS] embedding as the global image representation, $z = f(x) \in \mathbb{R}^d$ with $d = 768$, together with the grid of patch-token embeddings of the same block [8, 30]. Following feature-space OOD practice [42], we use the normalized representation $\tilde{z} = z/\|z\|_2$, for both SAE encoding and feature-space partitioning. This normalized CLS representation is the common input to the two branches of our pipeline, i.e., feature-space partitioning for local neighborhood selection and SAE encoding for concept extraction. Patch-token embeddings are not used to compute the OOD score and are only employed for the spatial grounding analysis in Section 4.5.

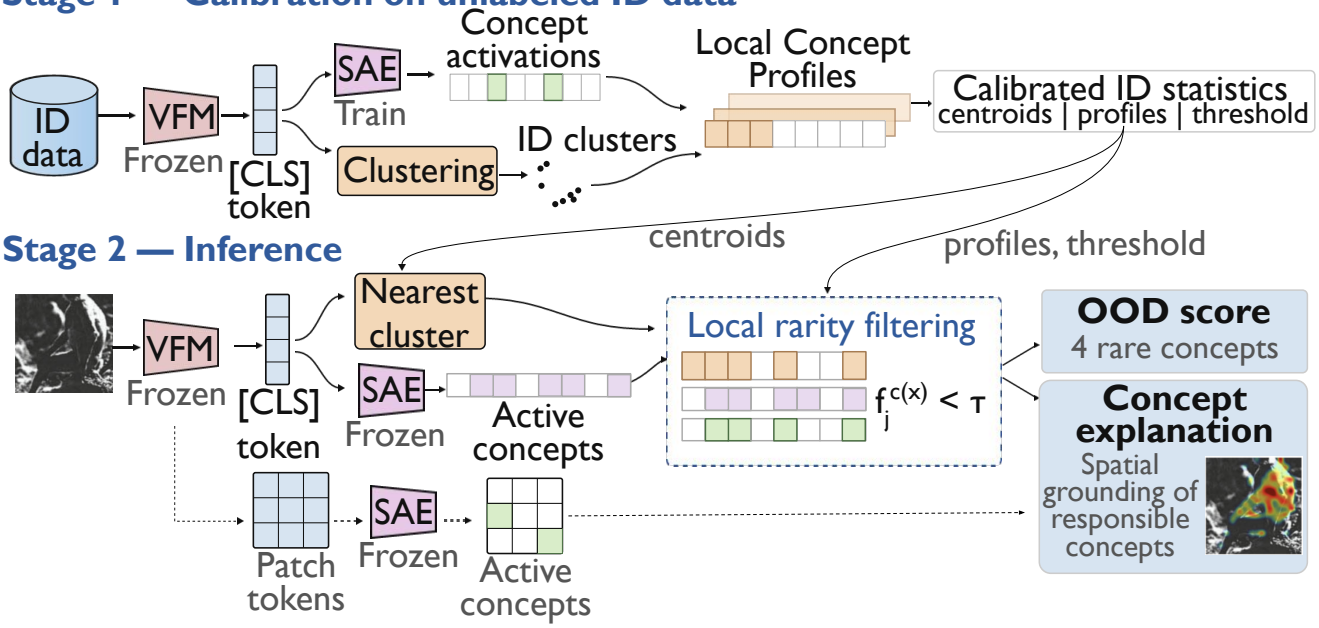


Figure 2. RareCount pipeline. The input image is encoded using a frozen vision foundation model and projected into a sparse concept space through a SAE. The activated concepts are compared with their frequencies in the local region of the in-distribution feature space to compute the OOD score and identify the concepts responsible for the detection.

SAE and Concepts. A sparse autoencoder reconstructs $\tilde{z}$ through a wide and sparsely activated bottleneck [18, 22, 31, 33], an encoder produces

$$a(\tilde{z}) \;=\; \sigma(W_{\text{enc}}(\tilde{z} - b_{\text{pre}}) + b_{\text{enc}}) \;\in\; \mathbb{R}^m,\; m \gg d\,, \quad (1)$$

and a decoder reconstructs $\hat{z} = W_{\text{dec}}\, a(\tilde{z}) + b_{\text{pre}}$, with the columns of W_{dec} constrained to unit norm. The four SAE families we study differ in the sparsifying nonlinearity σ, which leaves only a small fraction of the m latent coordinates nonzero [4, 18, 33, 34]. In line with existing work [18, 22], we refer to each latent coordinate as a *concept*. Each SAE family defines a firing rule $\phi_j(\tilde{z}) \in \{0, 1\}$, yielding

$$\mathcal{A}(x) = \{j : \phi_j(\tilde{z}) = 1\}\,, \quad (2)$$

defined as the *active set* [18]. RareCount depends only on concept presence through $\mathcal{A}(x)$. During the offline stage, the SAE is trained on normalized CLS representations from unlabeled ID data and then frozen. During inference, the same encoder maps each test representation to $\mathcal{A}(x)$.

3.1. Cluster-Conditioned Concept Rarity

Figure 2 shows the two stages of RareCount: ❶ an offline reference construction stage performed once on unlabeled ID data, and ❷ an inference stage that assigns each test sample to a local ID region and evaluates the "rarity" of its active concepts.

Building the reference. We characterize how concepts are typically used by ID data in different regions of the representation space. Let $\mathcal{D}$ denote the ID training set. We partition its ℓ_2-normalized CLS embeddings into K clusters with MiniBatch k-means [38], with centroids $\{\mu_1, \ldots, \mu_K\}$. Each input is assigned to the nearest centroid, $c(x) = \arg\min_k \|\tilde{z} - \mu_k\|_2$. For the subset of training samples assigned to cluster k, $\mathcal{D}_k = \{\, x \in \mathcal{D} : c(x) = k \,\}$ we define the activation frequency of concept j in region k as

$$f_j^k = \frac{1}{|\mathcal{D}_k|} \sum_{x \in \mathcal{D}_k} \phi_j(\tilde{z}(x))\,, \quad (3)$$

where $\phi_j(\tilde{z}(x))$ indicates whether concept j is active for input x. Collecting these statistics yields a frequency matrix $F \in [0, 1]^{K \times m}$, with $F_{kj} = f_j^k$, where the k-th row represents the empirical concept profile of the corresponding ID region. We denote this row as $F_k = (f_1^k, \ldots, f_m^k)$ and refer to it as the local concept profile of cluster k. The centroids $\mu_k{}_{k=1}^K$ and local concept profiles F constitute the ID reference statistics computed offline and stored for inference. The rarity threshold τ is selected separately on validation data as described in Section 3.2 for standard comparison with the literature, while we show that it can be selected via ID cluster–concept frequency statistics only in Appendix B.2.

Scoring. At inference, the normalized CLS representation $\tilde{z}$ is processed by two frozen branches shown in Fig. 2. The input sample x is assigned to its nearest reference cluster $c(x)$ which retrieves the corresponding local concept profile $F_{c(x)}$, while the frozen SAE produces the active concept set $\mathcal{A}(x)$. RareCount is then computed using the precomputed frequency profile $F_{c(x)}$, without requiring access to the ID

training samples at test time. An active concept $j \in \mathcal{A}(x)$ is considered *rare* if it is infrequently observed among ID samples in the same region, i.e., $f_j^{c(x)} < \tau$, where τ is a global rarity threshold. We refer to this comparison between the test sample active set and its retrieved local concept profile as *local rarity filtering*. The filtering step retains exactly the active concepts that violate the local ID profile. RareCount is the number of rare concepts active for input x,

$$\text{RareCount}(x) = \sum_{j \in \mathcal{A}(x)} \mathbf{1}\left[f_j^{c(x)} < \tau \right]. \quad (4)$$

As such, a high RareCount indicates that the input activates many concepts that are uncommon in its local ID context. Since different SAE families might activate different numbers of concepts, we use a normalized version when comparing across architectures

$$\text{RareFraction}(x) = \frac{\text{RareCount}(x)}{|\mathcal{A}(x)|}, \quad (5)$$

RareFraction measures the proportion of active concepts that are rare in the input's assigned region, making the score comparable across SAEs with different sparsity levels. For TopK SAE $|\mathcal{A}(x)|$ is fixed by construction, so RareCount and RareFraction differ only by a constant scaling factor.

Responsible concepts. Since RareCount Eq. (4) is defined as a sum of concepts, we can identify which concepts contributed to the score and define the responsible set as $\mathcal{R}(x) = \{j \in \mathcal{A}(x) : f_j^{c(x)} < \tau\}$. By construction, $\text{RareCount}(x) = |\mathcal{R}(x)|$ yielding an exact decomposition of the detection scores into the responsible concepts without requiring a post-hoc attribution method. For spatial grounding, we apply the same SAE trained on the CLS to the patch-token representations, the validity of this CLS-patch transfer is evaluated in Sec. 4.5 and Appendix H. For each responsible concept $j \in \mathcal{R}(x)$, its patch-level activations are arranged over the token grid to obtain a concept-specific localization map. Thus, the CLS representation determines *which* concepts are responsible for the detection, while the patch representations determine *where* those concepts are expressed in the image. We evaluate their interpretability at sample, dataset and spatial levels in Section 4.5.

3.2. Experimental Setup

Benchmark. We follow the OpenOOD protocol with ImageNet-1K as the in-distribution dataset [45, 47]. We evaluate on the near-OOD sources SSB-Hard and NINCO, and on the far-OOD sources iNaturalist, Textures, and OpenImage-O. ImageNet-C [13] is used to evaluate sensitivity to covariate shift, considering all corruption types and severity levels.

Backbones and SAEs. Our primary backbone is a frozen DINOv2 ViT-B/14 [30]; Sec. 4.2 evaluates transfer to ViT-MAE ViT-B/16 [12] and DINOv3 ViT-B/16 [39]. For each backbone, we use its publicly released pretrained checkpoints. ViT-MAE is pretrained in a self-supervised manner on ImageNet-1K, so in this setting the pretraining data coincide with the downstream ID reference distribution. DINOv2 and DINOv3 instead use larger curated pretraining datasets, LVD-142M and LVD-1689M, respectively. For these models, following the OpenOOD evaluation protocol [45, 47], which explicitly evaluates DINOv2 for OOD detection with ImageNet-1K as the ID reference distribution, we define the ID/OOD distinction with respect to the downstream ImageNet-1K reference distribution rather than the backbone pretraining distribution. All backbones remain frozen throughout our experiments, and RareCount operates directly on their CLS representations. SAEs are trained only on the CLS activations of the unlabeled ID training split, without OOD data. All SAE families use a matched dictionary of $m = 7680$ concepts ($10 \times d$), so that their comparison isolates the sparsity mechanism rather than dictionary capacity.

We consider four SAE families, whose firing rules define the active set in Eq. (2): TopK retains the $k = 128$ largest pre-activations; BatchTopK uses a global firing threshold calibrated on ID activations after training; Gated requires both an open gate and a positive magnitude; and ReLU fires whenever $a_j > 0$. Appendix C.1 ablates both the dictionary width m and the TopK activation budget k, showing that RareCount remains consistent across alternative dictionary capacities and sparsity levels. TopK, BatchTopK, and Gated operate at comparable sparsity, whereas ReLU is kept at its native, denser operating point.

Hyperparameters. They are the number of clusters K and the *rarity threshold* τ. We select both on the OpenOOD validation splits and freeze them before evaluating the held-out test sets. We use $K = 1000$, which controls only the granularity of the unsupervised partition and does not rely on ImageNet class labels, despite coinciding with the number of ImageNet-1K classes. The threshold τ is selected from the observed validation frequencies using a prespecified criterion balancing AUROC and FPR95. RareCount does not fundamentally require OOD data for threshold calibration: in Appendix B.2, we show that τ can instead be estimated exclusively from ID cluster–concept frequency statistics, while retaining most of the detection performance. Robustness to clustering algorithms, values of K, random seeds, alternative threshold-selection rules, and the joint (K, τ) grid is reported in Appendix B.

Baselines. We compare against post-hoc detectors spanning the three families of Sec. 2: output-based (MSP, TempScale, EBO, GEN), activation- and gradient-based (ReAct, ASH, GradNorm) and feature-based (MDS, RMDS, KNN), together with EPD [28]. All baselines use the same frozen backbone and the OpenOOD reference implementations. For output-, activation- and gradient- based scores that require

classifier predictions, we use the official ImageNet-1K linear classifier associated with each backbone (see Appendix A.1 for further details on the specific backbone models and classifiers used). Following the OpenOOD evaluation protocol, deterministic post-hoc baselines evaluated on a fixed backbone are reported as a single score, while mean and standard deviation are reported only for EPD and our variants, for which we can retrain the SAE across multiple runs [47].

Metrics. In line with existing literature, we compute the False Positive Rate at 95% true positive rate (FPR95) and Area Under Receiver Operating Curve (AUROC). A lower FPR95 and a higher AUROC indicate better performance.

4. Experimental Results

4.1. Benchmark Comparison

Table 1 reports detection performance on DINOv2. Three main observations emerge.

❶ **RareCount outperforms the closest SAE-based baseline despite removing its supervision requirements.** On far-OOD, TopK reduces mean FPR95 from 34.85 for EPD [28] to 22.89, while ReLU lowers it to 20.01; mean AUROC increases from 93.87 to 94.88 and 95.31, respectively. This improvement is obtained without the class labels and classifier head required by EPD. Since both approaches operate on the same frozen backbone and comparable SAE representations, the gain reflects the effectiveness of local concept rarity as a scoring mechanism.

❷ **RareCount is competitive with established post-hoc detectors under weaker assumptions.** Across near- and far-OOD sources, RareCount remains competitive with the strongest feature- and output-based baselines while requiring neither class annotations nor a trained classifier. Methods that achieve better results on selected metrics, such as MDS, rely on class-conditional statistics, whereas comparable label- and head-free alternatives such as KNN provide only a scalar distance. RareCount therefore offers a favorable trade-off between detection performance, supervision requirements, and interpretability.

❸ **RareCount does not depend on hard sparsity.** This is a key distinction from EPD, which is formulated around hard-sparse TopK representations. RareCount remains stable across TopK, BatchTopK, Gated, and ReLU SAEs, whose AUROCs differ by at most 0.18 points on near-OOD and 0.51 points on far-OOD. Most notably, ReLU activates 848 concepts per input, compared with 128 for TopK, yet achieves the strongest far-OOD result among our variants. Thus, local concept rarity is preserved across different sparsity regimes and does not require an activation budget.

4.2. Generalization Across Backbones

Table 2 reports detection performance on ViT-MAE and Tab. 3 evaluates RareCount across DINOv2, ViT-MAE, and DINOv3. Three main observations emerge. We evaluate standard OOD baselines across backbones, while EPD [28] serves as the closest SAE-specific baseline in the DINOv2 setting (Tab. 1) considered in its original work.

❶ **Local concept rarity transfers across different backbones.** RareCount is effective on all three representation spaces, with mean far-OOD AUROC ranging from 92.38 to 95.31 and near-OOD AUROC from 76.33 to 80.93 across TopK and ReLU. ViT-MAE provides the strongest test of this claim, since it is pretrained through masked reconstruction rather than the self-distillation objective used by the DINO family. The signal is therefore not specific to DINOv2 or to a particular pretraining objective.

❷ **RareCount becomes comparatively stronger on ViT-MAE.** As shown in Tab. 2, Gated and ReLU achieve the highest mean AUROC among all evaluated methods: ReLU ranks first on near-OOD (79.95) with Gated second (79.84) while Gated ranks first on far-OOD (92.99), with ReLU second (92.91). This is a stronger relative performance than on DINOv2, where supervised and output-based baselines remain preferable on selected metrics. Local concept rarity therefore remains effective even when classifier-based scores degrade under a different representation.

❸ **RareCount is more stable across backbones than classifier-dependent detectors.** From DINOv2 to ViT-MAE, far-OOD AUROC decreases by only 2.4–2.5 points across TopK and ReLU, whereas standard classifier-dependent baselines such as MSP, TempScale, EBO, GEN, and ReAct lose between 6.8 and 15.0 points. These results indicate that RareCount inherits the transferability of backbone representations while remaining independent of the classifier head.

4.3. Robustness to Covariate Shift

We evaluate RareCount under covariate shift using ImageNet-C with a TopK SAE and DINOv2 backbone. With the operating threshold fixed on clean ImageNet, the fraction of samples flagged as OOD increases from 4.85% on clean data to 43.07% at severity 5, while the mean number of rare concepts rises from 27.65 to 58.94. RareCount remains competitive with standard post-hoc detectors under mild-to-moderate corruptions, although feature-based baselines such as MDS are more robust at the highest severities. Full per-corruption and baseline results are reported in Appendix E.

4.4. Ablations and SAE Analysis

Two random ID images typically share few active concepts, making most concepts globally infrequent and therefore limiting the usefulness of global frequency as a measure of rarity. In contrast, ID images represented similarly by the backbone tend to share many of the same concepts. Cluster conditioning exploits this structure by defining rarity relative to ID samples with similar representations. OOD inputs can then be detected when they violate this local expectation,

Table 1. Post-hoc OOD detection on the OpenOOD benchmark with ImageNet-1K as in-distribution and DINOv2 (ViT-B/14) as backbone. Results for our SAE variants are reported as mean ± standard deviation averaged over ten random seeds for the TopK SAE and five random seeds for each of the other SAE variants; rankings are based on the mean performance. Colors indicate the **first**, **second**, and **third** best result in each column, respectively.

