Reconstruction, Not Localization: A Mechanistic Account of Sparse-Feature OOD Detection in Vision Transformers

Anonymous authors
Paper under double-blind review

Abstract

Sparse autoencoder (SAE) features have been proposed as an interpretable basis for out-of-distribution (OOD) detection, inviting the reading that SAEs contain localized "OOD units". We test the controls such a claim needs on frozen vision transformers (13 OOD sets, CIFAR-10-C, a 100-class in-distribution (ID) dataset, and ImageNet-1K scale). In both ReLU and exact-k TopK SAEs (three seeds each), frequently firing units are seed-idiosyncratic, fixed sets of up to five units add about one point of confidence collapse on ImageNet-O beyond the plain reconstruction round-trip, and the largest-contribution units are *in-distribution anchors*. The collapse is instead the SAE encode→decode round-trip: it lowers a frozen ImageNet head's confidence on ImageNet-O by 70% versus 11% on ID (47% with its native CLS input), whereas matched noise moves it by only 7%; a PCA round-trip reproduces the collapse (73%). It is head-dependent: at a fixed ID, the drop grows monotonically with the number of head classes (from −3% at 10 to 70% at 1,000), while heads trained on the ID label space stay inert (4.5%). We call this **Mismatched Wide-Head Reconstruction Sensitivity**; it persists with heads trained at each width and on a second backbone (DeiT-B/16). Against published baselines, SAE-based detection is competitive but not better, and a PCA basis matches it. We release a six-step **Sparse Reconstruction Audit** with margin-reported decision criteria, all JSON artifacts, and a runnable demo; on a released third-party SAE trained on ImageNet it finds no such effect, and a controlled comparison at the same hook ties the effect to an SAE fit on the audited ID. The contribution is not a stronger detector but a tested boundary on what sparse-feature OOD analysis can support.

1 Introduction

Sparse autoencoders (SAEs) have been proposed as an interpretable basis for out-of-distribution (OOD) detection in vision models (Olson et al., 2025; Oh et al., 2026). A natural reading of these results — one the cited works do not themselves claim — is that SAEs supply a vocabulary of localized "OOD units" whose firing flags off-manifold inputs. If such units existed and were stable, a deployed classifier could be audited by watching a handful of features, and unit-level interpretability claims increasingly carry safety arguments of exactly this kind (Hendrycks et al., 2021c). We ask what controls a unit-level OOD claim needs, run them on frozen vision transformers (ViTs), and report what actually drives sparse-feature OOD signals. Our contributions:

- **A negative result on OOD units.** In both a ReLU and an exact-k TopK SAE family (three seeds each), individual units are seed-idiosyncratic (~8% ReLU and ~10% TopK of units have a >0.7-correlated partner across seeds; frequently firing units stay near 15% with ~70× more data), fixed sets of up to five units add ~1 point of ImageNet-O confidence collapse beyond the plain reconstruction round-trip, and the largest-contribution units gate *in-distribution* (ID) confidence (ID anchors) (§5, §6).
- **The collapse is reconstruction.** The SAE encode→decode round-trip lowers a frozen ImageNet head's max-softmax probability (MSP) on ImageNet-O by 70% versus 11% on ID, while Gaussian noise of matched magnitude moves it by only 7%. A PCA round-trip re-

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

produces the collapse (73%), so the effect belongs to the encoder–code–head system as a whole (§5).

- **Head dependence, deconfounded.** At a fixed ID (CIFAR-10), fixed head weights, and a fixed SAE, the OOD drop grows monotonically with the number of ImageNet classes the head may choose between (from -3% at 10 classes to 70% at 1,000), while the ID drop stays at 9–14%; heads *trained* at each width show the same trend, and heads trained on the ID label space are inert (4.5% with a 10-class probe, 0.7% with a 100-class probe). With ImageNet-100 as a narrow in-domain ID, a matched head is inert while broader heads that include the ID classes drop OOD confidence by 10% at no ID cost. We call this **Mismatched Wide-Head Reconstruction Sensitivity**. It persists when the head receives its native CLS input (47% vs 11%) and replicates on a second supervised backbone, DeiT-B/16, though with weaker selectivity (§5.1).
- **A reusable audit.** A six-step **Sparse Reconstruction Audit** with margin-reported decision criteria, released with all JSON artifacts and a runnable demo (§3), and run on a released third-party SAE, where the confidence-collapse effect is absent; a controlled comparison at the same hook ties it to an SAE fit on the audited ID (§5.1).

In short, sparse-feature confidence collapse is a property of reconstruction and of the head's label space, not of individual units. Detection performance is secondary here: SAE reconstruction error is competitive with published baselines but not better than them, and it is matched by a PCA basis (§4); at ImageNet-1K scale a sparsity-matched SAE trails the best dense baseline on all seven OOD sets (§7). The sparse-specific residue is descriptive: a sharper attribution spectrum than dense or PCA bases, which is, however, largely a re-encoding of reconstruction error (§6).

What should change in practice. Unit-level claims about "OOD features" need seed-stability and ablation controls before they carry weight (§6 shows how quickly they dissolve). Confidence-based reconstruction diagnostics depend on the whole encoder–code–head system: the same code is diagnostic under a mismatched wide head and nearly inert under a matched one (§5.1).

2 Related Work

OOD scoring. MSP (Hendrycks and Gimpel, 2017), Energy (Liu et al., 2021), Mahalanobis (Lee et al., 2018), KNN (Sun et al., 2022), ViM (Wang et al., 2022), and ReAct (Sun et al., 2021) score dense embeddings or logits; see Yang et al. (2024) for a survey and Yang et al. (2022) for a standard benchmark whose ViM and Mahalanobis definitions our published-baseline rows follow. Reconstruction-error scoring with autoencoders and PCA (Guan et al., 2023) predates sparse dictionaries by decades, so what we add is the *control decomposition and head dependence*; label-space breadth is known to matter for vision–language OOD scoring (Jiang et al., 2024), and §5.1 isolates it for reconstruction.

Sparse features and OOD. SAEs decompose activations into interpretable units (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024; Gao et al., 2024); ViTs have also been interpreted with attention-based relevance propagation (Chefer et al., 2021), and language-model work trains decoders and oracles that explain activations (Huang et al., 2025; Li et al., 2025; Karvonen et al., 2025). Olson et al. (2025) report that vision-SAE features "can be used for OOD detection"; Oh et al. (2026) train TopK SAEs on ViT class tokens and read class-specific activation profiles as an ID signature that "OOD samples systematically disrupt"; and Chhetri et al. (2026) propose SAE-perturbation confidence probes; concurrent work reports that ID and OOD data activate distinct sets of sparse features (Karmacharya et al., 2026). None claims a seed-stable vocabulary of individual OOD units. What we test is the unit-level reading their aggregate signatures invite: the profile-level ID signature of Oh et al. (2026) is consistent with our ID-anchor finding, and Chhetri et al. (2026) perturb deeper-layer representations along class-conditioned SAE concept vectors and measure logit stability, which is related to, but distinct from, the SAE round-trip our §5 controls analyze. Superposition theory predicts that learned dictionaries are not canonical bases (Elhage et al., 2022); in language models, SAEs trained on the same data with different seeds share only about 30% of their features, with TopK SAEs less stable than ReLU SAEs (Paulo and Belrose, 2026), and random baselines can match trained SAEs on several interpretability and editing evaluations (Korznikov et al., 2026); SAEs on large vision models likewise produce sharply different dictionaries across training runs (Fel et al., 2025). We measure this instability in a vision-SAE OOD pipeline, where it bears

directly on unit-level OOD claims, and find that in our regime the TopK/ReLU ordering of Paulo and Belrose (2026) reverses (§6).

3 Setup and the Sparse Reconstruction Audit

Encoders and heads. The primary encoder is a frozen `vit_base_patch16_224` (Dosovitskiy et al., 2021) (timm `augreg2_in21k_ft_in1k` weights), mean-pooled over all 197 post-LayerNorm tokens (768-d); DeiT-B/16 (Touvron et al., 2021), ResNet-50 (He et al., 2016), and self-supervised DINOv2-B/14 (Oquab et al., 2024) are cross-architecture checks, and ImageNet-1K serves as a scale ID (§7). Images are resized (shorter side to 224, bilinear), center-cropped to 224, and normalized with the ImageNet mean and standard deviation for every backbone and dataset; this matches the DeiT and ResNet-50 configurations, whereas timm's configuration for the ViT specifies mean = std = 0.5. DeiT-B/16, for which this normalization is native, replicates the head-width effect (§5.1). Under timm's preprocessing (bicubic resizing, crop ratio 0.9, mean = std = 0.5), native-head ImageNet top-1 accuracy rises from 81.0% to 85.3% on 10,000 validation images, while the headline round-trip, re-extracted with the SAE retrained, is essentially unchanged (ImageNet-O 75.5% vs ID 9.1%, matched noise 6.7%; with our preprocessing 70.3%, 10.7%, and 7.2%; `outputs/results/normalization_sensitivity.json`). The main ID is CIFAR-10 (Krizhevsky, 2009). Detection (§4) scores MSP/Energy/ViM with a **CIFAR-10 linear probe** (logistic regression, temperature-scaled on a held-out split; ID accuracy 0.933), which removes the ImageNet-head leakage that makes DTD, Food-101, and STL-10 near-distribution. The mechanism audit (§5) perturbs the frozen **ImageNet-1k head**. That head was trained on the CLS token, so applying it to mean-pooled features is itself a mismatch; we report the CLS-pooled regime alongside (Appendix B.2).

SAE. A ReLU SAE (768→2048) trained only on CIFAR-10 train embeddings, with an L0-bisection sparsity scheduler whose λ ceiling stops it at $\bar{k} \approx 132$ active units (6.4% of h); a TopK SAE (Gao et al., 2024) at k=131 is the second family (Appendix C.4). Other IDs and backbones get their own SAEs.

Scorers. SAE-Recon (reconstruction error), SAE-ActMahalanobis (Mahalanobis distance in SAE activation space), dense baselines (MSP, Energy, KNN, and the published ViM (Wang et al., 2022) and Mahalanobis (Lee et al., 2018) with tied within-class covariance, all fit on ID data only; we also report the simplified ViM-style and total-covariance Mahalanobis variants of Appendix E.1, whose ViM weight was chosen on SVHN), PCA-Recon (k=256), ReAct-Energy, and a DINOv2-KNN reference. Attribution uses per-unit reconstruction contributions $\delta_i = a_i^2 \|W_{\mathrm{dec}[:,i]}\|^2$ and their normalized entropy $H_{\mathrm{norm}} = H/\log k(x)$. SAE-Recon is the detector-facing quantity; the MSP drop is mechanism-facing and is not proposed as a detector.

The Sparse Reconstruction Audit. A six-step protocol (`scripts/sparse_reconstruction_audit.py`) that produces one JSON report for an encoder + SAE + head: (1) SAE-Recon AUROC; (2) MSP drop under the round-trip and its AUROC as a diagnostic; (3) attribution concentration (H_{norm}, modal-unit share); (4) a matched-Gaussian noise control; (5) ID-vs-OOD selectivity; (6) a k=256 PCA round-trip control. Verdicts use explicit thresholds (e.g. OOD-selective if the OOD drop is $\geq 5\times$ the ID drop and ≥ 20 points; diffuse if H_{norm} exceeds the run's own ID value by ≥ 0.10) and carry their margins. The thresholds were set on the regimes studied here and are not calibrated for false-positive control on unseen systems; verdicts classify mechanism and are not a safety certification.

4 Detection Performance

Table 1. AUROC, CIFAR-10 as ID. OOD sets: SVHN (Netzer et al., 2011), LSUN (Yu et al., 2015), TinyImageNet (Le and Yang, 2015), CIFAR-100 (Krizhevsky, 2009), STL-10 (Coates et al., 2011), MNIST (LeCun et al., 1998), DTD (Cimpoi et al., 2014), Food-101 (Bossard et al., 2014), Places365 (Zhou et al., 2017), and the ImageNet-derived (Deng et al., 2009) sets ImageNet-O (Hendrycks et al., 2021a), ImageNet-R (Hendrycks et al., 2021b), and ImageNet-Sketch (Wang et al., 2019); SUN397 (Xiao et al., 2010) and iNaturalist (Van Horn et al., 2018) are added at scale (§7). SAE rows: seed-42 model, 3-seed stability in Appendix C.2; FPR95/AUPR in Appendix A.1. Best non-DINOv2 value per column in **bold**; our methods in *italics*. ViM (Wang et al.) and Mahalanobis (Lee et al.) are the published baselines; the ViM-style (simplified) and total-covariance Mahalanobis rows are the simplified variants of Appendix E.1. TImg = TinyImageNet; LSUN-r/c = LSUN resize/crop; IN-O/R/Sk = ImageNet-O/R/Sketch.