	Near-OOD						**Far-OOD**							
	SSB-Hard		NINCO		*Avg.*		iNaturalist		Textures		OpenImage-O		*Avg.*	
Method	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
MSP	75.43	74.23	47.19	86.10	61.31	80.17	22.91	94.82	52.76	86.41	35.52	90.93	37.07	90.72
TempScale	73.85	75.65	44.62	87.85	59.24	81.75	13.56	96.66	48.73	88.57	28.82	93.23	30.37	92.82
EBO	73.23	**77.02**	**41.64**	87.84	**57.44**	**82.43**	4.82	98.64	33.66	91.94	**16.92**	**95.68**	**18.47**	**95.42**
GEN	73.11	**77.04**	**41.12**	88.04	**57.12**	**82.54**	4.89	98.59	33.54	91.93	**16.63**	**95.70**	**18.36**	95.41
ReAct	76.05	74.20	46.26	86.21	61.16	80.21	7.29	97.82	32.92	91.84	18.68	95.13	19.63	94.93
ASH	95.57	48.92	98.15	42.77	96.86	45.85	97.70	40.85	99.90	31.61	97.68	45.93	98.43	39.46
GradNorm	95.40	51.13	90.32	63.70	92.86	57.42	33.57	92.56	91.54	64.43	64.82	82.21	63.31	79.73
kNN	82.39	73.11	53.29	87.01	67.84	80.06	**2.52**	**99.25**	**32.90**	**92.59**	21.93	95.47	19.12	**95.77**
MDS	77.47	**78.92**	**41.20**	**90.44**	59.33	**84.68**	**1.02**	**99.66**	**27.08**	**94.12**	**16.33**	**96.47**	**14.81**	**96.75**
RMDS	84.28	74.17	55.76	**88.47**	70.02	81.32	3.12	98.84	39.23	90.72	19.15	95.67	20.50	95.08
EPD	**68.13** ±0.25	75.95 ±0.07	50.90 ±0.82	**88.37** ±0.22	59.51 ±0.42	82.16 ±0.12	21.34 ±1.71	96.84 ±0.17	48.04 ±0.91	90.37 ±0.15	35.16 ±0.89	94.39 ±0.10	34.85 ±1.02	93.87 ±0.12
Ours														
TopK	69.39 ±0.90	74.49 ±0.17	48.85 ±1.19	87.38 ±0.22	59.12 ±0.68	80.93 ±0.14	8.14 ±0.57	98.14 ±0.08	35.25 ±0.98	91.57 ±0.10	25.28 ±0.54	94.93 ±0.06	22.89 ±0.58	94.88 ±0.07
BatchTopK	70.34 ±0.31	74.48 ±0.13	49.44 ±0.57	87.52 ±0.17	59.89 ±0.19	81.00 ±0.14	8.14 ±0.48	98.19 ±0.08	35.88 ±0.70	91.60 ±0.14	25.97 ±0.23	94.98 ±0.04	23.33 ±0.32	94.92 ±0.04
Gated	70.56 ±0.80	74.39 ±0.41	50.48 ±0.88	87.45 ±0.17	60.52 ±0.24	80.92 ±0.13	8.30 ±0.46	98.13 ±0.07	36.69 ±1.44	91.44 ±0.50	26.50 ±0.56	94.83 ±0.15	23.83 ±0.68	94.80 ±0.22
ReLU	70.55 ±0.82	74.30 ±0.38	47.67 ±0.53	87.33 ±0.30	59.11 ±0.58	80.82 ±0.20	3.05 ±1.54	98.87 ±0.36	32.73 ±1.47	91.95 ±0.38	24.24 ±0.35	95.10 ±0.05	20.01 ±0.73	95.31 ±0.19

Table 2. Post-hoc OOD detection on the OpenOOD benchmark with ImageNet-1K as in-distribution and ViT-MAE (ViT-B/16) as backbone. Results for our SAE variants are reported as mean ± standard deviation over 3 SAE seeds; rankings are based on the mean performance. Colors indicate the **first**, **second**, and **third** best result in each column, respectively.

	Near-OOD						**Far-OOD**							
	SSB-Hard		NINCO		*Avg.*		iNaturalist		Textures		OpenImage-O		*Avg.*	
Method	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
MSP	**78.14**	71.92	64.92	79.49	71.53	75.70	46.98	84.62	62.26	81.11	55.04	82.51	54.76	82.75
TempScale	**79.26**	72.75	69.76	80.35	74.51	76.55	49.92	86.97	70.04	82.32	64.43	83.51	61.46	84.27
EBO	85.43	71.03	83.11	75.41	84.27	73.22	60.60	86.94	82.72	76.72	81.63	77.52	74.98	80.40
GEN	**77.96**	**74.88**	58.00	84.43	**67.98**	**79.65**	36.39	90.80	56.98	86.74	47.03	88.37	46.80	88.64
ReAct	81.72	72.65	72.69	82.32	77.20	77.48	37.26	92.55	69.67	85.06	61.90	86.81	56.28	88.14
ASH	96.75	49.92	96.75	49.63	96.75	49.78	96.75	50.05	96.75	49.30	96.75	49.41	96.75	49.59
GradNorm	95.87	37.52	97.11	23.84	96.49	30.68	97.18	14.78	97.36	17.94	98.31	14.05	97.62	15.59
kNN	84.18	66.57	**55.05**	82.58	69.61	74.58	**24.26**	91.53	**33.01**	**91.24**	**29.39**	90.94	**28.88**	91.23
MDS	82.84	69.88	**50.51**	84.18	**66.67**	77.03	**19.12**	94.07	**34.84**	89.74	**26.21**	91.82	**26.72**	91.88
RMDS	81.85	72.59	**46.26**	**86.39**	**64.05**	79.49	**17.68**	**95.32**	**35.98**	89.99	**24.80**	**92.77**	**26.15**	**92.69**
Ours														
TopK	80.35 ±1.01	72.01 ±0.72	63.67 ±1.12	85.87 ±0.43	72.01 ±0.24	78.94 ±0.57	37.01 ±9.38	93.88 ±0.38	49.37 ±3.51	90.51 ±0.31	42.61 ±2.18	92.75 ±0.06	43.00 ±5.02	92.38 ±0.21
BatchTopK	81.16 ±1.11	71.92 ±0.85	67.82 ±0.27	85.39 ±0.62	74.49 ±0.49	78.66 ±0.73	52.89 ±6.23	92.49 ±0.62	57.58 ±0.55	89.28 ±0.20	50.01 ±0.70	91.80 ±0.25	53.50 ±2.13	91.19 ±0.24
Gated	79.48 ±0.56	72.96 ±0.40	61.94 ±1.03	86.72 ±0.27	70.71 ±0.31	79.84 ±0.18	27.16 ±0.67	95.02 ±0.17	46.09 ±1.42	90.94 ±0.21	40.20 ±0.72	93.02 ±0.13	37.82 ±0.43	92.99 ±0.14
ReLU	79.40 ±0.38	73.01 ±0.34	62.81 ±0.68	86.90 ±0.47	71.10 ±0.22	79.95 ±0.09	27.22 ±0.61	94.81 ±0.16	46.94 ±1.72	90.85 ±0.04	40.42 ±0.99	93.06 ±0.12	38.19 ±0.45	92.91 ±0.04

activating concepts that are atypical for their assigned region. For a fair comparison the rarity threshold τ is selected independently for the local and global variants on the validation split using the same selection criterion and is then frozen for test evaluation. As shown in Tab. 4, local conditioning improves AUROC by 26.06 points on near-OOD and 34.75 points on far-OOD, while reducing FPR95 on all datasets.

SAE Diagnostics. Table 5 shows that all four SAE families learn well-formed dictionaries, with no dead or near-dead features and no measured redundancy. ReLU achieves the best reconstruction metrics and the lowest Gini coefficient, indicating a more uniform use of the dictionary.

Robustness across SAE regimes. Detection performance is not explained by poor reconstruction or code density. ReLU combines the best reconstruction quality with the highest far-OOD AUROC, while TopK reconstructs less accurately but performs only 0.43 AUROC points worse. The architectures span a 6.6× range in ℓ_0, from 128 for TopK to 848 for ReLU,

Table 3. **Cross-backbone generalization.** Mean ± standard deviation of FPR95 (↓) and AUROC (↑) on OpenOOD across ten random seeds for the TopK SAE and five random seeds for each of the other DINOv2 SAE variants and three SAE seeds for ViT-MAE and DINOv3. Colors indicate the **first**, **second**, and **third** best results within each SAE family, based on mean performance.

		Near-OOD		Far-OOD	
SAE	**Backbone**	**FPR95↓**	**AUC↑**	**FPR95↓**	**AUC↑**
TopK	DINOv2	59.12 ±0.68	80.93 ±0.14	22.89 ±0.58	94.88 ±0.07
	ViT-MAE	72.01 ±0.24	78.94 ±0.57	43.00 ±5.02	92.38 ±0.21
	DINOv3	69.89 ±0.78	76.33 ±0.13	31.53 ±0.42	92.45 ±0.13
ReLU	DINOv2	59.11 ±0.58	80.82 ±0.20	20.01 ±0.73	95.31 ±0.19
	ViT-MAE	71.10 ±0.22	79.95 ±0.09	38.19 ±0.45	92.91 ±0.04
	DINOv3	69.44 ±0.45	77.09 ±0.16	29.15 ±0.25	93.09 ±0.13

Table 4. Ablation on the effect of local conditioning with TopK SAE. We report AUROC (↑) and FPR95 (↓) where concept rarity is computed using either the dataset-wide frequency f_j or the proposed cluster-conditioned frequency $f_j^{c(x)}$.

	AUROC ↑		FPR95 ↓	
Dataset	Local	Global	Local	Global
SSB-Hard	**74.49** ±0.17	53.96 ±0.90	**69.39** ±0.90	91.90 ±0.15
NINCO	**87.38** ±0.22	55.79 ±0.10	**48.85** ±1.19	91.01 ±1.50
iNaturalist	**98.14** ±0.08	60.79 ±1.91	**8.14** ±0.57	87.35 ±2.71
Textures	**91.57** ±0.10	59.81 ±2.12	**35.25** ±0.98	88.37 ±2.06
OpenImage-O	**94.93** ±0.06	59.80 ±0.75	**25.28** ±0.54	88.24 ±1.81

while far-OOD AUROC varies by only 0.51 points. Cross-SAE results use the normalized rare fraction in Eq. (5), and the ℓ_0-only control in Appendix D.3 confirms that code density alone does not account for the detection signal.

Default architecture. We use TopK as the default configuration as its fixed activation budget produces a compact and controlled responsible set, facilitating inspection. ReLU is reported as an accuracy-oriented alternative. Despite its denser latent representation, it achieves the strongest far-OOD performance. As such, RareCount does not depend on a specific reconstruction quality or sparsity regime.

4.5. From Score to Explanation

Since RareCount is computed directly from per-concept rarity indicators, it does not require a separate attribution method, i.e., every prediction decomposes exactly into the responsible set $\mathcal{R}(x)$ defined in Sec. 3.1. For each responsible concept, the detector exposes its activation a_j and its cluster-conditioned frequency $f_j^{c(x)}$, allowing a decision to be seen as: the input is anomalous since it activates concepts that are unusual for the ID region to which it is assigned.

Local concept-level explanation. The explanation is local by construction. A concept is in $\mathcal{R}(x)$ when

$$f_j^{c(x)} < \tau,$$

meaning that it is infrequent among ID samples represented similarly by the backbone. The responsible set therefore identifies concepts that are unexpected in the input's representation-level context, not merely concepts that happen to be rare in the dataset.

Dataset-level concept signatures. The same decomposition can be aggregated across samples to characterize an OOD source as a whole. Different datasets exhibit distinct profiles of responsible-concept activations: the concepts that occur on Textures, for example, differ from those on iNaturalist. Thus, $\mathcal{R}(x)$ provides not only sample-level evidence of anomaly, but also a source-level description of how the distribution differs from ID. We report the complete concept-signature analysis in Appendix G.

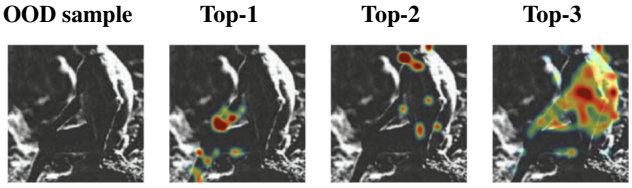


Figure 3. Visual grounding of the responsible set $\mathcal{R}(x)$. The leftmost column shows the input, while the others show the three rare responsible concepts with the highest activation. Warmer colors indicate stronger spatial activation. See Sec. 4.5 and Appendix H for additional examples and quantitative analysis.

Spatial grounding. Responsible concepts can also be localized in the image. Following prior global-to-local probing approaches in ViTs, where mappings learned on whole-image representations are reused on spatially dense representations to obtain local relevance maps [24], we reuse the patch tokens produced by the same backbone forward pass and encode them with the CLS-trained SAE, obtaining a spatial activation map for every $j \in \mathcal{R}(x)$. Figure 3 shows that different responsible concepts highlight complementary object parts, contextual regions, and texture patterns, decomposing the detection evidence into distinct spatial components.

Token Transfer Analysis. We empirically evaluate the transfer between CLS and patch embeddings. Table 6 tests three complementary properties. First, CLS–patch agreement measures whether concepts selected at image level are consistently recovered at patch level. Second, own-region occlusion tests whether the regions highlighted for a concept are selectively associated with that concept. Third, rare-map atypicality tests whether responsible-concept maps align with locally unusual patches specifically on OOD inputs. The three tests provide consistent evidence across all five OOD sources. CLS–patch agreement remains high, with AUROC between 0.811 and 0.879. Own-region occlusion produces a larger effect than cross-concept control occlusion

Table 5. Sparse autoencoder diagnostics across independently trained seeds for the four SAE families compared in Tab. 1. Lower MSE and higher cosine similarity and explained variance (EV) indicate better reconstruction; L_0 and density characterize the sparsity regime. Dead and near-dead units quantify inactive features, while Gini and decoder redundancy characterize dictionary usage. Values are reported as the mean and standard deviation across ten independently trained SAE seeds for TopK and five seeds for the other SAE variants.

SAE	Reconstruction			Sparsity		Inactive features		Dictionary usage	
	MSE ↓	Cos. ↑	EV ↑	L_0 ↓	Density ↓	Dead ↓	Near-dead ↓	Gini ↓	Redund. ↓
TopK	0.0007 ± 0.0000	0.667 ± 0.000	0.419 ± 0.001	128.0 ± 0.0	1.67 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%	0.409 ± 0.003	**0.000** ± 0.000%
BatchTopK	0.0006 ± 0.0000	0.728 ± 0.001	0.510 ± 0.003	194.9 ± 1.8	2.54 ± 0.02%	0.00 ± 0.00%	0.00 ± 0.00%	0.405 ± 0.001	**0.000** ± 0.000%
Gated	0.0007 ± 0.0000	0.691 ± 0.005	0.448 ± 0.008	160.7 ± 4.6	2.09 ± 0.06%	0.00 ± 0.00%	0.00 ± 0.00%	0.357 ± 0.004	**0.000** ± 0.000%
ReLU	**0.0004** ± 0.0000	**0.855** ± 0.001	**0.682** ± 0.004	848.6 ± 11.5	11.05 ± 0.15%	0.00 ± 0.00%	0.00 ± 0.00%	0.225 ± 0.001	**0.000** ± 0.000%

Table 6. **Quantitative analysis of concept-level spatial grounding.** Results for CLS–patch agreement, own-region occlusion, and rare-map atypicality across all OOD datasets. All three tests consistently support spatial grounding of the responsible concepts.

OOD dataset	CLS–patch AUC ↑	Own-region wins ↑	Rare-map atypicality ρ ↑ *ID-val:* 0.005 *(n.s.)*
SSB-Hard	0.827	61%	0.047
NINCO	0.842	79%	0.089
iNaturalist	0.811	59%	0.038
Textures	0.879	64%	0.066
OpenImage-O	0.837	70%	0.048

in a majority of independent images for every source (59–79%). Finally, responsible-concept maps correlate positively with patch-level atypicality on every OOD dataset, while the same association is negligible and non-significant on ID validation. As such, these results support spatial grounding of the concepts producing the RareCount score. Full controls, effect sizes, confidence intervals, and multiple-comparison corrections are reported in Appendix H and Appendix A.3.

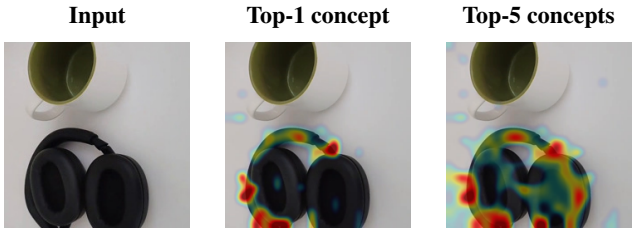


Figure 4. **Controlled object-level grounding.** A scene containing an ID mug and designated OOD headphones. Responsible-concept activation concentrates primarily on the headphones despite both objects being visible.

Controlled object-level demo. To complement the benchmark analysis, we evaluate RareCount on scenes containing ID objects with a clear correspondence to ImageNet classes, together with OOD objects. As shown in Fig. 4, the responsible-concept activations concentrate primarily on the headphones rather than on the accompanying mug, despite both objects being simultaneously visible. This provides a qualitative check that the detected evidence localizes on the designated OOD object rather than merely reflecting generic multi-object saliency. Additional object combinations and familiar-only controls are reported in Appendix J.

Fidelity to the detector. The explanation is a direct readout of the detector rather than a post-hoc approximation. RareCount depends only on which active concepts satisfy the local rarity condition, and $\mathcal{R}(x)$ contains exactly those concepts. Replacing the binary indicators with activation magnitudes or magnitude-weighted variants does not improve detection (Appendix D.1). The object being inspected is therefore the same object that determines the score, although this should not be interpreted as a causal explanation of the underlying backbone.

5. Conclusions

In this work, we have introduced RareCount, an interpretable, label- and head-free OOD detector using SAEs to identify unusually strong activation of rare concepts relative to similar samples. By operating in concept space, RareCount enables score decomposition and spatial grounding of the responsible concepts. On OpenOOD, RareCount is competitive with post-hoc detectors, improves mean far-OOD FPR95 over the only other SAE-based detector by up to 14.8 points, and generalizes across multiple VFMs and sparsity levels. We have shown that locally conditioned concept rarity provides effective and interpretable OOD detection for safer deployment of VFMs. By moving from the dense representation space of a VFM to a sparse concept space, RareCount takes a step toward addressing the limited interpretability of existing OOD detectors.

6. Limitations and Future Work

Our evaluation focuses on semantic OOD detection with ImageNet-1K as the reference distribution. Although we assess robustness on ImageNet-C and generalization across backbones and SAE architectures, evaluation remains limited to natural-image benchmarks. Future work should inves-

tigate RareCount in specialized domains, other modalities, and real-world distribution shifts. Moreover, while responsible concepts are faithful to the detector by construction, they should not yet be interpreted as causal explanations of the underlying backbone. Studying their causal relevance and semantic properties therefore remains a promising direction.

References

[1] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. pages 3319–3327, 2017. 1

[2] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. 1

[3] Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. 2

[4] Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2025. 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 2

[6] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[7] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. 1

[8] 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*, 2021. 2

[9] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[10] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. arXiv:2209.10652. 1

[11] Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 2, 4

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations.

In *International Conference on Learning Representations*, 2019. 4

[14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017. 1, 2

[15] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[16] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*, 2021. 1

[17] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 677–689, 2021. 2

[18] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3

[19] Mert Keser, Halil Ibrahim Orhan, Niki Amini-Naieni, Gesina Schwalbe, Alois C. Knoll, and Matthias Rottmann. Benchmarking vision foundation models for input monitoring in autonomous driving. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025. 1

[20] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[21] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[22] Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 3

[23] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21464–21475, 2020. 1, 2

[24] Andrew F. Luo, Jacob Yeung, Rushikesh Zawar, Shaurya Dewan, Margaret M. Henderson, Leila Wehbe, and Michael J. Tarr. Brain mapping with dense features: Grounding cortical semantic selectivity in natural images with vision transformers. *ArXiv*, abs/2410.05266, 2024. 7

[25] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1

[26] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *International Conference on Learning Representations (ICLR)*, 2023. 2

[27] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015. 1

[28] Ahyoung Oh, Wonseok Shin, and Songkuk Kim. Sparsity as a key: Unlocking new insights from latent structures for out-of-distribution detection. *arXiv preprint arXiv:2604.26409*, 2026. 2, 4, 5

[29] Matthew Lyle Olson, Musashi Hinck, Neale Ratzlaff, Changbai Li, Phillip Howard, Vasudev Lal, and Shao-Yen Tseng. Probing the representational power of sparse autoencoders in vision models. *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 6226–6236, 2025. 2

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 2, 4

[31] 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. 2, 3

[32] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 1, 2

[33] Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024. 2, 3

[34] Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024. 3

[35] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-OOD detection. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2021. 2

[36] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 1

[37] Anamta Sayyed, Rifat Shahriar, Milin Zhang, Ananthram Swami, Michael De Lucia, Nathaniel D. Bastian, and Francesco Restuccia. Out-of-distribution detection in computer vision: A comprehensive survey and research challenges. *ACM Computing Surveys*, 2026. 1

[38] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 1177–1178, 2010. 3

[39] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael 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, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. Featured Certification. 2, 4

[40] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models. *arXiv preprint arXiv:2502.06755*, 2025. 2

[41] Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 144–157, 2021. 2

[42] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. 1, 2

[43] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. 1, 2

[44] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning (ICML)*, 2022. 2

[45] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022. 2, 4

[46] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131:2607–2622, 2022. 2

[47] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2(3): 1–32, 2024. 4, 5

[48] Yukun Zhou, Mark A. Chia, Siegfried Karl Wagner, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622:156 – 163, 2023. 1

RareCount: Interpretable Out-of-Distribution Detection in Sparse Concept Space

Supplementary Material

Contents

A. Experimental Protocol and Reproducibility **13**
- A.1. Backbones and Feature Extraction . . . 13
- A.2. Sparse Autoencoder Training . . . 13
- A.3. Statistical Analysis . . . 13
- A.4. Computational Cost . . . 14

B. Hyperparameter Selection and Clustering Robustness **14**
- B.1. Robustness to the Threshold-Selection Criterion . . . 14
- B.2. ID-Only Threshold Selection . . . 15
- B.3. Clustering Ablation . . . 16

C. SAE Capacity, Sparsity, and Diagnostics **19**
- C.1. Dictionary Width and Activation Budget . . . 19
- C.2. Dictionary Diagnostics Across Configurations . . . 21

D. Score Form and Confounding Factors **21**
- D.1. Alternative Score Definitions . . . 21
- D.2. Rarity Among the Strongest Activations . . . 22
- D.3. ℓ_0 as a Potential Confounder . . . 23

E. Covariate Shift Study with ImageNet-C **23**

F. Extended OOD Detection Results **25**
- F.1. Statistical Significance Across SAE Seeds . . . 25
- F.2. Full Per-Dataset Results Across Backbones . . . 26
- F.3. Per-Dataset Score Distributions . . . 26
- F.4. Rare-Concept Distributions Across SAE Families . . . 28

G. Characterization of Rare and Responsible Concepts **28**

H. Quantitative Validation of Spatial Grounding **31**

I . Additional Qualitative Examples **32**

J . Controlled Object-Level Demo **35**

A. Experimental Protocol and Reproducibility

A.1. Backbones and Feature Extraction

We use the official pretrained checkpoints associated with each backbone. For DINOv2 ViT-B/14, we use `dinov2_vitb14_lc` distributed through the official DINOv2 Torch Hub repository.[1] For DINOv3 ViT-B/16, we use the Hugging Face checkpoint `facebook/dinov3-vitb16-pretrain-lvd1689m`.[2] For ViT-MAE ViT-B/16, we instantiate `vit_base_patch16_224.mae` through `timm` and load the official ImageNet-1K end-to-end fine-tuned checkpoint released by Meta. [3] For DINOv2 and ViT-MAE, images are resized to 256 pixels and center-cropped to 224×224; for DINOv3, images are resized to 520 pixels and center-cropped to 512×512. In all cases, bicubic interpolation and ImageNet normalization with mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$ are used. Feature extraction is performed in mixed precision. Unless otherwise specified, we extract the CLS representation returned by each backbone. Table 7 reports the probe accuracy on ImageNet-1K for each backbone.

Table 7. ImageNet-1K classification accuracy of the linear probe.

Backbone	**IN-1K Acc.**
DINOv2 ViT-B/14	84.63
ViT-MAE ViT-B/16	83.45
DINOv3 ViT-B/16	84.81

A.2. Sparse Autoencoder Training

For each backbone, we train the SAEs on the ℓ_2-normalized CLS activations of the ImageNet-1K training split. All architectures use an input dimension of $d = 768$ and a dictionary width of $m = 7680$ ($10 \times d$). We train for 100 epochs with Adam, a learning rate of 4×10^{-4}, and a batch size of 4096. TopK and BatchTopK use $k = 128$. For the sparsity-regularized variants, we use an ℓ_1 coefficient of 9×10^{-5} for Gated and 3×10^{-5} for ReLU. Decoder weights are Kaiming-initialized and normalized to unit norm, the encoder is initialized from the decoder transpose, and the input bias is initialized to the mean training activation. Table 8 summarizes the SAE-specific training hyperparameters.

Table 8. Sparse autoencoder training hyperparameters. Shared parameters are identical across all SAE architectures.

SAE	m	k	ℓ_1	**Epochs**
TopK	7680	128	–	100
BatchTopK	7680	128	–	100
Gated	7680	–	9×10^{-5}	100
ReLU	7680	–	3×10^{-5}	100

A.3. Statistical Analysis

RareCount results are reported as mean $\pm$ standard deviation across independently trained SAE runs. For the primary DINOv2 TopK configuration, we use ten seeds; BatchTopK, Gated, and ReLU experiments on DINOv2 use five seeds, while cross-backbone experiments use three seeds. Deterministic post-hoc baselines evaluated on a fixed backbone–classifier pair are reported as a single value. Near- and Far-OOD summary metrics are computed as unweighted arithmetic means over their constituent datasets, giving each dataset equal weight. Overall-OOD summary metrics are computed analogously as the unweighted arithmetic mean over all five OOD datasets. For the spatial-grounding analyses, the image is treated as the independent statistical unit. For the patch-atypicality analysis, 95% bootstrap confidence intervals are computed from 2,000 bootstrap resamples. Concept-specific occlusion is tested using a one-sided paired Wilcoxon signed-rank test on the per-image self–cross occlusion difference, with the alternative hypothesis that occluding a concept's own region produces a larger activation decrease than cross-concept occlusion. The five dataset-level p-values are jointly corrected using Holm–Bonferroni.

[1] https://github.com/facebookresearch/dinov2
[2] https://huggingface.co/facebook/dinov3-vitb16-pretrain-lvd1689m
[3] https://dl.fbaipublicfiles.com/mae/finetune/mae_finetuned_vit_base.pth

A.4. Computational Cost

We evaluate the computational overhead of RareCount using the default DINOv2 ViT-B/14 + TopK configuration on a single NVIDIA RTX A4000 GPU. We distinguish the one-time offline cost required to construct the detector from the recurring cost incurred at inference. Training the TopK SAE for 100 epochs requires approximately $1600\,\mathrm{s}$ ($26.7\,\mathrm{min}$), while constructing the local ID reference requires $146.2\,\mathrm{s}$ ($2.44\,\mathrm{min}$). The latter includes MiniBatch k-means fitting on $80\mathrm{k}$ ID embeddings and the subsequent estimation of the cluster-conditioned concept-frequency profiles from $400\mathrm{k}$ ID samples. Both operations are performed only once for a given backbone and reference distribution. At inference, the additional post-backbone computation consists of SAE encoding, nearest-centroid assignment, and local rarity scoring. We report the RareCount-specific post-backbone overhead separately from total end-to-end latency. This stage requires $0.346\,\mathrm{ms/image}$ at batch size one. End-to-end latency increases from $6.818\,\mathrm{ms/image}$ for the frozen VFM alone to $7.019\,\mathrm{ms/image}$ for the complete pipeline, corresponding to a 2.95% increase, while maintaining a throughput of 142.5 images/s. The dominant post-backbone computation amounts to approximately 13.33 MFLOPs per image. The stored SAE checkpoint occupies 45.09 MB, while the cluster centroids and local concept-frequency profiles require an additional 32.23 MB in FP32, resulting in 77.32 MB of additional deployment storage relative to the frozen VFM. Since RareCount only requires the SAE encoder at inference, the decoder can be discarded for deployment, reducing the SAE storage to 22.53 MB and the total additional storage to 54.76 MB. Overall, these results indicate that RareCount is lightweight at inference: the additional computation and storage remain small relative to the frozen backbone, while SAE training and ID-reference construction are one-time offline costs.

Table 9. Computational cost of the default DINOv2 ViT-B/14 + TopK configuration, measured on an NVIDIA RTX A4000. Offline costs are incurred once, while online costs are measured at batch size one. Deployment storage is reported in addition to the frozen VFM.

Quantity	Cost
One-time offline cost	
SAE training	26.7 min
ID reference construction	2.44 min
Online inference	
Frozen VFM latency	6.818 ms/img
RareCount post-backbone latency	0.346 ms/img
Full pipeline latency	7.019 ms/img
Full pipeline throughput	142.5 img/s
Post-backbone computation	13.33 MFLOPs/img
Additional deployment storage	
Full SAE checkpoint	45.09 MB
ID reference	32.23 MB
Total (full checkpoint)	**77.32 MB**
Encoder-only SAE	22.53 MB
Total (encoder-only)	**54.76 MB**

B. Hyperparameter Selection and Clustering Robustness

Unless otherwise stated, all analyses in this section are performed on the default DINOv2 ViT-B/14 + TopK SAE configuration, while varying only the component under study. This controlled setting isolates the sensitivity of RareCount to its calibration and clustering choices from variation due to the SAE architecture or backbone. In experiments involving random seeds for clustering, the trained SAE is kept fixed and only the construction of the ID partition and associated reference statistics is repeated.

Construction of the Rarity-Threshold Grid. Rather than sweeping arbitrary values of τ, we construct the candidate grid from the empirical distribution of cluster-conditioned concept frequencies f_j^k estimated on the ID reference set. This restricts the search to thresholds corresponding to frequencies actually observed in the learned local concept profiles. Candidate thresholds are evaluated exclusively on the validation split; the selected value is then frozen before test evaluation.

B.1. Robustness to the Threshold-Selection Criterion

We evaluate four validation-based rules for selecting τ: maximizing AUROC, minimizing FPR95, maximizing Youden's index, and selecting the center of the high-AUROC plateau. For the plateau criterion, we retain candidate thresholds whose validation

AUROC lies within 0.002 of the maximum and select the observed threshold closest to the midpoint of the retained range in log-τ space. As shown in Tab. 10, the selected thresholds differ, while validation AUROC remains within 0.13 percentage points across criteria. We use the minimum-FPR95 solution, $\tau = 0.0319$, as the default operating point. This value refers to the reference TopK configuration used in the sensitivity analyses; in multi-seed experiments, τ is selected independently on the validation split for each SAE seed.

Table 10. Validation-based selection of the rarity threshold τ.

Selection criterion	τ	Val AUROC	Val FPR95
Max AUROC	0.0460	94.88	28.4
Min FPR95	0.0319	94.75	25.5
Max Youden	0.0451	94.86	27.4
AUROC plateau center	0.0406	94.81	26.4

Stability to Threshold. Fig. 5 shows validation performance across the full rarity-threshold sweep. RareCount exhibits a broad stable region around the selected operating point rather than a narrow optimum: AUROC remains high over a substantial interval of τ, with corresponding FPR95 varying smoothly. Performance degrades mainly at the extremes, where the rarity criterion becomes respectively too restrictive or too permissive.

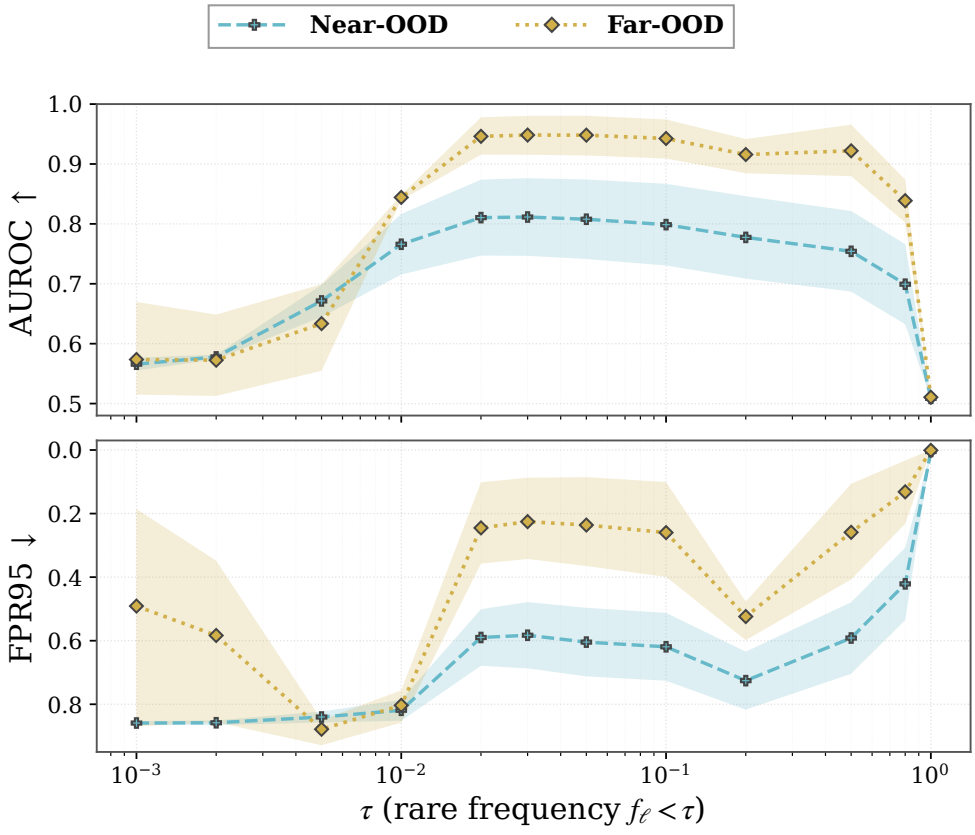


Figure 5. Sensitivity of RareCount to the rarity threshold τ.

B.2. ID-Only Threshold Selection

The main experiments select the rarity threshold τ on a validation split containing both ID and OOD samples. To assess whether RareCount fundamentally requires access to OOD data for calibration, we additionally consider a strictly ID-only

selection rule in which τ is estimated solely from the local concept-frequency statistics computed on the ID training data. For each cluster c and concept j, let $f_j^{(c)}$ denote the empirical frequency of concept j within cluster c. We collect all positive entries of the cluster–concept frequency matrix and operate in log-frequency space,

$$z_j^{(c)} = \log f_j^{(c)}, \qquad f_j^{(c)} > 0.$$

We then apply one-dimensional Otsu thresholding to $\{z_j^{(c)}\}$, selecting the split that maximizes the between-group variance. Zero-frequency entries are excluded from threshold estimation because they are already classified as rare for any positive threshold; Otsu is therefore applied only to the non-zero frequency distribution. Specifically, we define

$$t^\star = \arg\max_t \sigma_B^2(t),$$

where $\sigma_B^2(t)$ denotes the Otsu between-group variance for a split at t. The resulting rarity threshold in the original frequency space is

$$\tau_{\text{ID}} = \exp(t^\star).$$

This procedure is fully data-adaptive and introduces no OOD-dependent criterion: neither OOD samples nor AUROC/FPR95 are used to determine τ. We repeat the procedure independently for the ten TopK SAE seeds used in the DINOv2 experiments. The selected thresholds are remarkably consistent across independently trained SAEs,

$$\tau_{\text{ID}} = 0.016108 \pm 0.000016,$$

with individual values ranging only from 0.016094 to 0.016146. This low across-seed variability indicates that the ID frequency distribution induces a highly stable operating threshold, despite changes in SAE initialization.

Table 11. Comparison between validation-tuned and strictly ID-only threshold selection for TopK on DINOv2. Results are mean ± standard deviation over ten independently trained SAE seeds. AUROC and FPR95 are reported in percentage points.