Method	SVHN	LSUN-r	LSUN-c	TImg	Places365	CIFAR-100	STL-10	MNIST	DTD	Food-101	IN-O	IN-R	IN-Sk
MSP	0.932	0.925	0.926	0.870	0.901	0.904	0.571	0.918	0.916	0.911	0.914	0.861	0.881
Energy	0.958	0.936	0.956	0.875	0.913	0.920	0.571	0.962	0.928	0.910	0.925	0.868	0.882
Mahalanobis (total cov.)	0.821	0.829	0.720	0.955	0.990	0.843	0.792	0.996	**1.000**	**1.000**	0.996	0.997	0.997
KNN	0.855	0.754	0.822	0.930	0.993	0.852	0.766	0.766	**1.000**	**1.000**	0.993	0.996	0.991
ViM-style (simplified)	0.961	0.945	**0.959**	0.893	0.939	0.926	0.588	0.986	0.960	0.949	0.954	0.907	0.927
ViM (Wang et al.)	**0.964**	**0.949**	0.940	**0.965**	0.991	**0.946**	0.743	**0.997**	**1.000**	**1.000**	**0.998**	0.995	0.997
Mahalanobis (Lee et al.)	0.827	0.836	0.730	0.956	0.990	0.848	0.791	0.996	**1.000**	**1.000**	0.996	0.997	0.997
SAE-Recon	*0.812*	*0.812*	*0.743*	*0.942*	*0.992*	*0.835*	*0.777*	*0.988*	***1.000***	***1.000***	*0.997*	*0.997*	*0.999*
SAE-ActMahalanobis	*0.779*	*0.818*	*0.763*	*0.897*	***0.998***	*0.798*	***0.799***	*0.982*	***1.000***	***1.000***	*0.997*	***0.999***	***1.000***
PCA-Recon (linear basis)	0.835	0.830	0.737	0.954	0.989	0.842	0.790	0.994	**1.000**	**1.000**	0.996	0.996	0.996
ReAct-Energy	0.959	0.939	0.956	0.880	0.919	0.921	0.577	0.958	0.935	0.913	0.931	0.876	0.890
DINOv2-KNN (reference)	0.991	0.990	0.987	0.966	0.990	0.953	0.793	0.998	1.000	1.000	0.998	0.997	0.994

Finding 1 — the sparse SAE is competitive on natural-image semantic shift, not better. Against the published baselines, SAE scorers are best or tied-best on six of thirteen columns — Places365, STL-10 (a near-OOD set that shares 9 of CIFAR-10's 10 classes), ImageNet-R, ImageNet-Sketch (margins 0.002–0.008), and ceiling ties on DTD/Food-101 — while the published ViM is best on ImageNet-O (0.998 vs 0.997), TinyImageNet, MNIST, SVHN, LSUN-resize, and CIFAR-100; SAE-Recon trails the best dense baseline by 0.152 on SVHN and 0.112 on CIFAR-100. PCA-Recon matches SAE-Recon within 0.003 on ImageNet-O/R/Sketch and Places365: the signal comes from reconstruction itself. ReAct-Energy is markedly weaker on every semantic-shift set (ImageNet-O 0.931 vs 0.997). The strict-sparse SAE of Table 1 is a seed-42 point estimate (SAE-Recon seed-std ≤ 0.009); the 20-seed significance grid of Appendix C.3 uses a denser comparison SAE.

Finding 2 — CIFAR-10-C: no SAE advantage. Across the 19 CIFAR-10-C corruptions (Hendrycks and Dietterich, 2019) (severity 5) under the same protocol as Table 1, SAE-Recon (0.660 mean AUROC) trails the published ViM (0.746) and probe-based MSP (0.726), is best on only 2 of 19 corruptions, and leads no corruption group (Table 2, Appendix A.2).

Finding 3 — the probe removes the leakage confound. Under the ImageNet head, STL-10, Food-101, and DTD are near-distribution; with the CIFAR-10 probe they are genuine OOD (MSP 0.571/0.911/0.916). Our main semantic-shift claims rest on ImageNet-O.

5 The Confidence Collapse Is Reconstruction — and It Is Head-Dependent

Protocol. For each sample we encode z, optionally zero a unit subset, decode $\hat{z} = W_{\text{dec}}a' + b_{\text{dec}}$, pass $\hat{z}$ through the frozen head, and report the relative drop in mean MSP against the unperturbed embedding. The control battery separates the lossy round-trip itself (no units ablated) from the ablation of specific units; the "random active units" control must select genuinely active units (a common masking scheme selects inactive ones; Appendix B.1). Fixed unit sets are chosen on one half of ImageNet-O and evaluated on the other (held-out, n=1,000); full-set numbers use all 2,000 samples. ImageNet-O was built adversarially against ImageNet classifiers, so we also report SVHN and CIFAR-100.

Finding 4 — the collapse is the round-trip, not unit identity, and not generic noise. With no units ablated, the round-trip drops OOD MSP by 70.3% (full) / 73.3% (held-out) versus 10.7% on ID; the OOD drop is 3.9× (SVHN), 4.3× (CIFAR-100), and 6.6× (ImageNet-O) the ID drop (SVHN and CIFAR-100 thus fall below the audit's 5× criterion). At k = 5, every ablation protocol that leaves ID intact reproduces the ImageNet-O round-trip within ~1 point (SVHN: +2.8; more at larger k;

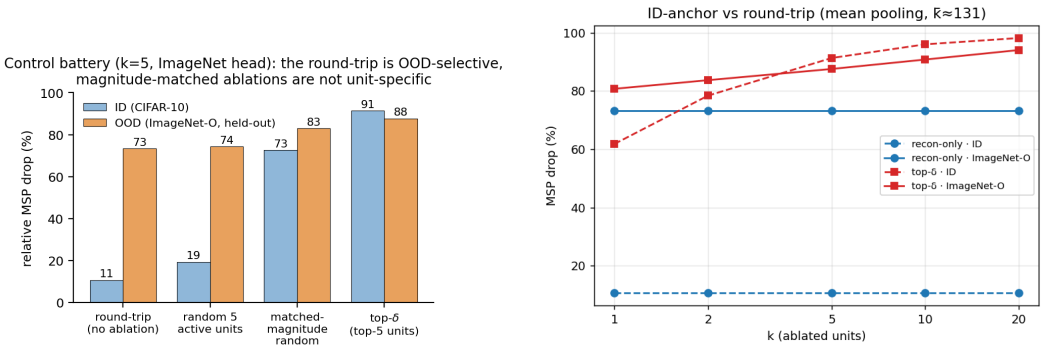


Figure 1: Control battery (ImageNet head, mean pooling). **Left** (k=5): the round-trip alone is OOD-selective; magnitude-matched random ablation reproduces most of the top-δ effect. **Right**: MSP drop vs number of ablated units k; the round-trip is flat in k while top-δ ID damage saturates near 100%.

Appendix E.2), and Gaussian noise matched to the reconstruction error drops ImageNet-O MSP by only 7.2% (per-sample difference +0.61 [0.58, 0.64]; Appendix C.3); per-sample logit-matched, residual-covariance, and swapped-residual noise give at most 10.5% (Appendix E.2). The drop is direction-driven (norm-preserving reconstruction: 73.8%), and it is not a base-confidence artifact: at matched ID-fitted base-MSP quintiles the OOD drop exceeds the ID drop by 3.7–11$\times$ in four of five bins, though not in the lowest-confidence bin (12.2% vs 14.8%; Appendix E.2). Generic low-rank or random maps behave differently: a random orthonormal projection and an untrained SAE destroy confidence everywhere (96.6%/99.6% on ImageNet-O, 92.1%/99.5% on ID), so the selectivity requires a *structured, ID-fit* reconstruction (Table 4b). Nor is the effect a resolution artifact: SVHN and CIFAR-100 are 32$\times$32 images like the CIFAR-10 ID and still drop 3.9–4.3$\times$ more than ID, and with ImageNet-100 as ID, at ImageNet-O's resolution, broad heads still show it (§5.1). A single-example walkthrough is in Appendix C.5.

Finding 5 — the largest-contribution units are ID anchors. The only protocol that adds to the round-trip is per-sample top-δ ablation, and a magnitude-matched random control reproduces most of it (magnitude-matched vs top-δ: ID 72.6% vs 91.3%; OOD 83.0% vs 87.6%): the effect tracks removed contribution *mass*, not unit identity. Removing the single largest-contribution unit drops ID MSP by 61.8% (k=1) while adding only +7.5 points on ImageNet-O (Fig. 1). The same structure holds in the TopK family and with CLS pooling, where top-δ ablation removes 91.4% of ID confidence, magnitude-matched random ablation 73.9%, and fixed or random unit sets add at most 1.1 points over the round-trip (Appendix C.4, E.2). It also holds on DeiT-B/16 (top-δ removes 95.5% of ID confidence, magnitude-matched random 93.6%) and with ImageNet-100 as ID (top-δ 80.4% of ID vs 31.7% of OOD confidence).

Finding 6 — the diagnostic is not SAE-specific. Scoring by the MSP drop gives AUROC 0.865 on held-out ImageNet-O, but a k=256 PCA round-trip gives a comparable drop (73.0%) and a stronger drop-AUROC (0.950 vs 0.859), and plain SAE-Recon (0.997) is a better detector than either. The mechanism is a property of the reconstruction pipeline; what is specific to the sparse code is the descriptive layer of §6.

5.1 Mismatched Wide-Head Reconstruction Sensitivity

Finding 7 — at a fixed ID, head width drives the collapse; heads matched to the ID are inert. Comparing heads across ID datasets confounds head width with the ID. To separate the two, we fix the ID (CIFAR-10), the OOD set, the SAE, and the head weights, and restrict the ImageNet head's softmax to random subsets of W classes (ten draws per W). The ImageNet-O drop rises monotonically with W ($-$3.0%, 14.3%, 35.1%, 57.0%, 70.3% at W = 10, 30, 100, 300, 1,000), while the ID drop stays at 8.8–13.9% (Fig. 2). The dependence is not specific to ImageNet-O: on all 13 OOD sets of Table 1 the W = 1,000 drop (25.5–77.4%) exceeds matched noise by 25.7–73.3 points and the drop rises monotonically with W on 11 of them, although only 5 reach the audit's 5$\times$ OOD/ID ratio

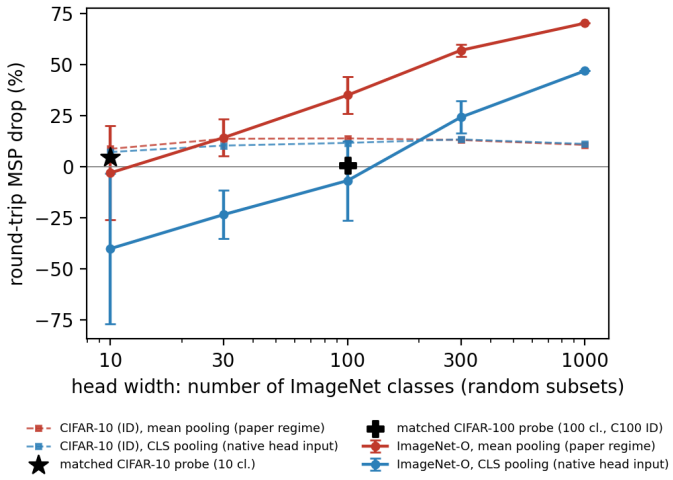


Figure 2: Round-trip MSP drop vs head width at fixed ID (CIFAR-10), fixed SAE, and fixed head weights: the ImageNet head's softmax is restricted to W random ImageNet classes (mean $\pm$ s.d. over ten draws). The OOD drop grows with W while the ID drop stays flat, under both mean and CLS pooling; heads trained on the ID label space (black markers) are inert. Source: `outputs/results/head_width_sweep.json`.