			Near-OOD		Far-OOD	
Selection	OOD?	τ	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Validation-tuned	✓	0.0337 ±0.0076	80.93 ±0.14	59.12 ±0.68	94.88 ±0.07	22.89 ±0.58
ID-only Otsu	×	0.01611 ±0.00002	80.42 ±0.08	61.82 ±0.45	93.39 ±0.11	33.76 ±1.26

As shown in Tab. 11, removing OOD data from threshold selection preserves most of the ranking performance of RareCount. Relative to the validation-tuned TopK setting, ID-only calibration decreases Near-OOD AUROC by only 0.51 points and Far-OOD AUROC by 1.49 points. Near-OOD FPR95 increases moderately, from 59.12 to 61.82 ($+2.7$ points). The larger difference appears for Far-OOD FPR95, which increases from 22.89 to 33.76 ($+10.87$ points). These results suggest that OOD validation is not necessary to recover a competitive RareCount ranking: an operating threshold derived entirely from ID concept statistics retains nearly the same AUROC, particularly on Near-OOD data. At the same time, FPR95 is more sensitive to calibration, especially on Far-OOD datasets, indicating that OOD-aware tuning primarily improves the high-recall operating regime rather than being necessary for the underlying separation between ID and OOD samples. Together with the threshold-sensitivity analysis in Fig. 5, this experiment shows that RareCount does not rely on a narrowly tuned OOD-dependent threshold. Instead, the local ID concept-frequency distribution provides a stable, reproducible estimate of the rarity boundary, while OOD-aware validation can be viewed as an optional calibration step for optimizing a specific operating metric.

B.3. Clustering Ablation

Number of Clusters. We vary the granularity of the ID partition over $K \in \{500, 1000, 1500, 2000\}$ while keeping $\tau = 0.0319$ fixed. Table 12 shows that performance changes only moderately across this range. Although $K = 1500$ reaches the highest validation AUROC, $K = 1000$ provides the best validation Youden index together with the lowest FPR95 and is therefore used as the default. Importantly, K controls only the granularity of the unsupervised partition and is unrelated to

the number of ImageNet classes. The comparable performance obtained for $K \in \{500, 1500, 2000\}$ further shows that the selected $K = 1000$ is not uniquely tied to the ImageNet-1K class count.

Table 12. Validation-based selection of the number of clusters K with fixed $\tau = 0.0319$.

K	Val AUROC ↑	Val FPR95 ↓	Val Youden ↑	Selection
500	94.37	27.57	0.7499	–
1000	**94.75**	**25.47**	**0.7511**	**Selected**
1500	<u>94.88</u>	26.09	0.7507	–
2000	94.36	25.87	0.7322	–

Joint Sensitivity to K and τ. Because the empirical frequency distribution changes with cluster granularity, we additionally re-select τ independently for each value of K. Table 13 shows that the preferred threshold shifts with K, as expected, while validation performance remains stable. The default pair $(K, \tau) = (1000, 0.0319)$ achieves the best Youden index and the lowest FPR95, indicating that RareCount does not rely on a narrowly tuned combination of the two hyperparameters.

Table 13. Joint validation selection of number of clusters and rarity threshold τ. For each value of K, τ is re-selected on validation using the same validation criterion adopted for the default configuration..

K	Selected τ	Val AUROC ↑	Val FPR95 ↓	Val Youden ↑
500	0.0234	94.43	25.69	0.7480
1000	**0.0319**	**94.75**	**25.47**	**0.7511**
1500	0.0363	94.49	26.21	0.7463
2000	0.0322	94.35	25.64	0.7326

Clustering Algorithm. We evaluate whether RareCount depends on the specific clustering algorithm used to construct the local ID regions. Keeping the remaining pipeline fixed, we compare MiniBatch k-means, full k-means with k-means++ initialization, k-means with random initialization, and a diagonal-covariance Gaussian mixture model. As shown in Fig. 6, the local-rarity signal is preserved across all clustering backends. Full k-means++ provides the strongest overall performance, but only modestly improves over MiniBatch k-means, particularly on Far-OOD. We therefore retain MiniBatch k-means as the default because it provides comparable detection performance while being substantially more scalable to large ID reference sets.

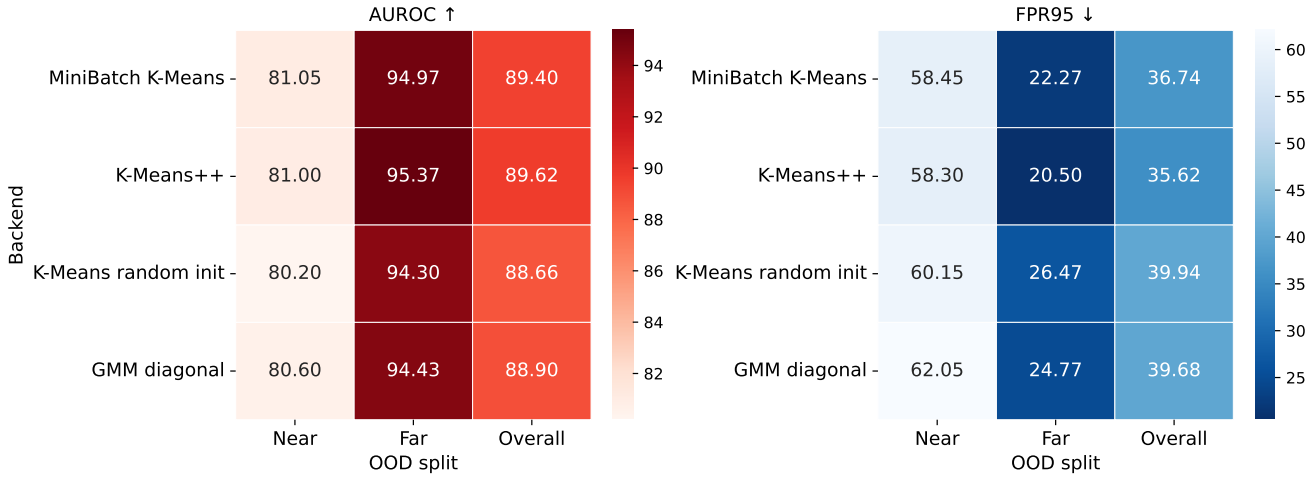


Figure 6. Sensitivity to the clustering algorithm used to construct the local ID regions. We compare MiniBatch k-means, full k-means with k-means++ initialization, randomly initialized k-means, and a diagonal-covariance Gaussian mixture model. RareCount remains effective across clustering backends, with k-means++ providing a modest improvement over the default MiniBatch k-means configuration.

Clustering Stability Across Random Seeds. We repeat construction of the ID partition with three independent clustering seeds while keeping $K = 1000$, 80k samples for centroid fitting, and 400k samples for concept-frequency estimation. As shown in Fig. 7, AUROC and FPR95 remain highly consistent across seeds on both Near- and Far-OOD. This indicates that the local-rarity signal does not depend on a particular random clustering subset or initialization.

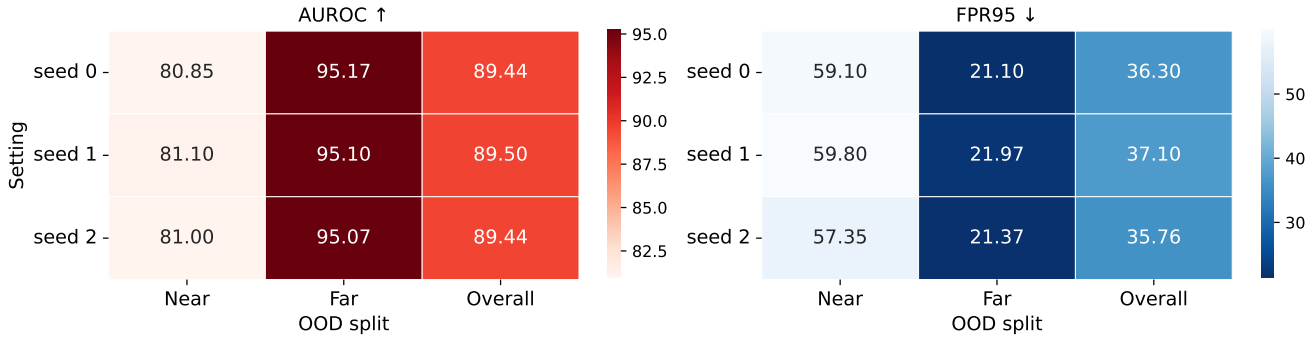


Figure 7. Robustness to the clustering seed. Each row corresponds to an independent MiniBatch k-means run with $K = 1000$, using 80k ID samples for centroid estimation and 400k samples for local concept-frequency statistics. Results remain stable across seeds.

Centroid Estimation. We first vary the number of ID samples used to fit the cluster centroids, using 80k, 160k, and 320k samples while keeping the concept-frequency reference set fixed at 400k. Figure 8 shows only minor variation in detection performance, indicating that 80k samples are already sufficient to estimate representative local regions.

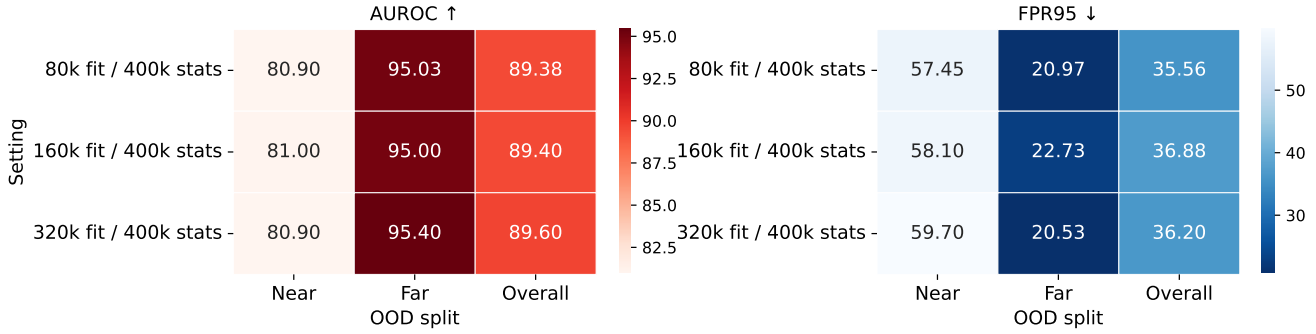


Figure 8. Sensitivity to the number of ID samples used for centroid estimation. Concept-frequency statistics are estimated from a fixed set of 400k ID samples.

Concept-Frequency Estimation. We next vary the number of ID samples used to estimate the local concept frequencies f_j^k, using 200k, 400k, and 800k samples while fixing the centroid-fitting set to 80k samples. Performance remains nearly unchanged (Fig. 9), showing that the estimated local concept profiles are stable over a broad range of reference-set sizes.

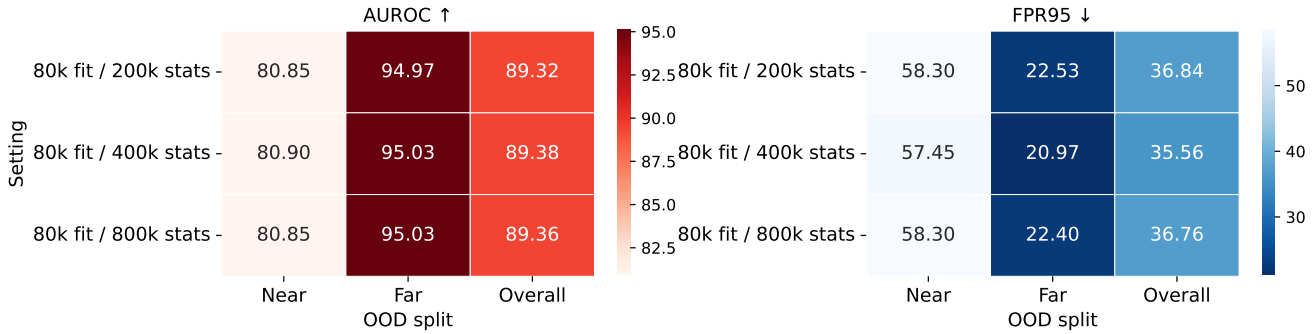


Figure 9. Sensitivity to the number of ID samples used to estimate the local concept-frequency profiles. Centroids are fitted on a fixed set of 80k ID samples, while the statistics set is varied over 200k, 400k, and 800k samples. Detection performance remains stable across reference-set sizes.

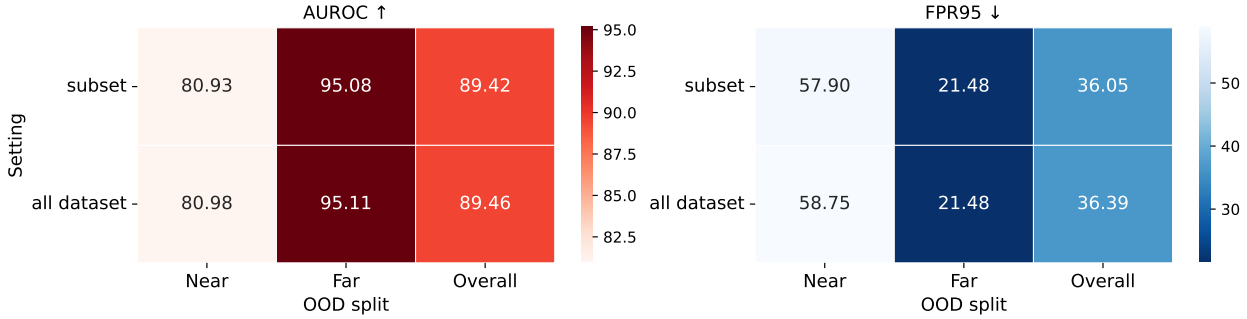


Figure 10. Comparison between subset-based and full-data estimation of the ID reference, averaged across clustering seeds. The two configurations produce nearly identical OOD detection performance.

Subset Versus Full-Data Estimation. Finally, we compare the scalable subset-based configuration with estimation using the full available ID training data. Figure 10 shows nearly identical AUROC and FPR95, confirming that the subset-based reference used in our default configuration does not materially alter the detector.

C. SAE Capacity, Sparsity, and Diagnostics

C.1. Dictionary Width and Activation Budget

We test whether RareCount depends on the capacity or sparsity of the underlying TopK SAE by jointly varying the dictionary width $m \in \{3840, 7680, 11520\}$ and the activation budget $k \in \{32, 64, 128, 256\}$. As shown in Fig. 11, detection performance remains stable across a broad range of both hyperparameters, with no systematic improvement from increasing either dictionary width or activation budget. The default configuration $(m, k) = (7680, 128)$ therefore represents an intermediate operating point rather than a narrowly tuned optimum.

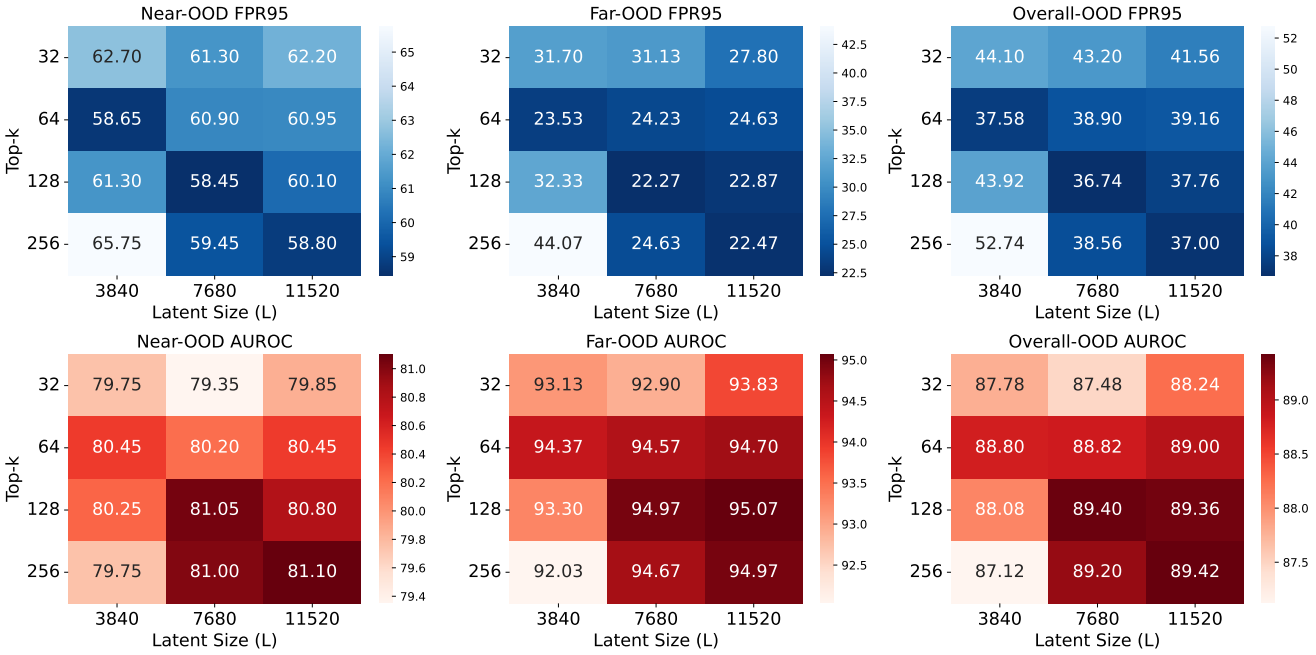


Figure 11. Joint sensitivity to the SAE dictionary width m and TopK activation budget k. We vary $m \in \{3840, 7680, 11520\}$ and $k \in \{32, 64, 128, 256\}$ while keeping the remaining RareCount pipeline fixed. Detection performance remains stable across substantially different dictionary capacities and sparsity levels.

Table 14. Sparse autoencoder diagnostics on DINOv3 ViT-B/16 across independently trained seeds. Lower MSE and higher cosine similarity and explained variance (EV) indicate better reconstruction; L_0 and density characterize the sparsity regime. Dead and near-dead units quantify inactive features, while Gini and decoder redundancy characterize dictionary usage.

SAE	Reconstruction			Sparsity		Inactive features		Dictionary usage	
	MSE ↓	Cos. ↑	EV ↑	L_0 ↓	Density ↓	Dead ↓	Near-dead ↓	Gini ↓	Redund. ↓
TopK	0.0008 ±0.0000	0.649 ±0.001	0.413 ±0.002	128.0 ±0.0	1.67 ±0.00%	0.00 ±0.00%	0.00 ±0.00%	0.523 ±0.006	**0.000** ±0.000 **%**
BatchTopK	0.0006 ±0.0000	0.716 ±0.003	0.506 ±0.004	206.3 ±3.2	2.69 ±0.04%	0.00 ±0.00%	0.00 ±0.01%	0.518 ±0.008	**0.000** ±0.000 **%**
Gated	0.0006 ±0.0000	0.750 ±0.004	0.544 ±0.008	237.3 ±5.9	3.09 ±0.08%	0.00 ±0.00%	0.00 ±0.00%	0.293 ±0.011	**0.000** ±0.000 **%**
ReLU	**0.0004** ±0.0000	**0.860** ±0.000	**0.709** ±0.000	999.3 ±0.6	13.01 ±0.01%	0.00 ±0.00%	0.00 ±0.00%	0.164 ±0.001	**0.000** ±0.000 **%**

Table 15. Sparse autoencoder diagnostics on ViT-MAE ViT-B/16 across independently trained seeds. Lower MSE and higher cosine similarity and explained variance (EV) indicate better reconstruction; L_0 and density characterize the sparsity regime. Dead and near-dead units quantify inactive features, while Gini and decoder redundancy characterize dictionary usage.