(Appendix E.2). Matched noise also moves wider heads more (ImageNet-O noise drop -15.4% at W = 10 to 7.2% at 1,000), so we also report the round-trip's excess over noise: it grows from 12 to 63 points with mean pooling and from -3 to 53 on DeiT-B/16 (CLS); with CLS pooling on the ViT it appears only at W $\geq$ 300 (18 and 33 points); and for trained heads it levels off beyond W = 100 (17, 27, 34, 37, 34 points). Width alone is not sufficient: heads trained on the ID label space are inert at similar widths (Table 3), and the full head on its native ImageNet-1K ID drops only 6.4% vs 3.2% (§7). The sensitivity therefore arises mainly with a wide head whose label space does not match the audited ID (one exception, DeiT with mean pooling, is in Appendix E.2). The same trend holds when the head receives its native CLS input (-40% at W = 10 to 47.0% vs 11.2% ID at W = 1,000), so it is not an artifact of mean pooling, though pooling changes its magnitude.

Nor is it an artifact of restricting one pretrained head: linear heads *trained* on mean-pooled ImageNet embeddings of W random classes, so that head and features share a pooling, show the same monotonic rise (ImageNet-O 3.0%, 12.6%, 28.8%, 38.0%, 40.3% at W = 10, 30, 100, 300, 1,000; ID 6.0–9.6%). Finally, in a realistic narrow-in-domain setting, with ImageNet-100 as ID and its own SAE, a head trained on the 100 ID classes is inert (1.0% OOD vs 0.2% ID), while heads covering the ID classes plus 200 or 900 others drop ImageNet-O confidence by 10.3% and 10.4% at essentially no ID cost (0.0%, -0.4%); on a second random class draw with three SAE seeds the broad head gives 9.0–10.3% and the matched head 0.4–2.2% (Appendix E.2). The trend replicates on a second supervised backbone, DeiT-B/16 with its native CLS input, though with a larger ID drop and so weaker selectivity (Table 3). It is also not a logit-scale artifact: on ImageNet-O it coincides with a margin collapse (top-1 minus top-2 logit 2.14 $\rightarrow$ 0.99) while the logit norm does not shrink, and fixed temperatures keep the OOD drop at 4.8–6.6$\times$ the ID drop. Width also does more than amplify the softmax: the relative OOD margin collapse itself grows from 0.7% at W = 10 to 53.9% at 1,000, while the ID margin change falls from 10.6% to -0.3% (Appendix E.2).

Table 3. Round-trip MSP drop on ImageNet-O as the head widens, at a fixed ID and fixed SAE. "Narrowest" is W = 10 random classes (for ImageNet-100, the matched 100-class head); "widest" is W = 1,000. Heads trained on the ID label space are inert: 10-class CIFAR-10 probe 4.5% (ID 0.5%), 100-class CIFAR-100 probe 0.7% (ID 4.5%). Full sweeps: Fig. 2 and Appendix E.2.

Backbone, head, pooling	ID	IN-O drop, narrowest	IN-O drop, widest	ID drop (all W)
ViT-B/16, restricted ImageNet head, mean pool	CIFAR-10	−3.0%	70.3%	8.8–13.9%
ViT-B/16, restricted ImageNet head, CLS pool	CIFAR-10	−40.0%	47.0%	7.3–13.4%
ViT-B/16, heads trained at each width, mean pool	CIFAR-10	3.0%	40.3%	6.0–9.6%
DeiT-B/16, restricted ImageNet head, CLS pool	CIFAR-10	−56.5%	71.3%	10.7–23.7%
ViT-B/16, trained heads (ID classes → ID + 900), mean pool	ImageNet-100	1.0%	10.4%	−0.4 to 0.2%

This models a common deployment pattern: a broad pretrained classifier applied to a narrower target domain. There, confidence drops under reconstruction are real and OOD-selective, but they are a property of the head–code system and give no evidence that the SAE contains OOD units, and they warn against reading a softmax collapse as intrinsic to the encoder. For auditing, the diagnostic is informative when a wide pretrained head meets a new, narrower ID and weak when the head is at home; on a self-supervised backbone with a probe (DINOv2), the round-trip even *raises* ImageNet-O confidence (−16.3%), so the direction of the change is itself head-specific (Appendix A.6). The aggregate structure is stable across three SAE seeds (round-trip ID / held-out ImageNet-O drop 9.7% [9.2, 10.7] / 73.9% [73.3, 74.4], range over seeds; Appendix E.2).

A third-party SAE. We also ran the audit on a released SAE we did not train: the Prisma (Joseph et al., 2025) SAE (layer-11 residual stream, 49,152 latents, trained by its authors on ImageNet-1k activations) for a CLIP ViT-B/32 (Radford et al., 2021) trained on DataComp (Gadre et al., 2023), spliced into the OpenCLIP (Ilharco et al., 2021) forward pass (with OpenCLIP's own preprocessing) with zero-shot text heads (clean zero-shot accuracy 95.8% on CIFAR-10 and 67.7% on ImageNet-100 over 1,000 classes). This general-purpose SAE reconstructs ID and OOD alike (fraction of variance explained 0.80 on CIFAR-10 vs 0.78 on ImageNet-O); its reconstruction error does not separate CIFAR-10 from ImageNet-O (AUROC 0.49), and the round-trip is not OOD-selective under any head (1,000-class head: 9.5% ID vs 5.6% OOD, below matched noise at 12.0%) and shows no width dependence. The largest-contribution units still gate ID confidence more than OOD confidence (47.9% vs 30.1%). Because this setting changes the backbone, the SAE's position, and the head type at once, a controlled comparison varies only the SAE's training data: at the same hook, with the same zero-shot heads, we trained SAEs with our recipe on equal numbers of CIFAR-10, ImageNet, or Food-101 activations (three seeds each). Only the CIFAR-10-fit SAE reproduces the effect: 60.0–63.6% OOD vs 5.2–10.1% ID under the 1,000-class head (matched noise $\leq$ 18.8%), with no collapse under the matched 10-class head. The ImageNet-fit SAE, fit on the OOD set's source domain, barely separates ID from OOD (reconstruction-error AUROC 0.46–0.50; 22.8–27.7% ID vs 30.5–31.1% OOD), and the Food-101-fit SAE, fit on neither and as sparse as the CIFAR-10 one, collapses ID more than OOD (51.9–59.0% vs 41.0–44.2%). In this setting, the sensitivity therefore requires an SAE fit on the audited ID, not merely one not fit on the OOD domain (Appendix E.2).

6 Attribution Spectrum: Diffuse Endpoints, Unstable Units

Finding 8 — semantic shift is diffuse, extreme domain shift is localized. Attribution entropy orders the sets consistently across metrics (Fig. 3): ID is the most concentrated (H_{norm} 0.636) and natural-image semantic shift the most diffuse (ImageNet-R 0.771, ImageNet-O 0.764; ID-vs-ImageNet-O Δ = +0.128 [0.123, 0.132], d = 1.38), with effective units rising from 27 to 90. Across samples, extreme domain shift concentrates on a single modal unit (MNIST 87.4%, ImageNet-Sketch 29.4%, SVHN 15.1% of samples) far more than semantic shift (8.5–12.8%) or ID (4.5%) (Appendix A.9).

Finding 9 — the spectrum is sharper than dense or PCA bases, but largely re-encodes reconstruction error. Dense decision attribution is flat (0.878–0.888) and a k=256 PCA attribution is uniformly more diffuse (0.84–0.93), so the dictionary concentrates attribution. But H_{norm} correlates with reconstruction error (+0.43 to +0.67) and active-unit count, and its residual OOD signal after controlling for reconstruction error is small (partial ρ = 0.18, residual AUROC 0.63). Its value as a detection signal is not established. The spectrum does not transfer to the TopK family (Appendix C.4).

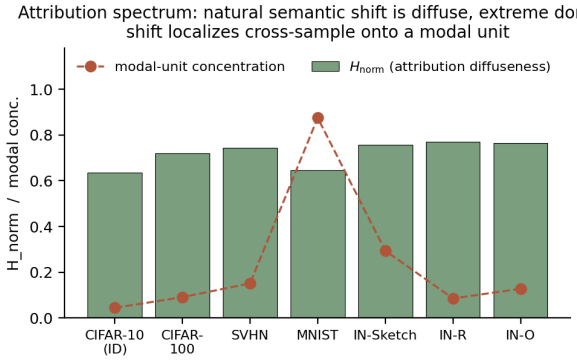


Figure 3: Attribution spectrum (ReLU SAE, seed 42): natural-image semantic shift is the most diffuse, while extreme domain shift (MNIST, SVHN, ImageNet-Sketch) concentrates on a modal unit across samples.

Units are not stable across seeds. Across three SAE seeds (data split fixed), only 8.2–8.8% of units have a >0.7-correlated partner in another seed's dictionary; TopK training raises this only slightly (9.7–10.4%). The dictionary is not a canonical decomposition, as superposition theory predicts (Elhage et al., 2022) and as Paulo and Belrose (2026) and Fel et al. (2025) found for language- and vision-model SAEs. Unlike Paulo and Belrose's setting, where TopK SAEs were the less stable family, here TopK is slightly *more* stable than ReLU (9.7–10.4% vs 8.2–8.8%), and a wider dictionary does not fix this: at fixed training data and sparsity (k/h = 6.4%), the fraction of units with a correlated partner *falls* from 12.4–14.3% at h = 1,024 to 6.4–6.8% at h = 8,192 (Appendix E.3). More training data does not fix it either: the ImageNet-1K TopK SAEs (1.15M rather than 16k embeddings, same h and k) reach 37.3% only through rarely firing units; among units firing on $\geq$1% of samples the rate is 14.0–15.6%, as at CIFAR scale (Appendix E.3). The aggregate endpoints, however, replicate (ID H_{norm} 0.636/0.619/0.624; ImageNet-O 0.764/0.752/0.766), so all our interpretability claims rest on aggregates, never on the identity of individual units. Concept labeling, failure-mode clustering, and unit visualizations are in Appendix E.3.

7 Robustness: Larger ID, Other Architectures, ImageNet-1K Scale

Finding 10 — CIFAR-100 as a 100-class ID (20 seeds). Against the published baselines, SAE scorers are best or tied-best on four of ten sets: Places365 (0.992 vs 0.984), ImageNet-O (0.988, tied with ViM), and the DTD/Food-101 ceiling ties; the published ViM leads TinyImageNet and STL-10, and Lee et al.'s Mahalanobis leads MNIST (Appendix A.3).

Finding 11 — the causal structure is architecture-dependent; ID anchors are not. On ResNet-50 the round-trip is small (15.7% vs 73.3% on held-out ImageNet-O) and magnitude-matched random ablation reproduces most of the top-δ effect; the largest-contribution units are ID anchors in both architectures. On DINOv2, detection transfers but the round-trip collapse is absent under a probe audit (ImageNet-O -16.3%). Attribution entropy also flips: ResNet-50 localizes semantic shift (0.808 vs ID 0.886) where the ViT diffuses it (Appendix A.4, A.6).

Finding 12 — at ImageNet-1K scale, the SAE trails the best dense baseline. With a native 1,000-class head and seven OOD sets, a ReLU scale SAE ($\bar{k} \approx 629$, 15.4% of h) is best on none; a sparsity-matched exact-k TopK SAE (k = 262, 6.4%) improves on it but trails the strongest dense baseline on all seven, by 0.006–0.007 on ImageNet-O/R and 0.030–0.051 on the rest (the published ViM is the strongest dense baseline on DTD, ImageNet-O, and ImageNet-R); the deficit holds for each of three TopK seeds (SAE-Recon seed range ≤ 0.013). The head-based baselines here (MSP, Energy, ViM) apply the native head to mean-pooled embeddings it was not trained on (Appendix B.2), so these gaps likely understate the SAE's deficit; even so handicapped, MSP is the strongest method on iNaturalist. With the head on its native ID, the round-trip drop nearly vanishes (ImageNet-O 6.4% vs ID 3.2%), as §5.1 predicts, and it is far smaller than a PCA round-trip (53.9%), so the MSP-drop diagnostic is weak at scale (drop-AUROC 0.47–0.56). Attribution partly transfers: the ID–OOD H_{norm} gap

narrows, but extreme domain shift still localizes (ImageNet-Sketch modal concentration 0.454 vs ID 0.029) (Appendix A.5).