SAE	Reconstruction			Sparsity		Inactive features		Dictionary usage	
	MSE ↓	Cos. ↑	EV ↑	L_0 ↓	Density ↓	Dead ↓	Near-dead ↓	Gini ↓	Redund. ↓
TopK	0.0007 ±0.0000	0.668 ±0.001	0.430 ±0.001	128.0 ±0.0	1.67 ±0.00%	0.00 ±0.00%	0.03 ±0.01%	0.593 ±0.005	**0.000** ±0.000 **%**
BatchTopK	0.0005 ±0.0000	0.793 ±0.001	0.614 ±0.002	306.7 ±2.8	3.99 ±0.04%	0.00 ±0.00%	0.01 ±0.01%	0.621 ±0.002	**0.000** ±0.000 **%**
Gated	0.0006 ±0.0000	0.746 ±0.001	0.531 ±0.002	247.6 ±1.6	3.22 ±0.02%	0.00 ±0.00%	0.00 ±0.00%	0.375 ±0.002	**0.000** ±0.000 **%**
ReLU	**0.0004** ±0.0000	**0.853** ±0.002	**0.689** ±0.005	954.8 ±1.5	12.43 ±0.02%	0.00 ±0.00%	0.00 ±0.00%	0.249 ±0.001	**0.000** ±0.000 **%**

C.2. Dictionary Diagnostics Across Configurations

We characterize the learned dictionaries in terms of reconstruction quality, sparsity, inactive features, and dictionary usage. Dead features have zero firing frequency, while near-dead features have a non-zero firing frequency below 10^{-4}. Gini is computed over the per-feature firing-frequency distribution, with lower values indicating more uniform dictionary usage, and decoder redundancy is the fraction of pairs of normalized decoder directions with cosine similarity above 0.9. Across ViT-MAE and DINOv3 configurations, all SAE variants exhibit stable reconstruction quality and negligible inactive-feature rates across seeds. Consistently across backbones, denser architectures achieve stronger reconstruction, while TopK preserves the most controlled sparse regime with a fixed $L_0 = 128$. In particular, ReLU attains the highest reconstruction quality on ViT-MAE, whereas BatchTopK improves reconstruction over TopK on DINOv3. No measurable decoder redundancy is observed across the evaluated configurations.

D. Score Form and Confounding Factors

We examine whether RareCount depends on the particular aggregation of rare concepts and whether its performance can instead be explained by variations in SAE activation density.

D.1. Alternative Score Definitions

Table 16 compares alternative ways of aggregating locally rare concepts. Weighting rare concepts by their activation magnitude does not improve detection, indicating that the presence of an unexpected concept is more informative than how strongly it activates. Similarly, using the unnormalized rare-concept count becomes less effective for SAE families whose number of active concepts varies across inputs. RareFraction provides the most consistent performance across SAE architectures by normalizing the number of locally rare concepts by the active-set size. For TopK, where $|\mathcal{A}(x)| = 128$ by construction, rare fraction, raw count, and truncated$_{128}$ are equivalent up to a constant scaling and therefore yield identical rankings. For the other SAE families, RareFraction consistently remains among the strongest variants, while restricting the score to the 128 strongest activations does not provide a systematic advantage. To test whether activation magnitude or presence provides additional information beyond the binary rarity indicator, we perform paired Wilcoxon signed-rank tests across the ten TopK SAE seeds. As shown in Tab. 17, raw count outperforms both rare-magnitude and rare-presence variants in 9–10 of the 10 seeds across Near- and Far-OOD metrics, with all comparisons remaining significant after Holm–Bonferroni correction. These results support the use of the binary local-rarity criterion underlying RareCount.

Table 16. **Alternative score definitions on DINOv2.** We compare different ways of aggregating locally rare SAE concepts. Results are mean ± standard deviation across SAE seeds (10 seeds for TopK and 5 for BatchTopK, Gated, and ReLU). RareFraction normalizes the rare-concept count by the number of active concepts, while truncated$_{128}$ restricts scoring to the 128 strongest activations.

		Near-OOD		Far-OOD	
SAE	**Score**	FPR95↓	AUROC↑	FPR95↓	AUROC↑
TopK	Rare fraction	**59.12** ±0.68	**80.93** ±0.14	**22.89** ±0.58	**94.88** ±0.07
	Rare magnitude	60.53 ±0.54	80.73 ±0.18	24.45 ±0.71	94.75 ±0.12
	Rare presence	60.38 ±0.68	80.81 ±0.17	24.30 ±0.40	94.78 ±0.07
	Raw count	**59.12** ±0.68	**80.93** ±0.14	**22.89** ±0.58	**94.88** ±0.07
	Truncated$_{128}$	**59.12** ±0.68	**80.93** ±0.14	**22.89** ±0.58	**94.88** ±0.07
BatchTopK	Rare fraction	59.89 ±0.19	**81.00** ±0.14	23.33 ±0.32	**94.92** ±0.04
	Rare magnitude	61.93 ±0.52	80.30 ±0.13	27.10 ±0.58	94.17 ±0.12
	Rare presence	61.99 ±0.40	80.24 ±0.22	27.17 ±0.78	94.12 ±0.18
	Raw count	61.92 ±0.43	80.22 ±0.13	27.38 ±0.60	94.04 ±0.11
	Truncated$_{128}$	**59.56** ±0.68	80.92 ±0.23	**23.11** ±0.47	94.90 ±0.03
Gated	Rare fraction	60.52 ±0.24	**80.92** ±0.13	**23.83** ±0.68	**94.80** ±0.22
	Rare magnitude	66.79 ±1.24	79.14 ±0.53	39.32 ±4.24	92.35 ±0.76
	Rare presence	66.72 ±1.36	79.22 ±0.60	39.32 ±4.55	92.38 ±0.82
	Raw count	66.93 ±1.41	79.04 ±0.56	40.08 ±4.55	92.11 ±0.85
	Truncated$_{128}$	**60.13** ±0.84	80.90 ±0.08	24.24 ±1.54	94.57 ±0.35
ReLU	Rare fraction	**59.11** ±0.58	**80.82** ±0.20	**20.01** ±0.73	**95.31** ±0.19
	Rare magnitude	61.42 ±0.59	80.12 ±0.17	25.00 ±0.33	94.56 ±0.12
	Rare presence	61.41 ±0.50	80.09 ±0.15	25.14 ±0.16	94.53 ±0.13
	Raw count	61.74 ±0.79	79.83 ±0.10	25.51 ±0.51	94.41 ±0.08
	Truncated$_{128}$	59.29 ±0.75	80.34 ±0.25	22.23 ±0.89	94.97 ±0.12

Table 17. Paired Wilcoxon signed-rank tests for the TopK SAE. "Wins" denotes the number of SAE seeds for which raw count performs better than the alternative. Holm–Bonferroni correction is applied jointly across all eight comparisons reported in the table.

Alternative	**Metric**	**Group**	**Wins**	$\mathbf{p}_{\mathrm{Holm}}$
Rare magnitude	AUROC	Near	9/10	0.0156
	FPR95	Near	10/10	0.0156
	AUROC	Far	9/10	0.0156
	FPR95	Far	10/10	0.0156
Rare presence	AUROC	Near	10/10	0.0156
	FPR95	Near	10/10	0.0156
	AUROC	Far	9/10	0.0156
	FPR95	Far	10/10	0.0156

D.2. Rarity Among the Strongest Activations

The local-rarity signal is not confined to weak activations in the tail of the sparse code. We rank active concepts by activation magnitude and measure the fraction satisfying the rarity criterion among the top-K concepts. As shown in Fig. 12, locally rare concepts are already overrepresented among the strongest activations of OOD inputs and remain consistently more frequent than for ID samples as K increases. This is important because it shows that the OOD signal is carried by highly activated components of the representation rather than being an artifact of weak or marginal features in the activation tail. Moreover, the separation between ID and OOD is already apparent when only a small number of the strongest active concepts is retained, indicating that local rarity provides a concentrated signal that does not require inspecting the entire sparse code.

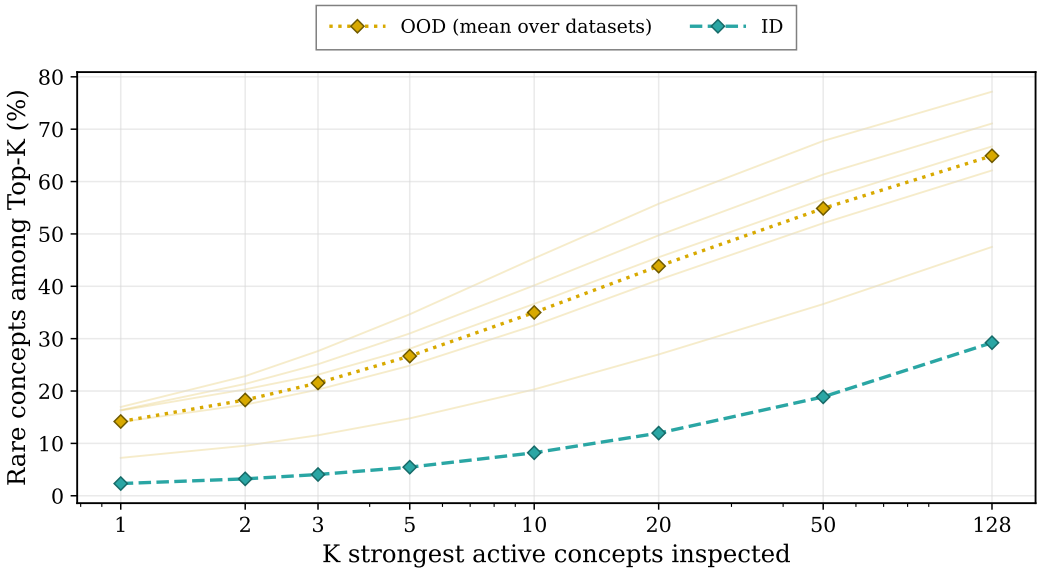


Figure 12. Fraction of locally rare concepts among the K strongest active concepts. The OOD curve denotes the average across the five evaluated OOD datasets, while individual dataset curves are shown as additional controls. Rare concepts are enriched among the strongest OOD activations rather than appearing only in the low-activation tail.

D.3. ℓ_0 as a Potential Confounder

Because the evaluated SAE families operate at substantially different sparsity levels, we test whether OOD detection could be explained simply by changes in the number of active concepts. We therefore use the per-sample $\ell_0 = |\mathcal{A}(x)|$ directly as an anomaly score. As shown in Tab. 18, ℓ_0 alone remains close to chance across all SAE families, with mean AUROC between 50.0 and 52.6. The OOD signal captured by RareCount therefore cannot be attributed to a generic increase or decrease in activation density. Instead, detection depends on *which* concepts are active relative to their local ID context.

Table 18. ℓ_0 **control.** OOD detection obtained by using only the number of active SAE concepts as the anomaly score. Results are mean $\pm$ standard deviation across SAE seeds. AUROC remains close to chance, showing that RareCount performance cannot be explained simply by changes in activation sparsity.

SAE	Near AUROC↑	Far AUROC↑
TopK	50.00 $\pm$0.00	50.00 $\pm$0.00
BatchTopK	51.18 $\pm$0.99	52.32 $\pm$0.43
Gated	52.10 $\pm$0.76	52.24 $\pm$0.51
ReLU	52.19 $\pm$0.83	52.55 $\pm$0.76

E. Covariate Shift Study with ImageNet-C

To assess how RareCount responds to non-semantic distribution shifts, we evaluate the frozen detector on the full ImageNet-C benchmark, comprising 15 corruption types at five severity levels. The SAE, ID clusters, cluster-conditioned concept frequencies, and rarity threshold are kept fixed throughout the experiment. We define a detection threshold γ as the 95th percentile of RareCount on the clean ImageNet validation split and report the fraction of corrupted ImageNet samples exceeding this threshold. Because rarity is defined relative to the clean ID reference, progressively stronger corruptions are expected to alter the backbone representation and activate concepts that are increasingly unusual for their assigned clean-ID region. Therefore, the relevant robustness criterion is not complete invariance to corruption, but whether mild shifts cause a disproportionate increase in OOD responses. As shown in Tab. 19, RareCount exhibits a gradual response to covariate shift. The fraction of samples flagged as OOD increases from 4.85% on clean ImageNet to 8.84% and 13.82% at severity levels 1 and 2, respectively, and then rises progressively to 19.76%, 26.59%, and 43.07% at severity levels 3–5. The corresponding

mean RareCount increases from 27.65 on clean ImageNet to 58.94 at severity 5. Sensitivity is strongly corruption-dependent: brightness, JPEG compression, and pixelation remain comparatively close to the clean reference, whereas blur and noise corruptions induce substantially larger shifts. Overall, these results indicate that RareCount does not indiscriminately treat mild covariate shifts as semantic OOD, while appropriately becoming more sensitive as the representation departs further from the clean-ID reference. As shown in Tab. 20, RareCount exhibits a covariate-shift response comparable to established post-hoc OOD detectors. It ranks second at both severity 1 (8.84%) and severity 2 (13.82%), while progressively increasing its response under stronger corruptions. Overall, RareCount remains relatively stable under mild non-semantic shifts without showing anomalous sensitivity compared with standard OOD methods.

Table 19. RareCount under the full ImageNet-C benchmark across all 15 corruption types and five severity levels. "Flagged" denotes the percentage of samples exceeding the OOD decision threshold, fixed at the 95th percentile of clean ImageNet validation RareCount. Clean ImageNet yields a mean RareCount of 27.65 and a flagged rate of 4.85%. All detector components are kept frozen.

Corruption	Metric	S1	S2	S3	S4	S5
Gaussian noise	Flagged (%)	10.58	8.80	19.09	30.34	62.72
	Mean RareCount	34.93	37.73	43.68	54.79	70.09
Shot noise	Flagged (%)	6.23	10.94	16.73	33.58	62.34
	Mean RareCount	34.29	38.20	44.03	55.84	67.49
Impulse noise	Flagged (%)	6.92	15.10	11.95	30.26	58.78
	Mean RareCount	34.36	40.76	41.22	53.47	66.84
Defocus blur	Flagged (%)	10.67	22.69	35.06	41.98	58.50
	Mean RareCount	41.51	48.47	57.76	61.79	70.27
Glass blur	Flagged (%)	11.99	20.00	51.62	63.78	76.36
	Mean RareCount	37.28	45.63	65.28	70.79	75.33
Motion blur	Flagged (%)	9.32	7.28	21.21	27.36	39.91
	Mean RareCount	36.32	37.68	48.18	53.33	61.15
Zoom blur	Flagged (%)	15.19	22.64	31.19	32.16	43.27
	Mean RareCount	39.98	48.05	53.58	54.87	61.19
Snow	Flagged (%)	8.98	18.77	13.45	24.90	30.29
	Mean RareCount	34.73	44.58	41.49	50.15	51.70
Frost	Flagged (%)	8.02	16.67	34.66	29.02	35.85
	Mean RareCount	34.94	42.06	54.78	51.22	55.19
Fog	Flagged (%)	10.20	6.80	12.46	12.60	34.00
	Mean RareCount	34.58	33.61	36.76	41.09	53.78
Brightness	Flagged (%)	5.68	5.43	5.34	8.84	9.57
	Mean RareCount	29.89	28.63	31.71	33.07	37.67
Contrast	Flagged (%)	4.48	6.67	11.66	11.19	37.86
	Mean RareCount	30.68	31.76	37.19	40.18	57.69
Elastic transform	Flagged (%)	9.89	30.71	15.58	25.56	62.84
	Mean RareCount	34.22	50.88	39.64	50.89	69.74
Pixelate	Flagged (%)	7.72	6.81	7.83	14.39	17.25
	Mean RareCount	31.79	31.89	34.79	40.38	41.35
JPEG compression	Flagged (%)	6.79	7.92	8.55	12.87	16.52
	Mean RareCount	33.61	34.35	36.43	39.02	44.57
Average	**Flagged (%)**	**8.84**	**13.82**	**19.76**	**26.59**	**43.07**
	Mean RareCount	**34.88**	**39.62**	**44.43**	**50.06**	**58.94**

Table 20. Fraction (%) of samples flagged as OOD on clean ImageNet and ImageNet-C at increasing corruption severity. For each detector, the decision threshold is calibrated on clean ImageNet such that approximately 5% of clean samples are flagged. Lower values indicate greater robustness to covariate shift. Best, second-best, and third-best results are highlighted.

Method	Clean	S1	S2	S3	S4	S5
ASH	4.87	9.97	16.50	22.12	30.11	46.19
ReAct	4.93	9.62	14.61	**19.62**	26.96	42.89
Temperature Scaling	5.06	8.96	15.67	20.53	27.63	42.51
MSP	5.14	**8.78**	**14.05**	**18.80**	**25.09**	**38.99**
EBO	4.87	9.97	16.50	22.12	30.11	46.19
GEN	4.74	9.54	15.47	20.68	28.70	44.65
GradNorm	5.19	10.38	15.34	19.77	**26.31**	**40.90**
MDS	4.72	**8.92**	**12.69**	**17.88**	**22.77**	**35.51**
RMDS	4.89	9.47	14.19	20.95	27.84	41.48
KNN	5.00	10.81	17.84	25.17	34.11	52.36
RareCount	4.85	**8.84**	**13.82**	19.76	26.59	43.07

F. Extended OOD Detection Results

F.1. Statistical Significance Across SAE Seeds

We assess whether the observed performance differences remain consistent across independently trained SAE instances. For TopK, RareCount and class-conditional EPD are evaluated using the same ten independently trained TopK SAE checkpoints, enabling a seed-wise paired comparison. We therefore use a two-sided Wilcoxon signed-rank test and apply the Holm–Bonferroni correction across the four Near/Far AUROC and FPR95 comparisons. We additionally evaluate the strongest Far-OOD configuration, ReLU RareFraction, against class-conditional EPD. Since these results are obtained from different SAE families and different numbers of independently trained instances ($n = 5$ for ReLU and $n = 10$ for TopK EPD), we use a two-sided Welch's t-test, which does not assume equal variances or equal sample sizes.