8 Discussion and Limitations

What sparsity buys. For raw OOD AUROC, a PCA basis often suffices and dense baselines are often stronger. For unit-level OOD localization, the answer is no in both SAE families. What sparse codes do add is a control-based decomposition of confidence collapse (round-trip versus ablation mass; ID anchors) and a sharper aggregate attribution description whose detection value is not established (Appendix C.7). An idealized variance proposition is in Appendix E.5. The round-trip mechanism and the unit instability hold across SAE families; the round-trip magnitude and the attribution spectrum depend on architecture and head.

Practitioner guide. Use sparse SAEs when an interpretable attribution *description*, or an audit of *why* confidence changes, is the goal; do not expect them to beat KNN, Mahalanobis, or ViM on raw AUROC. Test PCA reconstruction before concluding that sparse features are needed, and run the head-dependence check of §5.1 before interpreting confidence drops when a wide pretrained head is deployed on a narrow ID. Audit verdicts characterize mechanism, carry their threshold margins, and do not transfer across encoder families (Appendix A.6).

Limitations. (a) The ViT-B/16 is primary; DeiT-B/16, ResNet-50, and DINOv2 are checks, and on DeiT the width effect is less OOD-selective. (b) Head-width results use ImageNet-derived heads on two IDs (CIFAR-10, ImageNet-100), and the ImageNet-100 SAE is denser (13.8% of h) than the CIFAR SAE. (c) Dense-baseline hyperparameters were not swept (e.g. KNN k is fixed at 50), and the 20-seed significance grid of Appendix C.3 uses a denser comparison SAE than Table 1. (d) Audit thresholds are not calibrated for unseen systems, and the audit has been run on this paper's own SAEs and one third-party SAE; auditing the SAEs of Olson et al. (2025) and Oh et al. (2026) remains future work. (e) Small fixed unit sets do *detect* OOD within one model (five units: held-out ImageNet-O AUROC 0.90–0.94), but below SAE-Recon and active-unit count, with weak transfer (CIFAR-100 0.61–0.66) and seed-specific units (Appendix E.3); unit ablations were not run at ImageNet scale. Extended limitations and a practitioner guide are in Appendix E.6.

9 Conclusion

Sparse features give a useful aggregate attribution *description* for ViT OOD analysis, but not a stable vocabulary of localized OOD units, in either SAE family we test. The confidence collapse under sparse-code perturbation is the OOD-selective, direction-driven reconstruction round-trip, far stronger than matched noise, and it appears when a wide head whose label space does not match the ID audits the code of an SAE fit on that ID: at a fixed ID it grows with head width, and matched heads are inert. The largest-contribution units are ID anchors, and spectrum endpoints replicate across seeds even where units do not. The reconstruction signal is not SAE-specific; what sparse codes buy is the control-based explanation and the attribution description, released as the six-step Sparse Reconstruction Audit with explicit, margin-reported decision criteria.

AI Use Statement

Generative AI tools were used for software engineering (implementing and debugging the experiment scripts), for suggestions on experimental design, including control experiments that the authors implemented, ran, and verified, for literature search, and for drafting and editing substantial parts of the text. All reported results are produced by the released scripts from the released artifacts, and all AI-assisted code and references were verified by the authors. The authors take full responsibility for the content of this paper.

Reproducibility Statement

All results are produced from cached frozen-encoder embeddings by released scripts:
`scripts/run_all.py` (20-seed detection protocol), `scripts/score_*` (single-model rows),
`scripts/sparse_reconstruction_audit.py` (six-step audit, with `--cache-dir demo_data`

demo mode), `scripts/check_numbers.py` (prints the canonical artifact values for every key result and fails if an artifact is missing or malformed; manuscript numbers were additionally verified against these artifacts by a scripted cell-level diff with stated rounding tolerances). Porting the audit to a new encoder requires an embedding cache in the release format, an ID-trained SAE (L0-bisection recipe in the README), and a head-logits function; released compute envelopes are in the README. The supplement includes all code, the aggregate per-step JSONs (`outputs/results/`, `outputs/interpretability/`, `outputs/theory/`), and a demo bundle (an embedding subset and the paper's sparse SAE); full embedding caches, checkpoints, and per-sample dumps exceed the size limit and are regenerated by `scripts/regenerate_current_checkpoint.sh`. The head-width experiments use `scripts/head_width_sweep.py`, `scripts/trained_head_width.py`, and `scripts/second_backbone_deit.py`; the published baselines use `scripts/faithful_baselines.py`; the dictionary-size sweep uses `scripts/dictionary_size_sweep.py`; the preprocessing sensitivity uses `scripts/normalization_sensitivity.py`; the third-party SAE audit uses `scripts/external_sae_audit.py` and its controlled comparison `scripts/controlled_clip_sae.py`; the ImageNet-1K TopK seed summary uses `scripts/summarize_topk_scale_seeds.py` and its unit stability `scripts/scale_unit_stability.py`; correlation-matched top-10 overlap uses `scripts/matched_topk_overlap.py`; firing-stratified stability uses `scripts/unit_stability_by_firing.py`; fixed-unit-set detection uses `scripts/fixed_unit_detection.py`; the margin and subspace checks use `scripts/head_subspace_margin.py`; the all-set width sweep uses `scripts/width_sweep_all_ood.py`; the stronger noise nulls use `scripts/stronger_noise_nulls.py`; the three-arm CLIP comparison uses `scripts/controlled_clip_sae_seeds.py`. The TopK ablation uses `src/training/train_sae_mean_topk.py` (seeded) and `scripts/score_topk_sae_detection.py`; the TopK unit-level experiments use `src/analysis/unit_stability.py` and `src/evaluation/intervention_controls.py` with the `--sae-family topk` flag. Every published checkpoint is produced by `src/models/sae.py` (the single SAE implementation behind the ReLU, TopK, ResNet, DINOv2, and scale variants) via the `src/training/train_sae_.py` scripts; `src/sae.py` in the release root is a legacy duplicate retained for provenance of the earliest CLS-pooled baseline only and is not used by any canonical-path result.

Ethics Statement

This study uses only public computer-vision benchmark datasets (CIFAR-10/100, SVHN, STL-10, MNIST, LSUN, TinyImageNet, Places365, DTD, Food-101, SUN397, iNaturalist, ImageNet and its standard OOD derivatives). No human subjects, personal data, or novel data collection are involved. Improved OOD detection and mechanistic auditing of deployed classifiers support safer deployment; we note the dual-use consideration that any detector can also be evaded, and that our results *bound* what sparse-feature OOD scoring can justify (reconstruction, not localized anomaly units). A deployment-specific risk of this release is *false assurance*: audit verdicts characterize mechanism and are not safety certifications, and a verdict on one encoder family does not clear other families (the ViT/ResNet/DINOv2 contrasts in Appendix A.6 are the paper's own demonstration of this).

References

L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – Mining Discriminative Components with Random Forests," ECCV, 2014.

T. Bricken et al., "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning," Transformer Circuits Thread, 2023.

H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," CVPR, 2021. arXiv:2012.09838.

A. Chhetri, P. Shrestha, R. Rana, S. Philip, P. Gyawali, and B. Bhattarai, "When Confidence Lacks Concepts: Interpretable OOD Detection via Representation Perturbations," arXiv:2606.16196, 2026.

M. Cimpoi et al., "Describing Textures in the Wild," CVPR, 2014. (DTD.)

A. Coates, A. Ng, and H. Lee, "An Analysis of Single-Layer Networks in Unsupervised Feature Learning," AISTATS, 2011. (STL-10.)

H. Cunningham et al., "Sparse Autoencoders Find Highly Interpretable Features in Language Models," ICLR, 2024.

J. Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," CVPR, 2009.

A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.

N. Elhage et al., "Toy Models of Superposition," arXiv:2209.10652, 2022.

T. Fel, E. S. Lubana, J. S. Prince, M. Kowal, V. Boutin, I. Papadimitriou, B. Wang, M. Wattenberg, D. Ba, and T. Konkle, "Archetypal SAE: Adaptive and Stable Dictionary Learning for Concept Extraction in Large Vision Models," ICML, 2025. arXiv:2502.12892.

S. Y. Gadre et al., "DataComp: In Search of the Next Generation of Multimodal Datasets," NeurIPS Datasets and Benchmarks Track, 2023.

L. Gao et al., "Scaling and Evaluating Sparse Autoencoders," arXiv:2406.04093, 2024.

X. Guan, Z. Liu, W.-S. Zheng, Y. Zhou, and R. Wang, "Revisit PCA-based Technique for Out-of-Distribution Detection," ICCV, 2023.

K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," CVPR, 2016.

D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," ICLR, 2019. (CIFAR-10-C.)

D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," ICLR, 2017.

D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural Adversarial Examples," ICLR, 2021a. (ImageNet-O.)

D. Hendrycks et al., "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization," ICCV, 2021b. (ImageNet-R.)

D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, "Unsolved Problems in ML Safety," arXiv:2109.13916, 2021c.

V. Huang, D. Choi, D. D. Johnson, S. Schwettmann, and J. Steinhardt, "Predictive Concept Decoders: Training Scalable End-to-End Interpretability Assistants," arXiv:2512.15712, 2025.

G. Ilharco, M. Wortsman, R. Wightman, et al., "OpenCLIP," Zenodo, 2021. doi:10.5281/zenodo.5143773.

X. Jiang, F. Liu, Z. Fang, H. Chen, T. Liu, F. Zheng, and B. Han, "Negative Label Guided OOD Detection with Pretrained Vision-Language Models," ICLR, 2024.

S. Joseph, P. Suresh, L. Hufe, E. Stevinson, R. Graham, Y. Vadi, D. Bzdok, S. Lapuschkin, L. Sharkey, and B. A. Richards, "Prisma: An Open Source Toolkit for Mechanistic Interpretability in Vision and Video," CVPR Workshop on Mechanistic Interpretability for Vision, 2025. arXiv:2504.19475.

A. Karmacharya, L. Luschwitz, L. Romero, Y. Niu, and J. Campbell, "Sparse Autoencoders for Interpretable Out-of-Distribution Detection," arXiv:2607.12094, 2026.

A. Karvonen et al., "Activation Oracles: Training and Evaluating LLMs as General-Purpose Activation Explainers," arXiv:2512.15674, 2025.

A. Korznikov, A. Galichin, A. Dontsov, O. Rogov, I. Oseledets, and E. Tutubalina, "Sanity Checks for Sparse Autoencoders: Do SAEs Beat Random Baselines?," arXiv:2602.14111, 2026.

A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," Technical Report, University of Toronto, 2009. (CIFAR-10/100.)

Y. Le and X. Yang, "Tiny ImageNet Visual Recognition Challenge," CS231N course report, Stanford University, 2015.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, 1998. (MNIST.)

O. Ledoit and M. Wolf, "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices," J. Multivariate Analysis, 2004.

K. Lee, K. Lee, H. Lee, and J. Shin, "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," NeurIPS, 2018.

B. Li, Z. C. Guo, V. Huang, J. Steinhardt, and J. Andreas, "Training Language Models to Explain Their Own Computations," arXiv:2511.08579, 2025.

W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based Out-of-Distribution Detection," ICLR, 2021.

L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical Density Based Clustering," JOSS, 2017.

L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," arXiv:1802.03426, 2018.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading Digits in Natural Images with Unsupervised Feature Learning," NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011. (SVHN.)

A. Oh, W. Shin, and S. Kim, "Sparsity as a Key: Unlocking New Insights from Latent Structures for Out-of-Distribution Detection," CVPR, 2026. arXiv:2604.26409.

M. L. Olson, M. Hinck, N. Ratzlaff, C. Li, P. Howard, V. Lal, and S.-Y. Tseng, "Probing the Representational Power of Sparse Autoencoders in Vision Models," ICCV 2025 (Findings). arXiv:2508.11277.

M. Oquab et al., "DINOv2: Learning Robust Visual Features without Supervision," TMLR, 2024.

G. Paulo and N. Belrose, "Sparse Autoencoders Trained on the Same Data Learn Different Features," ICLR, 2026. arXiv:2501.16615.

A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.

Y. Sun, C. Guo, and Y. Li, "ReAct: Out-of-Distribution Detection With Rectified Activations," NeurIPS, 2021.

Y. Sun, Y. Guo, S. Vishwanathan, and Y. Li, "Out-of-Distribution Detection with Deep Nearest Neighbors," ICCV, 2022.

A. Templeton et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet," Transformer Circuits Thread, 2024.

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training Data-Efficient Image Transformers & Distillation through Attention," ICML, 2021.

G. Van Horn et al., "The iNaturalist Species Classification and Detection Dataset," CVPR, 2018. (Experiments use the iNaturalist-2021 mini-edition release.)

H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning Robust Global Representations by Penalizing Local Predictive Power," NeurIPS, 2019. (ImageNet-Sketch.)

H. Wang, Z. Li, L. Feng, and W. Zhang, "ViM: Out-of-Distribution with Virtual Logit Matching," CVPR, 2022.

R. Wightman, "PyTorch Image Models," 2019. (timm; https://github.com/huggingface/pytorch-image-models.)

J. Xiao et al., "SUN Database: Large-scale Scene Recognition from Abbey to Zoo," CVPR, 2010.

J. Yang, P. Wang, D. Zou, Z. Zhou, K. Ding, W. Peng, H. Wang, G. Chen, B. Li, Y. Sun, X. Du, K. Zhou, W. Zhang, D. Hendrycks, Y. Li, and Z. Liu, "OpenOOD: Benchmarking Generalized Out-of-Distribution Detection," NeurIPS Datasets and Benchmarks Track, 2022.

J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized Out-of-Distribution Detection: A Survey," Int. J. Comput. Vis., vol. 132, no. 12, pp. 5635–5662, 2024.

F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop," arXiv:1506.03365, 2015.

B. Zhou et al., "Places: A 10 Million Image Database for Scene Recognition," IEEE TPAMI, 2017.

A Appendix A. Full Tables and Additional Metrics

A.0.1 A.1 FPR95 and AUPR for Table 1 (all methods; lower FPR95 is better)

FPR95 (mean over 20 seeds; SAE rows use the seed-42 model) and AUPR-out for the 13-set suite. Best non-DINOv2 per column in **bold** (FPR95: lower is better; DINOv2-KNN is a stronger-backbone reference and is excluded from bolding). `outputs/aggregated_results.json`, `outputs/results/auroc_sparse_sae_mean.json`, `pca_recon_detection.json`, `react_baseline.json`, `dinov2_knn_results.json`.

FPR95:

Method	SVHN	C100	STL-10	MNIST	DTD	Food-101	LSUN-r	LSUN-c	TImg	Places365	IN-O	IN-R	IN-Sk
MSP	0.237	0.318	0.937	0.223	0.264	0.284	0.249	0.254	0.497	0.377	0.310	0.517	0.433
Energy	0.164	0.324	0.934	0.135	0.270	0.300	0.219	0.164	0.549	0.372	0.303	0.549	0.463
Mahalanobis (total cov.)	0.619	0.577	**0.691**	0.014	0.001	**0.000**	0.604	0.719	0.252	0.043	0.015	0.012	0.015
KNN	0.496	0.525	0.745	0.366	**0.000**	**0.000**	0.617	0.564	0.411	0.015	0.038	0.016	0.044
ViM-style (simplified)	0.149	0.287	0.922	0.061	0.163	0.188	0.197	**0.149**	0.502	0.281	0.194	0.414	0.319
ViM (Wang et al.)	**0.134**	**0.226**	0.770	**0.006**	**0.000**	**0.000**	**0.189**	0.159	**0.206**	0.026	**0.004**	0.017	0.010
Mahalanobis (Lee et al.)	0.599	0.565	0.693	0.014	0.001	**0.000**	0.584	0.697	0.249	0.040	0.015	0.012	0.015
SAE-Recon	0.587	0.571	0.721	0.055	**0.000**	0.001	0.592	0.662	0.317	0.033	0.013	0.010	0.005
SAE-ActMahalanobis	0.592	0.601	0.789	0.033	**0.000**	**0.000**	0.631	0.699	0.434	**0.005**	0.006	**0.004**	**0.002**
PCA-Recon	0.604	0.596	0.701	0.019	**0.000**	**0.000**	0.616	0.716	0.251	0.049	0.015	0.017	0.021
ReAct-Energy	0.160	0.314	0.928	0.142	0.247	0.286	0.207	0.164	0.528	0.338	0.278	0.517	0.430
DINOv2-KNN	0.039	0.229	0.696	0.005	0.000	0.000	0.043	0.054	0.230	0.031	0.003	0.013	0.022

AUPR-out:

Method	SVHN	C100	STL-10	MNIST	DTD	Food-101	LSUN-r	LSUN-c	TImg	Places365	IN-O	IN-R	IN-Sk
MSP	0.913	0.883	0.582	0.882	0.780	0.883	0.904	0.908	0.854	0.886	0.789	0.841	0.859
Energy	0.947	0.905	0.598	0.947	0.810	0.876	0.912	0.941	0.864	0.903	0.818	0.854	0.866
Mahalanobis (total cov.)	0.820	0.845	0.792	0.993	0.999	1.000	0.818	0.700	0.959	0.991	0.992	0.997	0.996
KNN	0.851	0.848	0.790	0.623	1.000	1.000	0.710	0.807	0.939	0.995	0.986	0.996	0.992
ViM-style (simplified)	0.949	0.912	0.607	0.982	0.897	0.934	0.927	0.946	0.888	0.936	0.891	0.903	0.923
ViM (Wang et al.)	0.957	0.942	0.759	0.993	1.000	1.000	0.936	0.914	0.970	0.993	0.996	0.995	0.997
Mahalanobis (Lee et al.)	0.825	0.850	0.791	0.993	1.000	1.000	0.824	0.707	0.960	0.991	0.992	0.997	0.997
SAE-Recon	0.799	0.832	0.773	0.984	1.000	1.000	0.789	0.717	0.948	0.993	0.993	0.998	0.999
SAE-ActMahalanobis	0.753	0.776	0.830	0.944	1.000	1.000	0.788	0.758	0.896	0.999	0.994	0.999	1.000
PCA-Recon	0.834	0.846	0.788	0.989	0.999	1.000	0.820	0.723	0.959	0.990	0.992	0.996	0.996
ReAct-Energy	0.947	0.905	0.601	0.940	0.822	0.877	0.916	0.941	0.868	0.908	0.828	0.861	0.872

(DINOv2-KNN appears in the FPR95 table only: `dinov2_knn_results.json` records AUROC/FPR95 but not AUPR.)

Table 2. CIFAR-10-C (severity 5; mean AUROC), Table-1 protocol. Head-based scorers use the CIFAR-10 probe; SAE rows use the strict sparse SAE (mean over seeds 42/1337/2024); best per row in bold; per-corruption values in Appendix A.2 (`outputs/results/cifar10c_aligned.json`).

Category (#)	MSP	Energy	Maha. (total)	KNN	ViM-style	ViM	Maha. (Lee)	*SAE-Recon*	*SAE-ActMah.*	DINOv2-KNN
Noise (4)	0.791	0.773	0.754	0.727	0.789	**0.837**	0.760	0.771	0.723	0.808
Blur (5)	0.762	**0.770**	0.569	0.655	0.769	0.750	0.575	0.593	0.619	0.761
Weather (5)	0.665	0.662	0.663	0.648	0.670	**0.706**	0.666	0.677	0.651	0.659
Digital (5)	0.699	0.703	0.595	0.658	0.707	0.711	0.599	0.622	0.640	**0.715**
Overall (19)	0.726	0.725	0.639	0.669	0.731	**0.746**	0.644	0.660	0.655	0.732

A.0.2 A.2 CIFAR-10-C per-corruption AUROC

AUROC over 10,000 corrupted test images per corruption (severity 5) under the Table-1 protocol: CIFAR-10 probe for MSP/Energy/ViM, the strict sparse SAE (mean over three seeds; seed std $\leq$ 0.015), deterministic dense baselines, and the DINOv2-KNN reference; best per row in bold. For comparison, scoring MSP/Energy/ViM with the ImageNet-1k head and using a denser 20-seed SAE (`outputs/cifar10c_aggregated.json`), a protocol inconsistent with Table 1, puts SAE-Recon first on the noise and weather groups; that ranking does not hold under the Table-1 protocol.

Corruption	MSP	Energy	Maha. (total)	KNN	ViM-style	ViM	Maha. (Lee)	*SAE-Recon*	*SAE-ActMah.*	DINOv2-KNN
brightness	0.600	0.595	0.573	0.534	0.604	**0.610**	0.575	0.588	0.581	0.556
contrast	0.623	0.639	0.613	**0.752**	0.638	0.686	0.614	0.658	0.709	0.655
defocus_blur	0.752	**0.760**	0.508	0.603	0.756	0.714	0.515	0.541	0.574	0.720
elastic_transform	0.779	0.791	0.583	0.695	0.790	0.769	0.590	0.609	0.634	**0.801**
fog	0.742	0.751	0.720	0.811	0.752	0.793	0.724	0.730	0.722	**0.822**
frost	0.669	0.659	0.665	0.640	0.669	**0.709**	0.668	0.674	0.661	0.652
gaussian_blur	0.772	**0.780**	0.546	0.644	0.779	0.750	0.553	0.580	0.618	0.743
gaussian_noise	0.814	0.796	0.748	0.730	0.810	**0.850**	0.755	0.760	0.740	0.807
glass_blur	0.815	0.829	0.704	0.759	0.833	0.847	0.710	0.712	0.712	**0.907**
impulse_noise	0.786	0.780	0.769	0.711	0.796	**0.850**	0.774	0.782	0.692	0.822
jpeg_compression	0.725	**0.729**	0.445	0.510	0.728	0.654	0.452	0.472	0.524	0.680
motion_blur	0.764	**0.774**	0.552	0.654	0.771	0.750	0.558	0.576	0.605	0.745
pixelate	0.767	0.776	0.657	0.714	0.779	0.787	0.663	0.683	0.671	**0.839**
saturate	0.598	0.581	0.676	0.620	0.599	0.658	0.678	**0.688**	0.659	0.598
shot_noise	0.790	0.767	0.744	0.729	0.783	**0.829**	0.750	0.763	0.731	0.798
snow	0.711	0.707	0.686	0.608	0.722	**0.749**	0.690	0.706	0.677	0.685
spatter	0.604	0.595	0.671	0.647	0.602	0.669	0.672	**0.689**	0.616	0.578
speckle_noise	0.775	0.749	0.754	0.740	0.766	**0.819**	0.760	0.778	0.729	0.806
zoom_blur	0.707	**0.710**	0.533	0.618	0.707	0.688	0.537	0.559	0.587	0.688

A.0.3 A.3 CIFAR-100-as-ID detection (Finding 10)

20-seed protocol with the CIFAR-100 strict sparse SAE ($\bar{k} \approx 151$); 100-class probe for MSP/Energy/ViM. FPR95 in parentheses. Best per column among methods shown in **bold**. `outputs/cifar100_id/aggregated_results.json` (20-seed protocol).