Table 21. Seed-wise statistical comparison between class-conditional EPD and RareCount using TopK SAEs. Results are mean ± standard deviation across the same ten independently trained SAE seeds. Δ denotes RareCount minus EPD. Reported p-values are obtained with a paired two-sided Wilcoxon signed-rank test and corrected using Holm–Bonferroni.

Group	**Metric**	**EPD**	**RareCount**	**Δ**	p_{Holm}
Near	AUROC ↑	82.16 $\pm$0.12	80.93 $\pm$0.14	-1.23	**0.0078**
Near	FPR95 ↓	59.51 $\pm$0.42	59.12 $\pm$0.68	-0.39	0.1309
Far	AUROC ↑	93.87 $\pm$0.12	94.88 $\pm$0.07	**$+1.01$**	**0.0078**
Far	FPR95 ↓	34.85 $\pm$1.02	22.89 $\pm$0.58	**-11.96**	**0.0078**

The paired analysis shows that RareCount provides a consistent advantage over class-conditional EPD under Far-OOD shifts. Across ten TopK SAE seeds, RareCount improves mean Far-OOD AUROC by approximately 1.0 point and reduces FPR95 by 12.0 percentage points. Both improvements remain statistically significant after Holm–Bonferroni correction ($p_{\text{Holm}} = 0.0078$). For Near-OOD data, EPD retains a higher AUROC, while the small difference in FPR95 is not statistically significant ($p_{\text{Holm}} = 0.1309$).

Table 22. Far-OOD comparison between TopK class-conditional EPD and ReLU RareFraction. EPD is evaluated across ten independently trained TopK SAEs, while RareFraction is evaluated across five independently trained ReLU SAEs. Since the two groups are not naturally paired and have unequal sample sizes, statistical significance is assessed using a two-sided Welch's t-test. Δ denotes ReLU RareFraction minus EPD.

Metric	**EPD**	**ReLU RareFraction**	**Δ**	**Welch p**
AUROC ↑	93.87 $\pm$0.12	95.31 $\pm$0.19	**$+1.44$**	**8.2×10^{-6}**
FPR95 ↓	34.85 $\pm$1.02	20.01 $\pm$0.73	**-14.84**	**1.35×10^{-12}**

The stronger ReLU configuration further amplifies the Far-OOD advantage. RareFraction improves mean AUROC by 1.44 points and reduces FPR95 by 14.84 percentage points relative to class-conditional EPD, with both differences highly

significant under Welch's t-test. Thus, the Far-OOD improvement is robust to SAE initialization and persists across different sparsity mechanisms, reaching an approximately 15 percentage-point reduction in FPR95 over class-conditional EPD.

F.2. Full Per-Dataset Results Across Backbones

Table 23 reports the complete per-dataset results for DINOv3, complementing the aggregate cross-backbone comparison reported in the main paper. RareCount remains effective across all evaluated OOD sources and across the four SAE architectures. Among our variants, Gated and ReLU provide the strongest overall performance, with ReLU reaching a mean Far-OOD AUROC of 93.09 and FPR95 of 29.15. Although the relative performance of individual post-hoc baselines changes with the backbone, the local concept-rarity signal is consistently preserved when moving from DINOv2 to DINOv3. This supports the conclusion that RareCount is not tied to a specific backbone representation.

Table 23. Post-hoc OOD detection results on the OpenOOD benchmark with ImageNet-1K as in-distribution, using DINOv3 (ViT-B/16, 512×512) as backbone. We report FPR@95 (↓) and AUROC (↑) on the Near-OOD (SSB-Hard and NINCO) and Far-OOD (iNaturalist, Textures, and OpenImage-O) splits, together with their averages. Our SAE variants use the rare-fraction score, with the rarity threshold selected exclusively on the validation set and then frozen for test evaluation. Results for our SAE variants are reported as mean ± standard deviation over 3 SAE seeds; rankings are based on mean performance. The **best**, **second-best**, and **third-best** results are highlighted accordingly. Tied results are assigned the same rank.

	Near-OOD						**Far-OOD**								**ID**
	SSB-Hard		NINCO		*Avg.*		iNaturalist		Textures		OpenImage-O		*Avg.*		ImageNet
Method	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	ACC↑
MSP	77.37	74.44	51.89	**86.42**	64.63	80.43	11.53	97.68	40.22	91.41	25.64	94.87	25.80	94.65	84.81
TempScale	76.69	75.32	48.69	**87.56**	62.69	**81.44**	8.60	98.18	36.55	92.38	**20.27**	**95.65**	21.81	**95.40**	84.81
EBO	**68.62**	**78.33**	**48.48**	83.70	**58.55**	81.01	21.54	91.92	**19.73**	**95.33**	21.70	92.99	**20.99**	93.41	84.81
GEN	**75.38**	**76.29**	**41.81**	**88.62**	**58.60**	**82.46**	**4.30**	**99.06**	29.42	93.78	**15.76**	**96.71**	**16.49**	**96.52**	84.81
GradNorm	83.52	72.10	64.40	85.17	73.96	78.63	**5.10**	**98.55**	**29.41**	**93.99**	**16.60**	**96.10**	**17.04**	**96.21**	84.81
ASH	95.08	49.02	94.51	47.71	94.80	48.36	96.44	45.36	99.10	38.07	94.29	50.25	96.61	44.56	84.81
ReAct	**68.42**	**78.38**	**48.54**	83.68	**58.48**	**81.03**	21.98	91.74	**19.66**	**95.33**	21.82	92.94	21.15	93.34	84.81
MDS	83.86	72.04	64.84	85.25	74.35	78.65	6.81	98.43	42.44	89.47	22.20	95.13	23.82	94.34	84.81
RMDS	84.22	70.03	57.94	81.13	71.08	75.58	**3.93**	**98.58**	33.49	91.74	32.85	92.95	23.42	94.42	84.81
KNN	82.93	67.80	65.13	79.66	74.03	73.73	15.39	95.91	47.80	88.58	31.28	92.26	31.49	92.25	84.81
Ours (TopK)	78.50 ±0.96	70.20 ±0.44	61.28 ±0.67	82.45 ±0.20	69.89 ±0.78	76.33 ±0.13	17.63 ±0.40	95.68 ±0.11	42.31 ±1.11	88.96 ±0.26	34.67 ±0.50	92.69 ±0.05	31.53 ±0.42	92.45 ±0.13	84.81
Ours (BatchTopK)	78.32 ±0.81	70.49 ±0.42	61.30 ±0.59	82.86 ±0.52	69.81 ±0.70	76.68 ±0.46	15.38 ±1.01	96.22 ±0.15	43.37 ±1.15	88.91 ±0.24	34.82 ±0.73	92.88 ±0.16	31.19 ±0.87	92.67 ±0.18	84.81
Ours (Gated)	77.98 ±0.39	70.96 ±0.13	60.65 ±0.58	83.53 ±0.14	69.31 ±0.48	77.24 ±0.13	12.89 ±0.52	96.72 ±0.21	42.24 ±0.26	88.90 ±0.22	32.77 ±0.40	93.39 ±0.07	29.30 ±0.25	93.00 ±0.10	84.81
Ours (ReLU)	77.51 ±0.67	71.06 ±0.28	61.37 ±0.23	83.13 ±0.12	69.44 ±0.45	77.09 ±0.16	13.11 ±0.28	96.69 ±0.12	41.51 ±0.77	89.17 ±0.16	32.83 ±0.25	93.40 ±0.15	29.15 ±0.25	93.09 ±0.13	84.81

F.3. Per-Dataset Score Distributions

Beyond aggregate AUROC and FPR95, we inspect the full distribution of RareCount scores for the default DINOv2 TopK configuration. Figure 13 compares the ID score distribution against each OOD source separately. The separation is visible at the distribution level rather than arising only from a small subset of extreme samples. Far-OOD datasets exhibit the largest shifts, while the Near-OOD sources remain closer to the ID distribution, consistent with their more challenging detection performance.

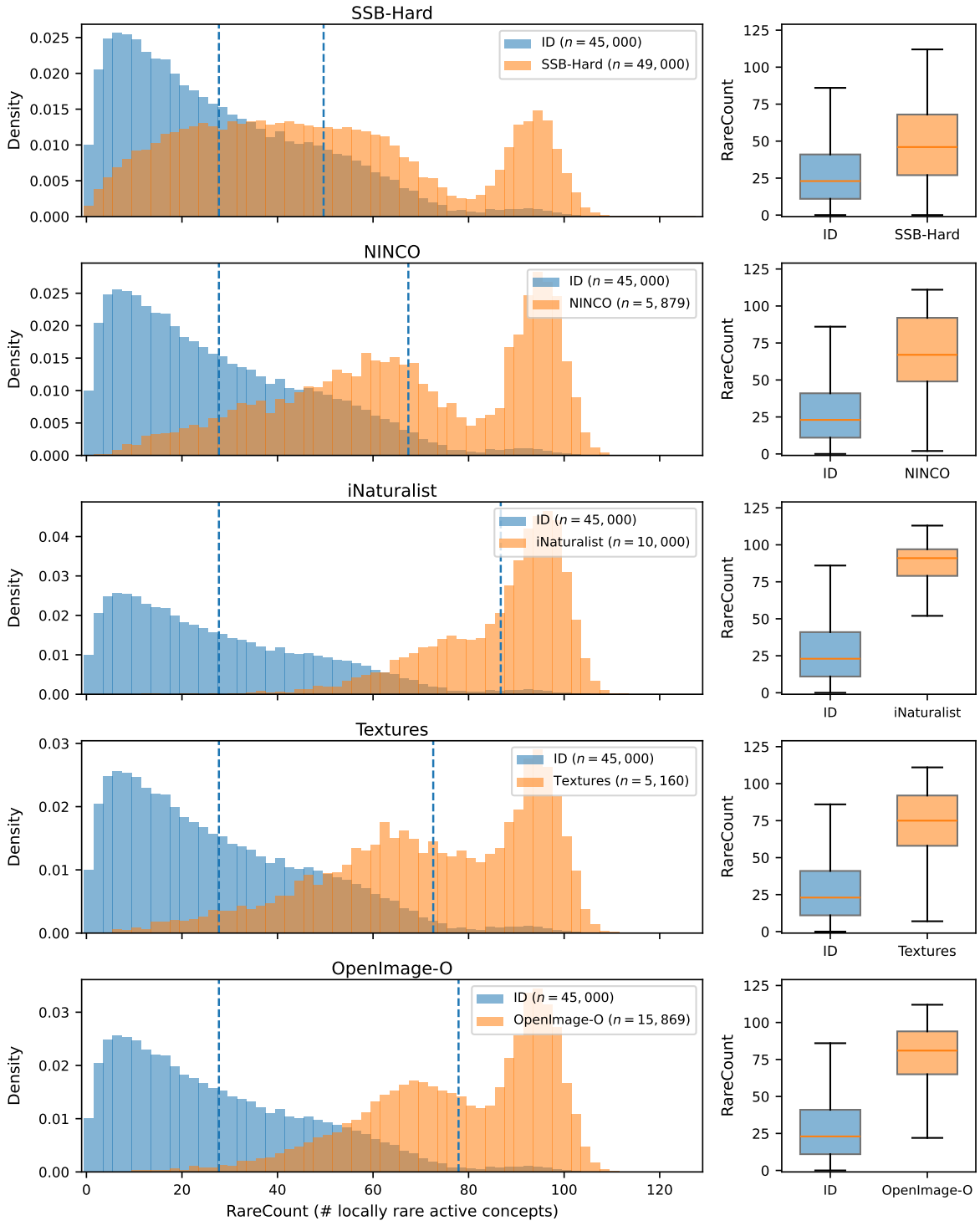


Figure 13. Per-dataset RareCount distributions for the default DINOv2 TopK configuration. Each row compares ImageNet-1K ID samples with one OOD source; histograms show the full score density and the corresponding boxplots provide a compact distribution summary. Dashed vertical lines indicate the mean RareCount of each distribution. Far-OOD datasets show the strongest displacement from the ID distribution, while Near-OOD sources exhibit more overlap.

F.4. Rare-Concept Distributions Across SAE Families

To further examine whether the local-rarity signal depends on a particular sparsity mechanism, we compare the distribution of RareFraction across the four SAE families used in the main experiments. Unlike the raw RareCount, RareFraction normalizes the number of locally rare concepts by the number of active concepts and is therefore directly comparable across architectures with substantially different activation densities. Figure 14 shows the distributions for ID, Near-OOD, and Far-OOD samples. In all four architectures, OOD inputs exhibit a systematic shift toward a larger fraction of locally rare concepts, with the strongest shift generally observed for Far-OOD. This qualitative separation persists from the fixed-sparsity TopK representation to the substantially denser ReLU representation, providing an additional view of the robustness reported quantitatively in the main paper.

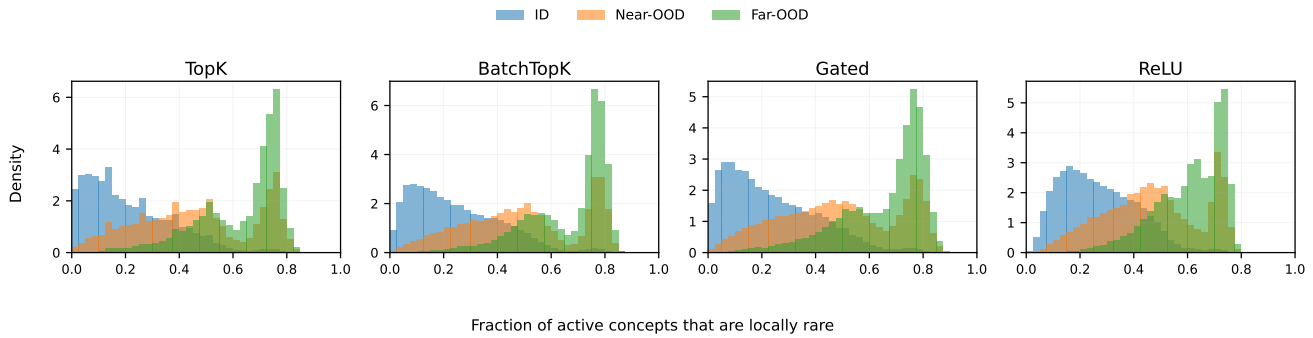


Figure 14. Distribution of RareFraction across the four SAE families. RareFraction measures the fraction of active concepts that are locally rare under the cluster-conditioned ID reference. Near-OOD pools SSB-Hard and NINCO with equal dataset weight, whereas Far-OOD pools iNaturalist, Textures, and OpenImage-O with equal dataset weight. The shift toward larger RareFraction values persists across TopK, BatchTopK, Gated, and ReLU despite their substantially different sparsity regimes.

G. Characterization of Rare and Responsible Concepts

Rare-Concept Enrichment. To characterize the concepts responsible for detection at the dataset level, we aggregate the occurrence of locally rare concepts across samples from each OOD source. For each dataset, this yields a rare-concept frequency profile, which we compare with the corresponding ID profile. Table 24 reports the cosine similarity between the two profiles, the mean enrichment of rare concepts relative to their ID frequency, and the average RareCount. For concept j and dataset D, we define its enrichment over ID as

$$E_j^D = \frac{p_D(j)}{p_{\mathrm{ID}}(j)},$$

where $p_D(j)$ and $p_{\mathrm{ID}}(j)$ denote the fractions of samples in which concept j occurs as a responsible concept. Enrichment is computed only for concepts with non-zero ID frequency. The results show that different OOD sources depart from the ID concept distribution to different degrees. iNaturalist and Textures exhibit the largest mean enrichment ($26.9\times$ and $27.1\times$) and comparatively low similarity to the ID profile, while SSB-Hard remains substantially more ID-like, with a cosine similarity of 0.925 and a mean enrichment of only $6.2\times$. NINCO and OpenImage-O occupy intermediate regimes. These differences indicate that RareCount captures not only the amount of locally unexpected evidence in an image, but also dataset-specific patterns in which concepts become unexpectedly frequent.

Table 24. Dataset-level rare-concept profiles. Cosine similarity is computed with respect to the ID profile, while enrichment measures the increase in concept occurrence relative to ID. Different OOD sources exhibit distinct degrees and patterns of departure from the ID rare-concept distribution.

Dataset	Band	Cosine to ID	Mean enrichment	Mean RareCount
iNaturalist	Far	0.748	26.9×	87
Textures	Far	0.746	27.1×	73
OpenImage-O	Far	0.904	9.8×	78
SSB-Hard	Near	0.925	6.2×	50
NINCO	Near	0.829	15.8×	67

Dataset-Specific Concept Signatures. We next inspect which individual concepts account for these dataset level shifts. For each OOD source, Tab. 25 reports the concepts with the largest increase in occurrence relative to ID, together with their frequency on the OOD dataset and on ID data. The resulting sets differ substantially across datasets: for example, the most enriched concepts on iNaturalist and Textures reach more than $40\times$ their ID frequency, whereas the strongest enrichments on SSB-Hard and OpenImage-O are considerably smaller. Some concepts recur across multiple OOD sources, while others are strongly source-specific, showing that the responsible set can be aggregated into a characteristic signature of the distribution shift.

Table 25. Top enriched rare concepts per OOD dataset. For each concept, we report enrichment over ID frequency, dataset frequency, and ID frequency.