Method	SVHN	STL-10	MNIST	DTD	Food-101	LSUN-r	LSUN-c	TImg	Places365	IN-O
MSP	0.921 (0.304)	0.855 (0.501)	0.736 (0.540)	0.839 (0.510)	0.846 (0.424)	0.796 (0.578)	0.906 (0.368)	0.762 (0.759)	0.785 (0.632)	0.837 (0.548)
Energy	**0.967** (**0.129**)	0.888 (0.400)	0.707 (0.556)	0.842 (0.494)	0.843 (0.456)	0.812 (0.528)	**0.946** (**0.221**)	0.777 (0.735)	0.794 (0.640)	0.845 (0.537)
Mahalanobis (total cov.)	0.620 (0.874)	0.919 (0.307)	0.895 (0.193)	0.998 (0.006)	0.998 (0.006)	0.641 (0.820)	0.522 (0.898)	0.897 (0.496)	0.973 (0.140)	0.978 (0.115)
KNN	0.708 (0.792)	0.957 (0.202)	0.294 (0.808)	0.997 (0.004)	0.999 (0.003)	0.538 (0.855)	0.659 (0.856)	0.856 (0.610)	0.984 (0.079)	0.964 (0.255)
ViM-style (simplified)	0.958 (0.151)	0.899 (0.369)	0.776 (0.469)	0.897 (0.376)	0.902 (0.302)	**0.836** (**0.499**)	0.940 (0.249)	0.800 (0.698)	0.837 (0.571)	0.889 (0.447)
ViM (Wang et al.)	0.909 (0.274)	**0.974** (**0.115**)	0.809 (0.356)	0.998 (0.003)	**1.000** (**0.002**)	0.804 (0.547)	0.866 (0.352)	**0.920** (**0.452**)	0.977 (0.141)	**0.988** (**0.034**)
Mahalanobis (Lee et al.)	0.631 (0.855)	0.919 (0.307)	**0.899** (0.184)	0.998 (0.006)	0.998 (0.007)	0.646 (0.813)	0.529 (0.892)	0.897 (0.497)	0.973 (0.143)	0.978 (0.111)
SAE-Recon	0.623 (0.854)	0.934 (0.360)	0.856 (0.372)	**0.999** (0.004)	**1.000** (0.010)	0.657 (0.803)	0.565 (0.868)	0.900 (0.557)	0.988 (0.111)	**0.988** (0.086)
SAE-ActMahalanobis	0.573 (0.884)	0.881 (0.529)	0.897 (**0.165**)	**0.999** (**0.002**)	0.999 (0.006)	0.634 (0.830)	0.520 (0.931)	0.854 (0.700)	**0.992** (**0.024**)	0.984 (0.100)

Note: KNN on MNIST (0.294) is below chance — a manifold-density artifact: MNIST forms an extremely compact cluster closest to a single CIFAR-100 class ("train", 86.9% of MNIST points), so its k-th-nearest-neighbor distance is *smaller* than CIFAR-100's own test points (`scripts/diagnose_knn_mnist_cifar100.py`). Distance-based scores are not calibrated when the OOD distribution is a compact manifold inside a diverse ID manifold.

A.0.4 A.4 DINOv2-B/14 detection (§7 transfer check, FPR95 in parentheses)

CIFAR-10 probe head for MSP/Energy/ViM; single sparse SAE ($\bar{k} \approx 44$); dense rows regenerated under the current temperature-calibrated probe protocol (`outputs/results/dinov2_detection.json`). The DINOv2 SAE is not matched to the ViT's $\bar{k} \approx 132$ (DINOv2's well-clustered embeddings reach a much lower reconstruction floor); many columns are near-ceiling, so this is a transfer check, not a head-to-head.

Method	SVHN	C100	STL-10	MNIST	DTD	Food-101	LSUN-r	LSUN-c	TImg	Places365	IN-O	IN-R	IN-Sk
MSP	0.977 (0.077)	0.956 (0.172)	0.684 (0.741)	0.994 (0.023)	0.970 (0.101)	0.978 (0.074)	0.984 (0.061)	0.968 (0.115)	0.926 (0.399)	0.967 (0.142)	0.976 (0.081)	0.913 (0.413)	0.947 (0.289)
Energy	0.993 (0.021)	0.964 (0.172)	0.718 (0.696)	0.998 (0.006)	0.982 (0.063)	0.994 (0.024)	0.994 (0.019)	0.991 (0.041)	0.939 (0.409)	0.981 (0.086)	0.989 (0.035)	0.948 (0.291)	0.964 (0.204)
Mahalanobis (total cov.)	0.994 (0.028)	0.924 (0.415)	0.793 (0.569)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.001)	0.979 (0.103)	0.980 (0.082)	0.951 (0.320)	0.981 (0.078)	0.997 (0.012)	0.990 (0.043)	0.989 (0.056)
KNN	0.991 (0.039)	0.953 (0.229)	0.792 (0.696)	0.998 (0.005)	**1.000** (0.000)	**1.000** (0.000)	0.990 (0.043)	0.987 (0.054)	**0.966** (0.230)	0.990 (0.031)	0.998 (0.003)	**0.997** (0.013)	**0.994** (0.022)
ViM-style (simplified)	0.995 (0.016)	0.968 (0.153)	0.721 (0.697)	**1.000** (0.001)	0.992 (0.028)	0.995 (0.017)	**0.995** (0.017)	0.992 (0.033)	0.943 (0.390)	0.982 (0.078)	0.992 (0.022)	0.953 (0.256)	0.968 (0.181)
ViM (Wang et al.)	**0.998** (0.007)	**0.971** (0.144)	0.829 (0.490)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**0.995** (0.024)	**0.995** (0.018)	0.963 (0.239)	0.991 (0.035)	0.998 (0.002)	0.994 (0.030)	0.993 (0.041)
Mahalanobis (Lee et al.)	0.995 (0.025)	0.928 (0.391)	0.793 (0.569)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.001)	0.982 (0.089)	0.983 (0.070)	0.952 (0.320)	0.983 (0.070)	0.997 (0.010)	0.990 (0.042)	0.990 (0.056)
SAE-Recon	*0.997* (0.013)	0.933 (0.354)	0.818 (0.509)	0.999 (0.002)	**1.000** (0.000)	**1.000** (0.000)	0.980 (0.093)	0.984 (0.069)	0.954 (0.297)	0.991 (0.044)	0.998 (0.005)	0.992 (0.035)	0.989 (0.067)
SAE-ActMahalanobis	0.892 (0.336)	0.888 (0.540)	***0.934*** (0.451)	0.999 (0.002)	**1.000** (0.000)	**1.000** (0.000)	0.959 (0.241)	0.893 (0.435)	***0.966*** (0.211)	***0.999*** (0.001)	***0.999*** (0.001)	**0.997** (0.013)	**0.994** (0.026)

A.0.5 A.5 ImageNet-1K-as-ID: detection, audit, and FPR95

Native 1,000-class ImageNet head; ImageNet-1K scale SAE (h=4096, $\bar{k} \approx 629$); embedding scorers fit on a 200k-subset of ID train; OOD capped at 5,000. The ID *is* ImageNet, so the three ImageNet-derived sets are near-distribution stress tests.

Table 9a. Primary non-ImageNet OOD (AUROC). Best per column in bold; our methods in italics.

Method	Places365	DTD	SUN397	iNat
MSP	0.766	0.762	0.772	**0.691**
Energy	0.722	0.742	0.732	0.641
Mahalanobis (total cov.)	0.774	0.877	0.770	0.669
KNN	0.700	0.867	0.712	0.354
ViM-style (simplified)	**0.777**	0.792	**0.783**	0.675
ViM (Wang et al.)	0.764	**0.909**	0.758	0.690
Mahalanobis (Lee et al.)	0.775	0.880	0.776	0.676
SAE-Recon	*0.698*	*0.829*	*0.683*	*0.608*
SAE-ActMahalanobis	*0.680*	*0.845*	*0.678*	*0.563*

Table 9b. ImageNet-derived stress tests (AUROC).

Method	IN-O	IN-R	IN-Sk
MSP	0.717	0.804	0.838
Energy	0.715	0.803	0.790
Mahalanobis (total cov.)	0.784	0.829	0.833
KNN	0.687	0.748	0.599
ViM-style (simplified)	0.750	0.830	**0.854**
ViM (Wang et al.)	**0.791**	**0.836**	0.814
Mahalanobis (Lee et al.)	0.790	0.834	0.833
SAE-Recon	*0.754*	*0.773*	*0.806*
SAE-ActMahalanobis	*0.694*	*0.756*	*0.722*

FPR95 (`outputs/results/imagenet1k_detection.json`; primary columns first, then stress tests; lower is better; DTD's very low Mahalanobis FPR95 reflects texture-class separability):

Method	Places365	DTD	SUN397	iNat	IN-O	IN-R	IN-Sk
MSP	0.733	0.784	0.764	**0.776**	0.759	0.683	0.633
Energy	0.811	0.882	0.816	0.829	0.881	0.704	0.756
Mahalanobis (total cov.)	0.682	0.470	0.694	0.847	0.625	0.648	0.592
KNN	0.745	0.577	**0.661**	0.966	0.838	0.784	0.836
ViM-style (simplified)	0.729	0.773	0.740	0.807	0.736	0.657	0.621
ViM (Wang et al.)	0.721	**0.379**	0.723	0.879	0.648	**0.636**	0.657
Mahalanobis (Lee et al.)	**0.679**	0.480	0.680	0.848	**0.616**	0.647	**0.586**
SAE-Recon	0.826	0.574	0.826	0.900	0.705	0.721	0.704
SAE-ActMahalanobis	0.855	0.494	0.845	0.955	0.863	0.760	0.827

The scale audit numbers are in Appendix A.5 (`outputs/results/imagenet1k_audit.json`); the PCA column there is the audit's Step-6 control: at CIFAR scale the PCA round-trip matched the SAE (73% vs 71% on ImageNet-O); at scale the SAE round-trip is far smaller (6.4% vs 53.9%) — the scale SAE is a high-fidelity autoencoder on its native domain — and the MSP-drop diagnostic is weak (drop-AUROC 0.47–0.56), consistent with Finding 6.

Sparsity-matched scale validation (TopK, k = 262 = 6.4% of h). Same protocol as Tables 9a–9b (native head; embedding scorers fit on the 200k subset), with the exact-k TopK SAE trained on the same 1.15M-embedding cache (`outputs/results/imagenet1k_detection_topk.json`). Dense rows are unchanged from Tables 9a–9b (deterministic). SAE-Recon beats the denser ReLU scale SAE on all seven sets but trails the best dense baseline (now including the published ViM and Mahalanobis) on all seven: DTD 0.879 vs 0.909, ImageNet-O 0.785 vs 0.791, ImageNet-R 0.829 vs 0.830 (published ViM 0.836). Across three SAE seeds (42, 1337, 2024; `imagenet1k_topk_seed_summary.json`) SAE-Recon varies by at most 0.013 AUROC per set and trails the best dense baseline in every seed, by 0.005–0.007 on ImageNet-O/R and 0.029–0.056 on the other five sets. TopK SAE-ActMahalanobis (not tabulated) is weaker than the ReLU row on every set (0.48–0.69; `imagenet1k_detection_topk.json`):

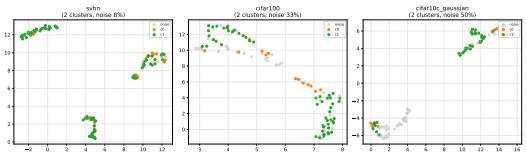


Figure 4: HDBSCAN-clustered 4D features projected with UMAP, colored by cluster label: clusters are driven by reconstruction error, not by distinct OOD mechanisms.

Method	**Places365**	**DTD**	**SUN397**	**iNat**	**IN-O**	**IN-R**	**IN-Sk**
Best dense (from 9a–9b)	0.777	0.909	0.783	0.691	0.791	0.836	0.854
ReLU scale SAE ($\bar{k} \approx 629$, 15.4%)	0.698	0.829	0.683	0.608	0.754	0.773	0.806
TopK scale SAE ($k = 262$, 6.4%)	*0.742*	*0.879*	*0.732*	*0.651*	*0.785*	*0.829*	*0.808*

Its audit (`outputs/results/imagenet1k_audit_topk.json`; kbar exactly 262.0): ID recon drop 1.8% vs ImageNet-O 5.5% [4.8, 6.2] with matched-Gaussian 1.5%; PCA round-trip 53.9%; H_norm ID 0.822 vs OOD 0.832–0.874; ImageNet-Sketch modal concentration 0.218 vs ID 0.059 — the reconstruction-over-noise structure, the PCA excess, and the extreme-shift localization all persist at matched sparsity, and in all three seeds (ImageNet-O round-trip 4.8–5.5% vs matched Gaussian 1.5% and ID 1.8–1.9%; ImageNet-Sketch modal concentration 0.218–0.323 vs ID 0.059–0.160).

A.0.6 A.6 Cross-architecture audit (§5.1)

Appendix Table A.6. Cross-architecture audit. OOD = ImageNet-O (2,000 samples; full-set convention, matching the 70.3% figure; the held-out-1,000 eval half gives 73.3% ViT / 15.7% ResNet, as in §5). The CLS row feeds the head its native CLS input and uses the CLS-trained SAE (`roundtrip_controls_cls_s42_imagenet.json`; the round-trip battery does not compute step 1).