Dataset	Concept	Enrichment	Dataset freq.	ID freq.
iNaturalist	3975	43.1×	0.094	0.002
	2094	34.7×	0.050	0.001
	3329	23.5×	0.031	0.001
	410	23.2×	0.072	0.003
	5965	23.1×	0.098	0.004
	2025	22.9×	0.098	0.004
	1711	22.9×	0.092	0.004
	892	21.9×	0.031	0.001
Textures	7028	45.2×	0.148	0.003
	2776	31.9×	0.067	0.002
	7516	26.0×	0.042	0.002
	139	25.2×	0.051	0.002
	4202	25.1×	0.112	0.004
	5414	22.2×	0.050	0.002
	2289	20.8×	0.074	0.004
	4741	20.3×	0.066	0.003
OpenImage-O	3329	11.0×	0.014	0.001
	6357	10.2×	0.012	0.001
	1263	10.0×	0.038	0.004
	6722	9.7×	0.019	0.002
	1090	9.7×	0.037	0.004
	5345	9.6×	0.014	0.001
	7506	9.3×	0.033	0.003
	6962	9.1×	0.045	0.005
SSB-Hard	6224	6.8×	0.023	0.003
	2935	6.2×	0.016	0.003
	988	6.2×	0.017	0.003
	1631	6.2×	0.016	0.003
	3329	6.1×	0.008	0.001
	7279	6.0×	0.013	0.002
	535	6.0×	0.022	0.004
	1090	6.0×	0.023	0.004
NINCO	5542	20.9×	0.042	0.002
	614	17.5×	0.053	0.003
	2878	16.2×	0.030	0.002
	6071	14.8×	0.055	0.004
	7612	14.7×	0.029	0.002
	7446	14.3×	0.028	0.002
	7010	14.0×	0.018	0.001
	1658	13.8×	0.037	0.003

Dataset-Level Responsible-Concept Signatures. The aggregate responsible-concept profiles also reveal structure across different distribution shifts. Figure 15 summarizes each OOD source using its similarity to the ID rare-concept profile and its mean concept enrichment. Near-OOD datasets remain comparatively closer to the ID signature, whereas Far-OOD datasets exhibit stronger enrichment and a larger departure from the reference profile.

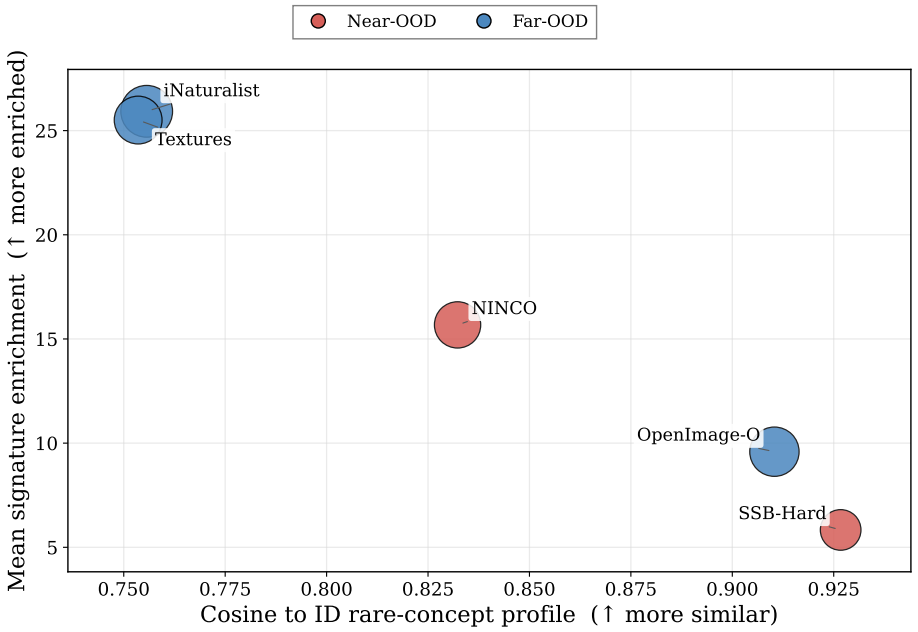


Figure 15. Dataset-level geometry of responsible-concept signatures. The horizontal axis measures cosine similarity between each dataset's rare-concept profile and the corresponding ID profile, while the vertical axis reports mean concept enrichment over ID.

To complement the axis-aligned summary above, we embed the complete responsible-concept signatures in two dimensions using multidimensional scaling on cosine dissimilarities. Figure 16 provides a direct visualization of their relative organization.

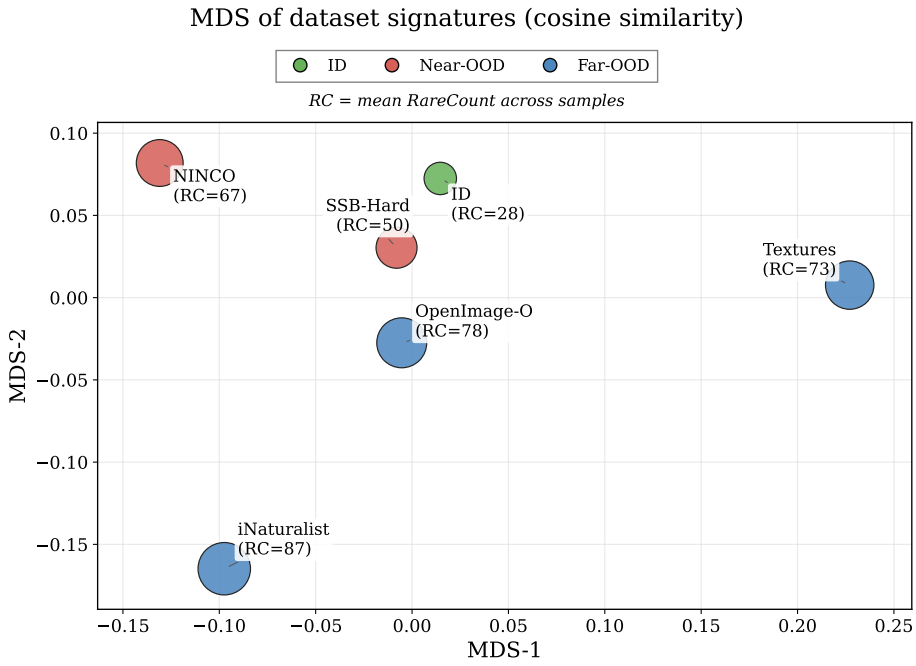


Figure 16. Two-dimensional multidimensional-scaling embedding of dataset-level responsible-concept signatures using $1 - \text{cosine}$ as the dissimilarity measure. ImageNet-1K provides the ID reference; colors distinguish Near-OOD and Far-OOD sources, and marker size is proportional to mean RareCount.

H. Quantitative Validation of Spatial Grounding

Grounding Evaluation Protocol. To evaluate whether the concepts responsible for RareCount are spatially grounded, we apply the same CLS-trained SAE to the patch tokens produced by the corresponding frozen backbone. For each image, the concepts in the responsible set $\mathcal{R}(x)$ are selected from the CLS representation, while their patch-level activations define spatial concept maps over the image. We evaluate these maps through three complementary tests: CLS–patch consistency, concept-specific occlusion, and alignment with patch-level atypicality. We perform these analyses using a TopK-SAE trained on DINOv2 CLS representations.

CLS–Patch Transfer. Table 26 verifies that transferring the CLS-trained SAE to patch tokens produces concept-specific rather than generic spatial maps. Correlations with low-level patch statistics and generic spatial patterns remain small, while CLS–patch agreement is consistently high across all OOD datasets, with AUROC ranging from 0.811 to 0.879. This supports the use of patch activations as a spatial proxy for the concepts identified at the CLS level.

Table 26. Sanity checks for concept maps. C1 measures whether maps follow low-level patch statistics such as norm or the first PCA direction; lower is better. C2 measures whether maps collapse to generic spatial/objectness patterns; lower is better. C3 measures CLS–patch agreement; higher AUROC indicates stronger consistency between image-level concept activation and patch-level localization.

Dataset	C1 norm ↓	C1 PCA1 ↓	C2 spec. ↓	C3 AUROC ↑
iNaturalist	0.078	0.177	0.064	0.811
Textures	0.109	0.218	0.103	0.879
OpenImage-O	0.097	0.199	0.082	0.837
NINCO	0.097	0.252	0.094	0.842
SSB-Hard	0.091	0.222	0.087	0.827

Concept-Specific Occlusion. We next test whether the identified regions are specifically associated with the corresponding concept. As shown in Tab. 27, occluding a concept's own high-activation region suppresses its activation more strongly than occluding regions associated with other concepts. The median self–cross effect is positive on every dataset, with self occlusion winning on 59–79% of images, and the effect remains significant for all five OOD sources after Holm–Bonferroni correction.

Table 27. Image-level significance of occlusion selectivity. Each image is treated as an independent unit. We report the median self–cross difference, the fraction of positive differences, and one-sided Wilcoxon signed-rank p-values corrected jointly across the five datasets using Holm–Bonferroni.

Dataset	n_{img}	Median self–cross	Frac. > 0	p_{Holm}
SSB-Hard	112	+0.0153	0.61	8.37×10^{-3}
NINCO	118	+0.0501	0.79	1.77×10^{-10}
iNaturalist	107	+0.0211	0.59	3.74×10^{-3}
Textures	119	+0.0186	0.64	1.06×10^{-4}
OpenImage-O	106	+0.0254	0.70	2.21×10^{-5}

Responsible-Map Alignment with Patch Atypicality. Table 28 evaluates whether responsible-concept maps concentrate on locally atypical image regions. The correlation is negligible on ID validation ($\rho = 0.005$) but consistently positive on all OOD datasets ($\rho = 0.038$–0.089). These associations are stronger than both active non-rare concept controls and spatially permuted null maps, indicating that the concepts contributing to RareCount preferentially localize atypical regions of OOD inputs.

Table 28. Patch-level OOD relevance measured through Mahalanobis atypicality. Rare/correct reports the median concept-level correlation between rare-concept heatmaps and patch atypicality; Rand/correct uses active non-rare concepts, and Rare/null uses spatially permuted atypicality. Frac. > 0 is the fraction of positive rare/correct correlations. For Δ vs ID, correlations are first averaged across rare concepts within each image, and the reported value is the difference between the OOD and ID medians of these per-image averages. POS indicates a 95% bootstrap confidence interval entirely above zero.

Split	Rare/correct	Rand/correct	Rare/null	Frac. > 0	Δ vs ID	Verdict
ID	+0.005	-0.027	+0.001	0.51	–	n.s.
SSB-Hard	+0.047	+0.013	+0.006	0.64	+0.050	POS
NINCO	+0.089	+0.041	+0.002	0.70	+0.099	POS
iNaturalist	+0.038	+0.013	+0.002	0.62	+0.048	POS
Textures	+0.066	+0.026	+0.004	0.67	+0.075	POS
OpenImage-O	+0.048	+0.012	+0.004	0.63	+0.050	POS

Spatial Structure of Concept Maps. Finally, Tab. 29 characterizes the spatial organization of the concept maps. Positive Moran's I values across all datasets indicate that concept activations are spatially structured rather than randomly scattered.

Table 29. Relationship between spatial structure and occlusion selectivity. Med. diffus. measures how spatially spread a concept map is. Moran's I measures spatial coherence/autocorrelation. The Spearman correlations ρ(diff., select.) and $\rho(I$, select.) indicate whether more diffuse or more coherent concepts are more locally attributable under occlusion. The pooled row is computed by concatenating all valid concept–image observations across the five OOD datasets; medians and Spearman correlations are then recomputed on the resulting pooled sample.

Dataset	Med. diffus.	ρ(diff., select.)	Med. Moran's I	$\rho(I$, select.)
iNaturalist	0.090	+0.278	0.162	+0.226
Textures	0.240	-0.058	0.203	+0.007
OpenImage-O	0.129	+0.146	0.200	+0.177
NINCO	0.146	+0.078	0.305	+0.291
SSB-Hard	0.128	+0.174	0.278	+0.157
Pooled	0.149	+0.100	0.234	+0.182

I. Additional Qualitative Examples

We provide additional qualitative examples of the spatial grounding induced by the responsible set $\mathcal{R}(x)$ across all OOD sources. For each input, we visualize the four responsible concepts with the largest image-level activation. Activation magnitude is used only to rank the displayed responsible concepts and does not enter the RareCount score. These examples complement the quantitative grounding analysis by showing that responsible concepts capture distinct object parts, structural regions, and local texture patterns across diverse OOD inputs.

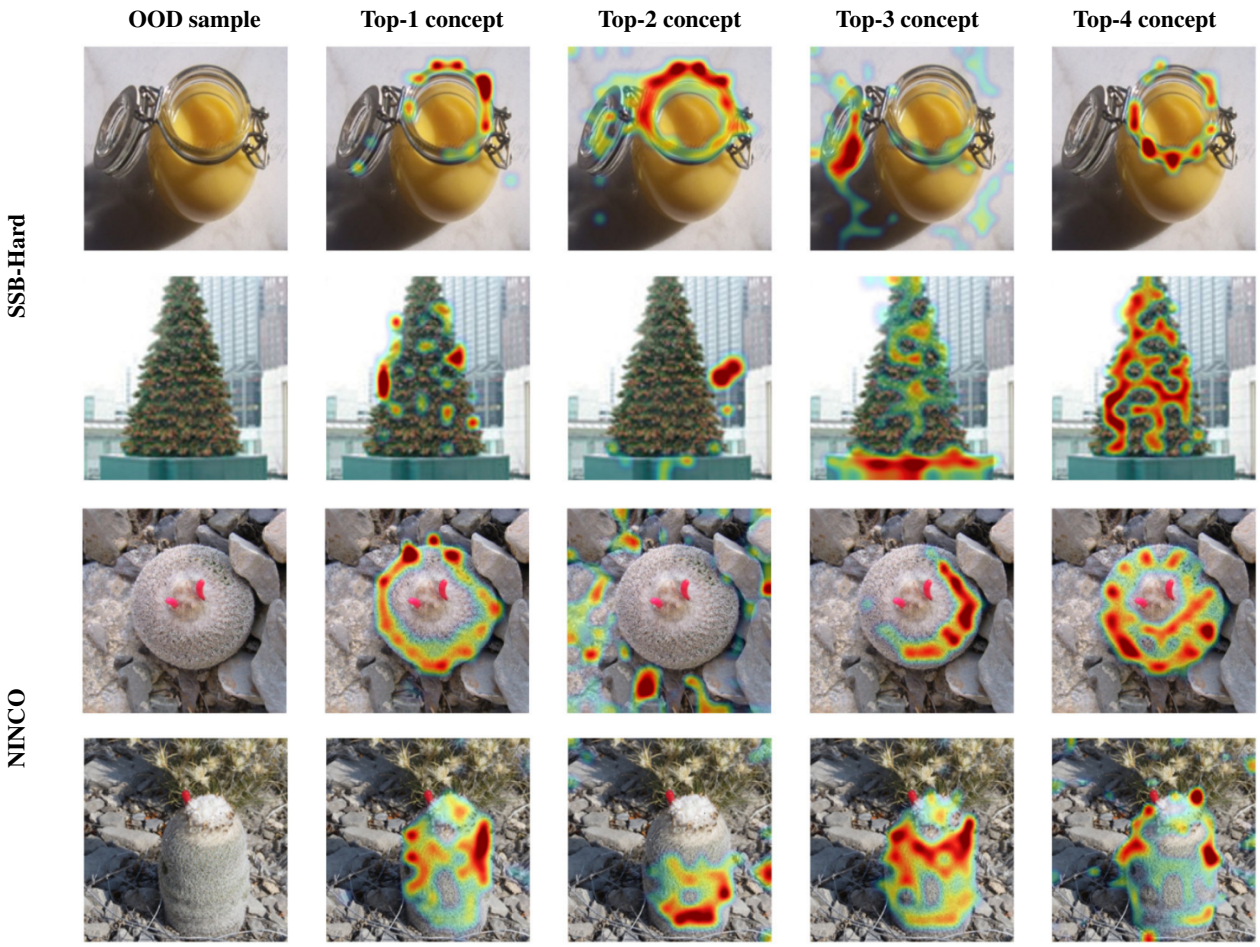


Figure 17. **Additional qualitative grounding examples on Near-OOD data.** Each row shows an input followed by the spatial activation maps of the four highest-activation concepts within its responsible set $\mathcal{R}(x)$. The examples span SSB-Hard and NINCO and show that different responsible concepts capture complementary structural regions rather than collapsing to a single generic spatial pattern. Activation magnitude is used only to rank the displayed concepts and does not enter the RareCount score.

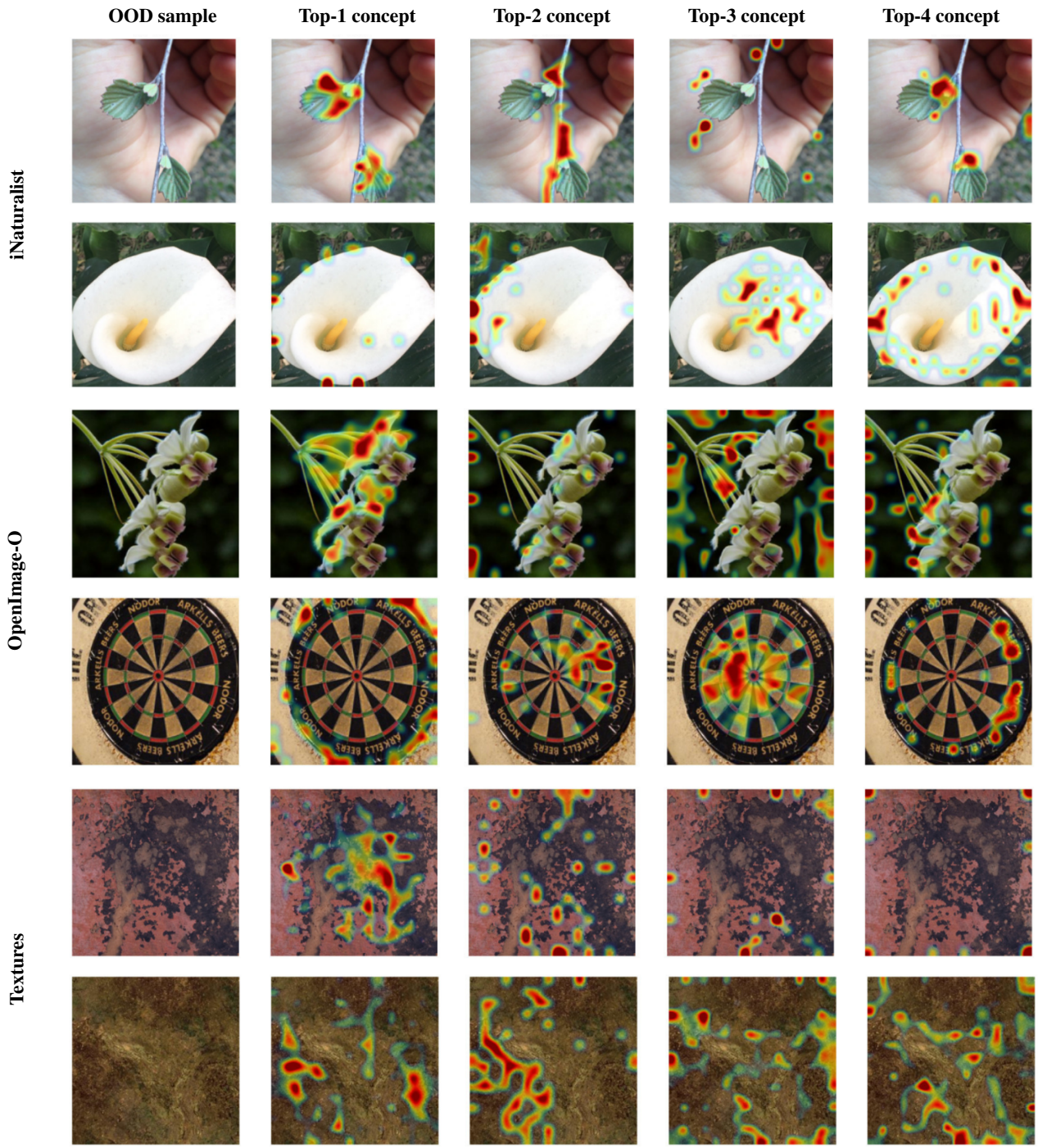


Figure 18. **Additional qualitative grounding examples on Far-OOD data.** Each row shows an input followed by the spatial activation maps of the four highest-activation concepts within its responsible set $\mathcal{R}(x)$. The examples span iNaturalist, OpenImage-O, and Textures, illustrating that responsible-concept maps remain spatially structured across natural objects, geometric structures, and texture-only inputs. Activation magnitude is used only to rank the displayed concepts and does not enter the RareCount score.

Example Score Decompositions. Figure 19 illustrates the sample-level decomposition underlying RareCount. The ID example contains only 23 locally rare concepts, whereas the SSB-Hard sample contains 91.

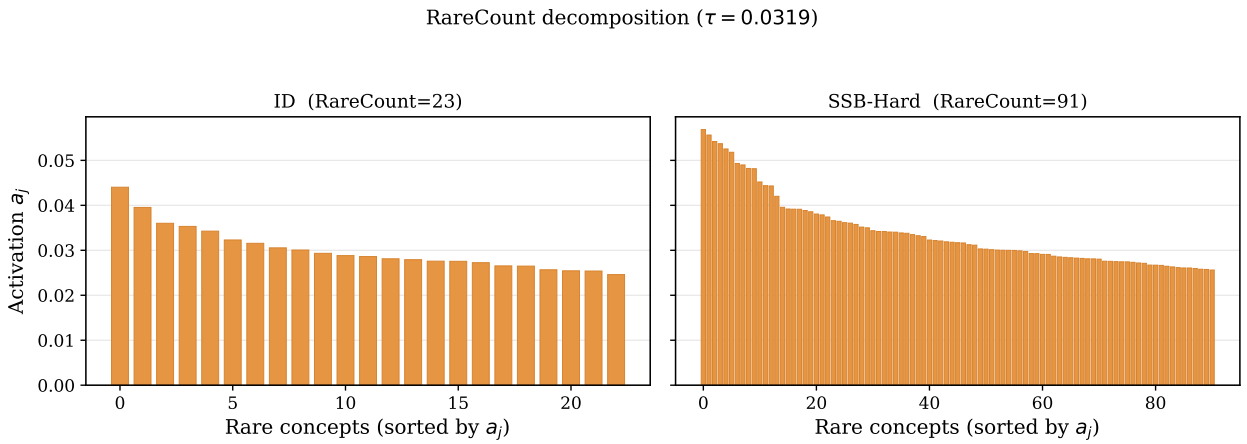


Figure 19. Example decomposition of RareCount for an ID sample (left) and an SSB-Hard OOD sample (right). Bars show the activation magnitude a_j of concepts classified as locally rare, sorted by activation. Although both inputs activate the same number of concepts under the TopK SAE, the OOD sample contains substantially more locally rare concepts (91 versus 23). Activation magnitude is shown only for visualization and does not enter the RareCount score.

J. Controlled Object-Level Demo

We additionally provide a controlled object-level demonstration to directly inspect RareCount behavior in simple real-world scenes.

Acquisition Protocol. We record short videos using a fixed tabletop setup and extract representative frames from each sequence. Each frame is processed independently using the same pipeline adopted throughout the paper, without exploiting temporal information. We additionally provide representative video sequences in the supplementary material, showing the corresponding RareCount responsible-concept heatmaps throughout the recorded scenes. We use a light, approximately uniform background and a fixed camera viewpoint to reduce visual clutter and facilitate spatial inspection of the responsible concepts. Multi-object scenes are constructed so that ID and OOD objects are simultaneously visible. These scenes allow us to examine whether the responsible concepts localize on the OOD object rather than on other image regions or on the accompanying ID objects.

Object Selection. We use six ordinary objects spanning different geometries, sizes, appearances and semantic functions: a *mug*, a *pen*, a *bottle*, a *mouse*, *headphones* and a *toothbrush*. The mug, pen, bottle, and mouse are used as ID objects, whereas the headphones and toothbrush are used as OOD objects. The selected ID objects have a clear correspondence with ImageNet classes. The headphones and toothbrush are semantically distinct from these familiar controls while providing complementary visual configurations. The headphones form a large articulated structure whose headband and ear cups occupy multiple spatial regions, whereas the toothbrush is considerably smaller and elongated. These two objects therefore allow us to inspect whether the qualitative grounding behavior persists across different object geometries and scales.

Score Decomposition and Visualization. For each frame we compute the responsible set of rare active concepts defined in Sec. 3.1. For qualitative inspection, we follow the visual grounding procedure described in Sec. 4.5. In addition to the original frame, we report four complementary views: the two responsible concepts with the largest CLS activation, an aggregate map of the *Top-5* highest activation responsible concepts, and an aggregate over the complete responsible set. The latter is expected to be more diffuse, since RareCount also counts responsible concepts with relatively small activation magnitudes, and is therefore used mainly as a sanity check rather than as the primary explanatory view.

Single-Object Characterization. We first inspect each object independently, isolating its response before introducing compositional effects. Figure 20 shows the grounding of the responsible concepts for the considered objects. The ID objects generally exhibit weaker and more diffuse activation maps, without a consistent concentration on specific object regions. In contrast, the headphones and toothbrush show more localized responsible evidence on characteristic structural regions. For the headphones, for example, the strongest concepts concentrate on the headband and ear cups.

Familiar-Composition Controls. We evaluate two scenes containing only familiar objects *pen* + *mug* and *mug* + *bottle*. These scenes are important controls since a multi-object composition is not intrinsically OOD. They therefore test whether simply increasing the number of visible objects is sufficient to reproduce the spatial patterns observed in the mixed ID/OOD scenes.

Fixed OOD Object under Varying Familiar Context. A central part of the demonstration keeps the OOD objects fixed while varying the accompanying familiar object. This controlled variation is informative because the OOD object remains approximately unchanged while its surrounding semantic context varies. If the responsible-concept grounding merely reflected generic saliency or the presence of multiple objects, its spatial support would be expected to change substantially with the accompanying object. Instead, across the selected scenes the dominant responsible evidence remains associated with the OOD object, while the ID object is comparatively weakly involved. The *mug + headphones* scene provides the clearest example. The highest activation responsible concept localizes primarily on a restricted structural region of the headphones, particularly along the headband. When considering the five strongest responsible concepts, the spatial support extends over a larger portion of the headphones, while the mug remains largely uncovered. The *toothbrush + pen* scene is more challenging and its grounding is generally more diffuse than that of the headphones. Nonetheless, the responsible evidence remains preferentially associated with the region containing the toothbrush rather than spreading uniformly across the entire scene. This comparison suggests that the qualitative behavior is not restricted to a particular geometry or scale.

Occlusion Stress Test. We additionally consider a simple occlusion setting to qualitatively inspect how the responsible-concept grounding changes when part of the visible scene is progressively obscured. This experiment is intended only as a qualitative stress test: it allows us to examine whether the spatial support of the responsible concepts remains associated with the visible evidence rather than shifting arbitrarily to unrelated image regions. The corresponding sequence is included in the supplementary video material.

Complete Responsible-Set Sanity Check. Grounding obtained by aggregating all the responsible concepts is consistently less selective than the Top-5 and individual-concept views. This is expected because the complete set also includes concepts with smaller activation magnitudes. The resulting increase in visual diffusion does not invalidate the individual groundings; rather, it motivates separating the detector definition from the visualization strategy. All responsible concepts contribute equally to RareCount, whereas activation magnitude provides a natural criterion for prioritizing the most strongly expressed concepts when constructing a compact visual explanation.

Discussion. Taken together, the controlled examples probe three complementary properties. First, the single-object experiments characterize the response associated with each object independently. Second, the familiar multi-object scenes verify that the presence of multiple objects alone does not reproduce the explanatory pattern observed in mixed ID/OOD scenes. Third, in mixed scenes, the dominant responsible-concept grounding remains associated with the OOD object across different familiar contexts. These observations complement the quantitative grounding experiments reported in Sec. 4.5. They should not be interpreted as establishing a causal explanation of the underlying VFM. Rather, they demonstrate that, in a directly inspectable setting, the concepts that constitute the RareCount score can be spatially associated with the region introducing the OOD evidence, while visible familiar content can remain largely unaffected.

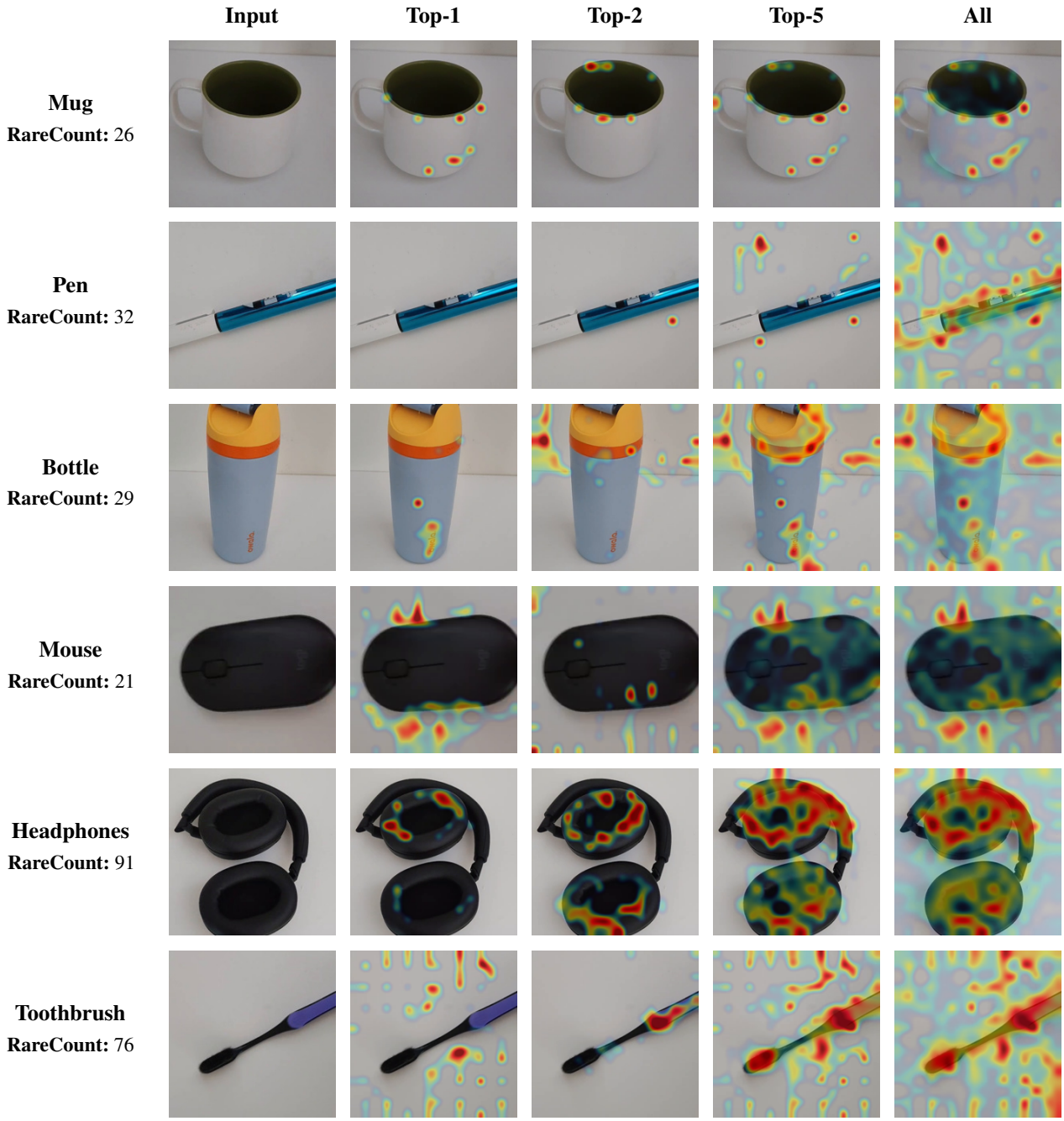


Figure 20. **Qualitative grounding results on individual objects from the controlled tabletop demo.** Each row shows an input frame together with the activation maps of the highest-ranking responsible concepts. The leftmost label reports the object shown in the scene and the corresponding RareCount score. Top-1 and Top-2 show individual responsible concepts, while Top-5 and All aggregate the activation maps of the five highest-ranking and all responsible concepts, respectively.

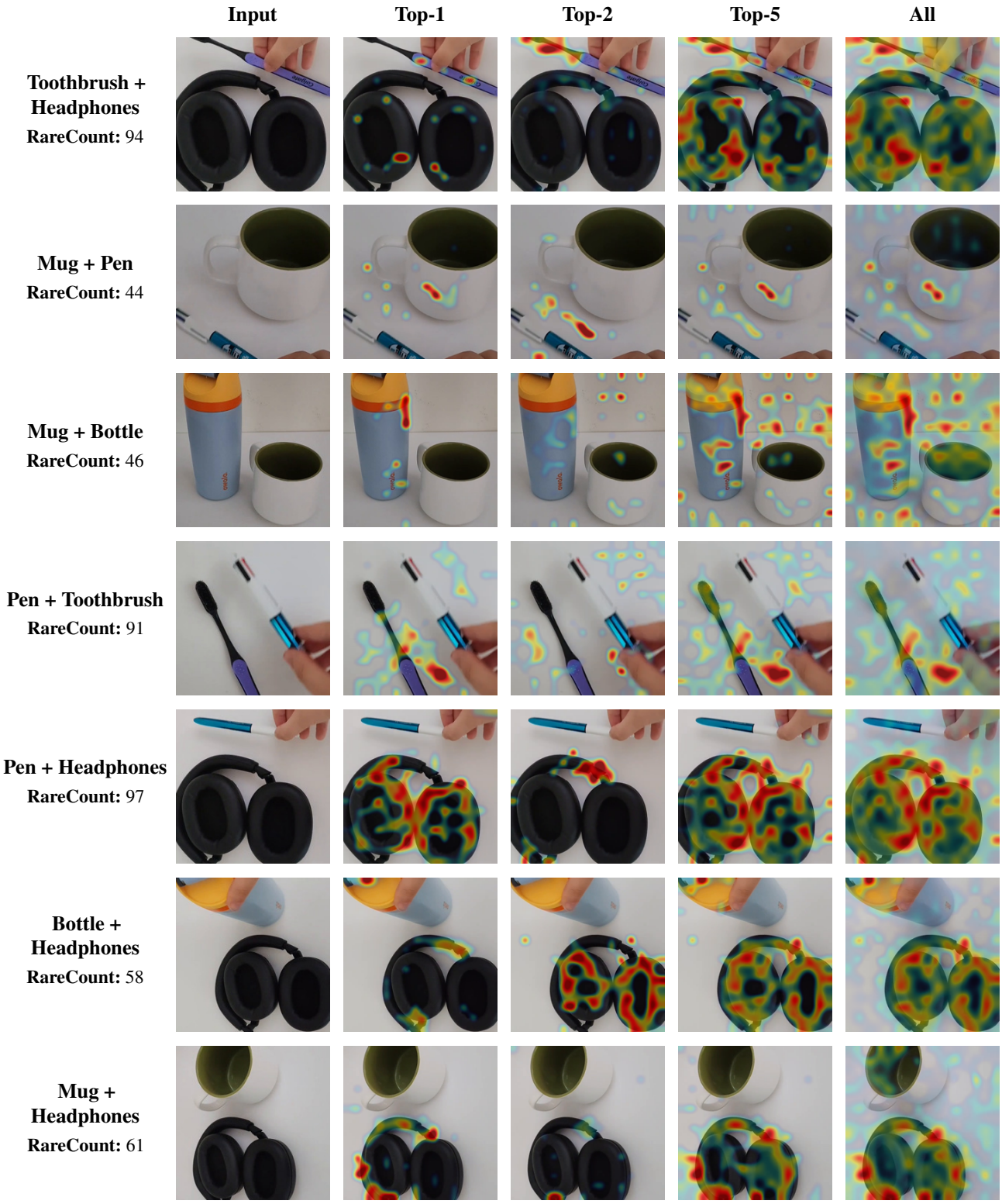


Figure 21. **Qualitative grounding results on the controlled tabletop demo.** Each row shows an input frame together with the activation maps of the highest-ranking responsible concepts. The leftmost label reports the object pair appearing in the scene and the corresponding RareCount score. Top-1 and Top-2 show individual responsible concepts, while Top-5 and All aggregate the activation maps of the five highest-ranking and all responsible concepts, respectively.