Backbone	**Head**	**SAE $\bar{k}$**	**Recon AUROC**	**MSP-drop AUROC**	**ID drop**	**OOD drop**	**Noise**	**PCA drop**	**Interpretation**
ViT-B/16	ImageNet-1k (1000)	133	0.997	0.859	10.7%	70.3%	7.2%	73.0%	reconstruction-dominant
ViT-B/16, CLS pool	ImageNet-1k (native input)	—	—	0.737	11.2%	47.0%	14.1%	74.3%	smaller round-trip
ViT-B/16	CIFAR-10 probe (10)	133	0.997	0.541	0.5%	4.5%	−16.4%	6.7%	head-dependent (inert)
DINOv2-B/14	CIFAR-10 probe (10)	44	0.998	0.355	0.6%	−16.3%	−16.4%	6.1%	no collapse under probe
ResNet-50	ImageNet-1k (1000)	437	1.000	0.849	2.0%	14.1%	0.3%	4.8%	small round-trip
ViT-B/16	ImageNet-1k head, IN-1K ID	629	0.754	0.561	3.2%	6.4%	0.7%	53.9%	native-head inert (high-fidelity recon)

A.0.7 A.7 Concept-labeling basis (mechanistic explanation)

We label units by CIFAR-10 class purity rather than CLIP zero-shot mapping because the two live in different, misaligned embedding spaces. CLIP maps text and images into a *shared* text-image space, whereas the ViT's final LayerNorm embedding lives in the ViT's own feature space; a CLIP text embedding cannot be inserted into the sparse code without a subspace alignment step, and the CLIP mapping we tested was degenerate (it collapsed to a single token across units; `outputs/concept_labeling.txt`). CIFAR-10 class purity is measured entirely within the ViT's own representation: for each frequent unit, top-activating test images give the dominant-class fraction, with a label-shuffle null (0.116) as chance level. A second labeling protocol (`src/analysis/concept_labeling.py`; 1,591 units) yields higher absolute purity (mean 0.533; 41.4% of units with purity > 0.5), consistent with partial class alignment.

A.0.8 A.8 Additional figures

Figure 4 — failure-mode UMAP (§6; `outputs/figures/fig_appendix_failure_umap.pdf`): OOD manifests as a continuous reconstruction-error continuum rather than discrete failure modes.

Figure 5 — attribution spectrum across datasets (`outputs/figures/fig_attribution_spectrum.png`):

Figure 6 — per-sample attribution-entropy distributions (`outputs/figures/attribution_hnorm_dists.png`):

Figure 7 — unit visualizations (outputs/figures/unit_visualizations.png):

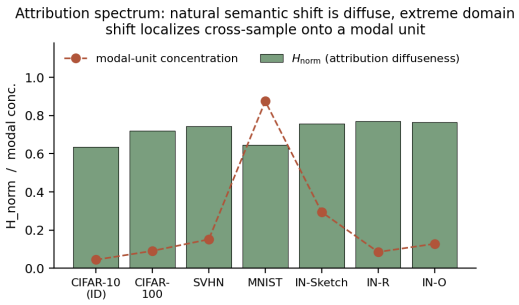


Figure 5: Attribution spectrum across datasets: H_norm rises with semantic shift while modal-unit concentration separates extreme domain shift (SVHN, MNIST) from ID and natural semantic shift.

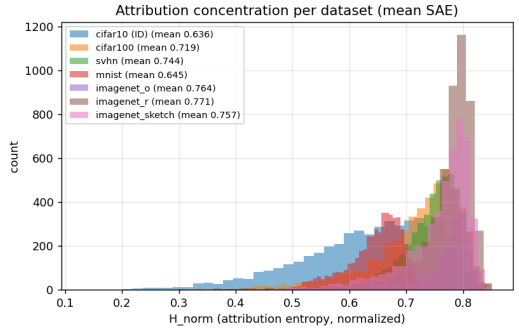


Figure 6: Per-sample H_norm distributions: ID concentrated at low values, ImageNet-O at high values, other sets in between.

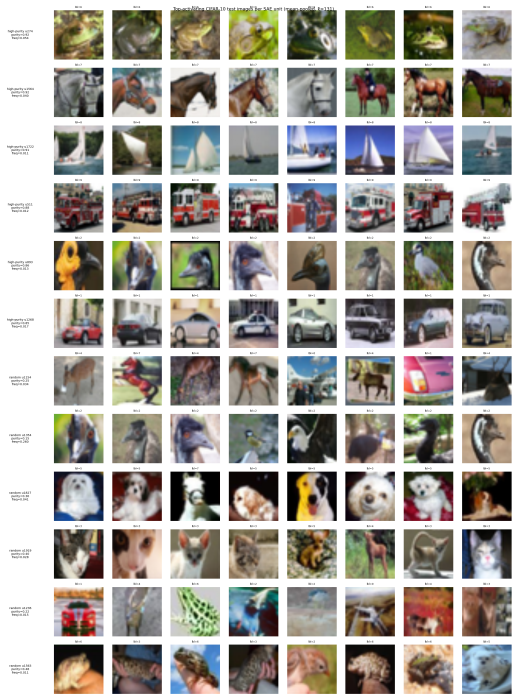


Figure 7: Top-activating CIFAR-10 test images for the highest-purity units (visually coherent) vs. randomly chosen units (polysemantic) — consistent with the moderate-but-real class purity of the sparse code.

A.0.9 A.9 Attribution concentration (§6)

Appendix Table A.9. Attribution concentration (unified mean representation, $\bar{k} \approx 132$). $H_{\text{norm}} = H/\log k$; top-1/top-5 = largest per-sample contribution mass; participation ratio PR and effective units e^H are alternative concentration measures; modal conc. = fraction of samples whose modal top-1 unit is the dataset's modal unit. N = all test samples (H columns: `outputs/interpretability/attribution_baselines_mean.json`; remaining columns: `attribution_robustness_mean_s42.json`; 3-seed replicates in `_s1337`/`_s2024`).

Dataset	SAE H_{norm}	$\bar{k}$	top-1	top-5	PR	eff. units	modal conc.	Dense H_{norm}	PCA H_{norm}
CIFAR-10 (ID)	**0.636**	132	0.250	0.550	14.0	27.1	0.045	0.884	0.913
CIFAR-100	0.719	163	0.173	0.411	22.4	42.6	0.091	0.879	0.921
SVHN	0.744	168	0.135	0.367	26.1	47.5	0.151	0.887	0.896
MNIST	0.645	193	0.211	0.537	13.3	30.7	0.874	0.878	0.841
ImageNet-O	**0.764**	326	0.105	0.267	48.6	89.5	0.128	0.886	0.932
ImageNet-R	0.771	320	0.098	0.254	49.5	90.2	0.085	0.888	0.921
ImageNet-Sketch	0.757	314	0.116	0.282	43.5	83.4	0.294	0.884	0.921

B Appendix B. Audit-Protocol Corrections

Each correction was introduced because a naive implementation either inverted the OOD convention or silently weakened a control; all are reproducible from the released code.

B.1 Random-active masking. The scheme $\texttt{rand} \cdot \texttt{active} + (1 - \texttt{active}) \cdot 10^9$ + top-k silently selects *inactive* (zero-valued) units whenever fewer than k units are active, making the control a no-op that merely approximates the round-trip — and can falsely suggest OOD-specificity. We use masking that selects genuinely active units (`src/evaluation/intervention_controls.py`).

B.2 Head convention. Cached embeddings are post-LayerNorm (timm, (Wightman, 2019): `forward_features` ends with `self.norm`; `fc_norm` is identity for token pooling), so the frozen head is applied as `softmax(head(z))` without re-normalization; re-applying `model.norm` double-normalizes and inflates baseline MSP. The model's native pooling is the CLS token: on CLS-pooled embeddings `head(z)` reproduces the true-forward logits to 5×10^{-6}, but on the mean-pooled embeddings used throughout the paper it does not (max logit error 3.5; top-1 agreement with the true forward pass 80.8%; mean ID MSP 0.381 vs 0.448). The wide-head audit therefore measures a head applied to a representation it was not trained on. With CLS pooling (CLS-trained SAE, native head input) the round-trip still drops ImageNet-O MSP far more than ID MSP (47.0% vs 11.2%; matched noise 14.1%), so the qualitative selectivity survives, but its magnitude is smaller than in the mean-pooled regime (`outputs/results/roundtrip_controls_cls_s42_imagenet.json`).

B.3 ViM sign. The residual/logit combination in ViM must be signed so that a larger residual implies more OOD; the opposite sign inverts the convention and collapses AUROC to ~0.18 (`src/scoring/vim.py`).

B.4 Matched-magnitude random ablation. To separate *mass* from *identity*, the random-ablation control removes random active units, in random order, until the removed contribution mass reaches the top-δ ablation's mass; it removes at least k and at most $10k$ units, so when the cap binds the removed mass is smaller than the target.

B.5 Pooling convention. All caches use CLS+patch mean pooling over the 197 post-LayerNorm tokens; the initial ImageNet-1K caches used patch-only pooling and were re-extracted with all scale results regenerated.

B.6 ID split-cache isolation. Split caches were previously keyed only by seed and fractions and could be silently shared across datasets with the same sample count (CIFAR-100's SAE training had reused CIFAR-10's split indices). Split caches are now keyed by sample count + a label fingerprint and validated on every load (`src/data/splits.py`); all CIFAR-100-as-ID results were regenerated. CIFAR-10 is unaffected (its split recomputes byte-identically).

C Appendix C. Ablations, Stability, and Statistical Support

C.0.1 C.1 Sparsity–AUROC trade-off (denser comparison SAE)

Sweeping the L1 penalty at fixed width (mean-pooled, h=2048) shows milder sparsity buys raw AUROC on low-level shift but blurs the unit-level resolution that §6 requires (`outputs/ablations/*/metrics.json`):

SAE variant	SVHN	CIFAR-100
Strict sparse (L0-bisection, achieved $\bar{k} \approx 132$)	0.819	0.835
Fixed L1 = 5×10^{-4}	0.824	0.845
Fixed L1 = 1×10^{-3}	0.851	0.849
Fixed L1 = 2×10^{-3}	0.858	0.852

C.0.2 C.2 Detection seed stability

Across three SAE training seeds (42/1337/2024; data split fixed), SAE-Recon AUROC seed-std is $\leq$ 0.009 on every set (SVHN 0.817±0.004, Places365 0.992±0.000, ImageNet-O 0.997±0.000, ImageNet-Sketch 0.999±0.000); SAE-ActMahalanobis is noisier on SVHN (0.755±0.020) but stable elsewhere. FPR95 and AUPR aggregates are included (`outputs/results/sparse_detection_seed_stability.json`; the pairing unit for the §C.3 tests is the seed index against the deterministic dense baseline, i.e., a one-sample signed-rank test on per-seed deltas; the FDR family is the full 2-method $\times$ 13-dataset BH grid). The TopK SAE family shows the same detection stability (`outputs/results/topk_sae_detection.json`) and comparable unit-level seed-idiosyncrasy (`outputs/interpretability/unit_stability_topk.json`; Appendix C.4).

C.0.3 C.3 Statistical support

Across the 20-seed protocol, per-dataset paired Wilcoxon tests against the best dense baseline, FDR-corrected over the full method $\times$ dataset grid (`outputs/significance_tests.json`): SAE-Recon is FDR-significantly above the best dense baseline on STL-10 (+0.049), MNIST (+0.0005), DTD (+0.0001), Food-101 (+0.0000), TinyImageNet (+0.006), and Places365 (+0.004). The tests are one-sided (SAE $>$ dense), so they cannot establish equivalence; where dense baselines lead, the SAE is significantly *worse* (paired 95% CIs exclude zero: SVHN -0.110 [-0.113, -0.107], CIFAR-100 -0.073, LSUN-resize -0.094, LSUN-crop -0.163). Effect-size context: the MNIST/DTD/Food-101 "significant" cells sit at ceiling where seed variance is $\approx$0, so statistical significance there reflects the protocol's sensitivity, not practical discriminability — the practical evidence is STL-10 (+0.049), TinyImageNet (+0.006), and Places365 (+0.004). Re-running the grid with the published ViM and Mahalanobis added as (deterministic) competitors (`significance_tests.py --published-baselines`; `outputs/significance_tests_published.json`) leaves SAE-Recon FDR-significantly above the best dense baseline on STL-10 (+0.049), Places365 (+0.004), and the DTD/Food-101 ceiling cells, but not on MNIST (-0.0008) or TinyImageNet (-0.0044), where the published ViM leads. (These paired tests use the 20-seed dense-comparison SAE — the only SAE run across 20 seeds; the strict-sparse SAE of Table 1 is evaluated at seed 42 with 3-seed stability in C.2, and its raw deficits vs the best dense baseline, including the published ViM, are -0.152 SVHN / -0.112 CIFAR-100.) For §5's interventions, per-sample paired tests (`outputs/results/paired_stats.json`; effect sizes and bootstrap CIs are the primary quantities — the p-values, over 1,000–5,000 per-sample pairs, are guaranteed tiny for any nonzero effect and are reported only to bound the null): round-trip vs matched noise +0.61 [0.58, 0.64] (p < 1e-200, d_z = 0.96); top-δ vs magnitude-matched random +0.16 [0.15, 0.17] (d_z = 0.53); the head-dependence contrast has non-overlapping CIs (70.3% [68.7, 71.9] wide head vs 4.5% [2.7, 6.3] probe).

C.0.4 C.4 TopK SAE ablation (modern SAE family, sparsity-matched)

To test whether the paper's findings are a ReLU-SAE artifact, we train a **TopK SAE** ((Gao et al., 2024); exact top-k activations, no L1/λ scheduling) on the same mean-pooled CIFAR-

10 train embeddings with the same protocol (15 epochs, lr 1e-3, batch 256, split seed 42; SAE seeds 42/1337/2024 seeded), with **k = 131 (within rounding of the ReLU SAE's achieved** $\bar{k} \approx 132$; 6.4% of h). Detection (SAE-Recon AUROC, mean±std over 3 seeds; `outputs/results/topk_sae_detection.json`):

Set	**ReLU strict ($\bar{k} \approx 132$)**	**TopK (k = 131)**
SVHN	0.812	0.830 ± 0.001
CIFAR-100	0.835	0.834 ± 0.002
STL-10	0.777	0.758 ± 0.000
MNIST	0.988	0.991 ± 0.001
DTD / Food-101	1.000 / 1.000	1.000 / 1.000
LSUN-r / LSUN-c	0.812 / 0.743	0.831 / 0.747
TinyImageNet	0.942	0.939 ± 0.001
Places365	0.992	0.987 ± 0.000
ImageNet-O	0.997	0.995 ± 0.000
ImageNet-R	0.997	0.996 ± 0.000
ImageNet-Sketch	0.999	0.997 ± 0.000

The full audit under both heads reproduces the reconstruction, noise, and head-dependence findings (`outputs/results/sparse_reconstruction_audit_mean_topk_{imagenet,probe}.json`): reconstruction AUROC 0.995; round-trip ImageNet-O drop **75.9%** vs ID **3.7%**; matched-Gaussian **4.6%**; under the 10-class probe 10.9% vs 0.2% (head-dependent, as in §5.1). The TopK SAE is a purer autoencoder at the same k, and its MSP-drop diagnostic is at least as strong (0.943 vs 0.859) — consistent with Finding 6: the confidence-drop lens tracks reconstruction fidelity of the code, not the SAE family. The attribution spectrum does *not* transfer: TopK attribution is far more diffuse everywhere (ID H_{norm} 0.861, ImageNet-O 0.913), and the ID–ImageNet-O gap (0.051) falls below the audit's 0.10 diffuseness threshold, so Finding 8 is specific to the ReLU SAE.

TopK unit-level claims. We also run the two unit-level measurements for the TopK family (seeds 42/1337/2024, same caches; `outputs/interpretability/unit_stability_topk.json`, `outputs/results/intervention_controls_mean_s42_{imagenet,probe}_topk.json`). *(i) Unit stability:* after mapping each unit to its best-correlated unit in the other seed, TopK per-sample top-10 overlap across seeds is Jaccard 0.17–0.18 (ReLU: 0.23–0.24; raw unit indices carry no cross-seed order and give only the chance level, ~5/h); 9.7–10.4% of units have a >0.7-correlated partner in another seed's dictionary (ReLU: 8.2–8.8%). Exact-k training slightly raises the fraction of functional counterparts but does not produce a stable unit-level vocabulary — the seed-idiosyncrasy conclusion is *not* a ReLU-SAE artifact, and it persists in the modern SAE family. *(ii) ID-anchor ablation (ImageNet head, k=5, held-out ImageNet-O):* the same qualitative structure replicates — reconstruction-only ID 3.7% vs IN-O 80.2%; top-δ ID 65.9% vs 91.8%; magnitude-matched random ID 61.3% vs 90.7%; the largest-contribution units gate ID confidence (ID anchors), although the single largest unit matters less than in the ReLU SAE (k=1 ID drop 27.9% vs 61.8%), and a fixed ID-anchor set adds only +1.0 point over the round-trip (81.2% vs 80.2%). Under the 10-class probe the round-trip is again nearly inert (10.6% OOD vs 0.2% ID). Reconstruction-dominance, noise-specificity, head-dependence, ID-anchor structure, and the seed-idiosyncrasy of unit identity transfer across both SAE families; the diffuse-attribution finding does not.

C.0.5 C.5 Extended intervention tables (full sample sets)

Table 4. MSP confidence drop under the control battery (k=5; held-out ImageNet-O, n=1,000).

Protocol	ID drop	ImageNet-O drop	SVHN drop	CIFAR-100 drop
Reconstruction-only (no ablation)	10.7%	73.3%	41.7%	46.5%
Top-δ (per-sample top-5 by δ_i)	91.3%	87.6%	80.6%	88.1%
Random 5 active units (active-only masking)	19.4%	74.3%	46.3%	50.5%
Matched-magnitude random	72.6%	83.0%	77.6%	78.4%
OOD-discriminative fixed set	10.5%	74.4%	44.5%	47.1%
ID-anchor fixed set	20.8%	74.1%	48.1%	52.4%

Table 4b. Reconstruction vs perturbation controls (full ImageNet-O, n=2,000).

Perturbation	ID drop	ImageNet-O drop	SVHN drop	CIFAR-100 drop
SAE reconstruction-only	10.7%	70.3%	41.7%	46.5%
Gaussian noise matched to recon MSE	1.0%	7.2%	−0.4%	1.9%
Norm-preserving SAE recon (direction only)	−15.2%	73.8%	14.2%	21.6%
PCA round-trip (k = 256)	4.6%	73.0%	35.9%	33.4%
Random orthonormal projection round-trip	92.1%	96.6%	88.9%	93.1%
Untrained (random) SAE round-trip	99.5%	99.6%	99.0%	99.4%

Table 4c. Single-example audit walkthrough (one test sample per set; ImageNet-1k head; `outputs/results/case_study_audit.json`). MSP = max-softmax probability; drops are relative to the example's own base MSP; a negative "drop" is an increase.

Example	Recon MSE/dim	Active units	H_{norm}	Top-1 unit (mass)	MSP base → round-trip	Drop	Matched-Gaussian drop	PCA drop
CIFAR-10 (ID) #0	0.245	131	0.769	u351 (0.16)	0.281 → 0.154	+45.3%	−13.5%	+17.5%
ImageNet-O #0	1.193	242	0.779	u374 (0.06)	0.913 → 0.040	**+95.6%**	+13.3%	+62.7%
SVHN #0	0.728	151	0.749	u1385 (0.10)	0.086 → 0.025	+70.2%	−8.2%	+58.1%

C.0.6 C.6 The two quantities the paper studies (main-text §3 pointer)

Two quantities, deliberately separated. The paper studies two related but distinct objects, and no result conflates them:

Quantity	Definition	Used for	Main finding
SAE-Recon	reconstruction error as an OOD score	detection (§4)	strong on semantic shift, not universally dominant; PCA basis matches it
MSP drop	confidence change after the round-trip	mechanistic probe (§5)	head-dependent, reconstruction-driven; *not proposed as a detector*
Attribution entropy	spread of per-unit contributions	interpretability (§6)	semantic shift diffuse, extreme shift localized
PCA-Recon	linear-basis reconstruction	control	matches much of the raw detection signal

SAE-Recon is detector-facing; the MSP drop is mechanism-facing — a controlled probe of how the sparse code interacts with a downstream head, not a proposed replacement score.

C.0.7 C.7 What sparsity buys, and what it does not buy (main-text §8 pointer)

Appendix Table C.7. What sparsity buys.

Task	PCA sufficient?	SAE useful?	Why
Raw OOD AUROC	often yes (§4)	sometimes	reconstruction/manifold displacement carries the signal
Semantic-shift detection	partly (§4)	competitive, best on some sets	SAE scorers lead Places365, STL-10, IN-R/Sk by 0.002–0.008; the published ViM leads IN-O
Unit-level OOD localization	—	**no**	no stable ablatable OOD units (the main negative result)
Aggregate attribution geometry	no (§6)	**descriptive, not established as detection-relevant**	sparse basis concentrates attribution (Spearman 0), but H_norm correlates with recon error (+0.43 to +0.67) and active count (+0.28 to +0.69); residual OOD signal after partialling (AUROC 0.59–0.68 across seeds) is modest and not shown to add detection value
Explaining confidence collapse	partly (§5)	**yes**	decomposes round-trip vs ablation mass; exposes ID anchors
Deployment-ready detector	—	not necessarily	dense baselines (ViM/Mahalanobis/KNN) often stronger

D Appendix D. Variance-Concentration Proof Sketch and Synthetic Validation

Proof sketch. Write the whitened distance $D = \sum_i (a_i - \mu_i)^2 \sigma_i^{-2}$ with $\Sigma = \mathrm{diag}(\sigma_i^2)$ the shrunk covariance on active coordinates (the diagonal form is the worst case for concentration; full shrinkage only helps). Expanding: $D = \sum_i a_i^2 \sigma_i^{-2} - 2\sum_i a_i \mu_i \sigma_i^{-2} + \sum_i \mu_i^2 \sigma_i^{-2}$. Given the support S (|S| = k), the active coordinates are conditionally independent (the decoder columns are unit-normed so σ_i are bounded), so

$$\mathrm{Var}[D] \;=\; \sum_{i \in S} \mathrm{Var}\big[(a_i - \mu_i)^2 \sigma_i^{-2}\big] \;+\; 4 \sum_{i \in S} \mu_i^2 \sigma_i^{-2}\, \mathrm{Var}[a_i] \;=\; O(k),$$

using bounded fourth moments of the ID-fitted activation distribution; concentration follows because k $\ll$ h with per-term variance O(1). A novel unit j $\notin$ S enters with $\mu_j \approx 0$ and small fitted σ_j, contributing a term with variance O(σ_j^{-4})-scale — far larger than in-support terms and not cancelled by any ID-fitted moment; hence OOD variance does not concentrate. Magnitude drift within S changes only per-term means, not the variance order, so in-support OOD concentrates like ID.

Known gaps in the sketch. The main idealization is the conditional independence of active coordinates, which the trained SAE does not satisfy (activations share the input through the encoder; the sketch treats it as the first-order story). The fixed-support idealization is the weakest link: per-sample k varies and OOD sets activate different supports, so the deployed regime is closer to the novel-unit case than to claim (3). Finally, Ledoit–Wolf shrinkage imposes a floor on fitted σ_j that tempers the O(σ_j^{-4}) argument — the mechanism survives qualitatively (the synthetic validation reproduces it), but the sketch should be read as motivation, not as an analysis of the deployed model.

Synthetic validation (`outputs/theory/synthetic_proposition_validation.json`). Activations from a known k-sparse dictionary model (h = 2048, d = 768, unit-normed columns; ID = fixed support with gamma-distributed magnitudes; OOD = 70% drifted in-support mass + 30% novel-unit mass; Gaussian fitted by Ledoit–Wolf (Ledoit and Wolf, 2004)):

k	k/h	Var[D_ID]	Var[D_OOD]	OOD/ID ratio
16	0.8%	1.03	173,214	168,824
32	1.6%	1.13	48,535	43,054
64	3.1%	1.17	12,271	10,520
128	6.2%	1.22	3,161	2,601
256	12.5%	1.23	786	640

All four predictions hold: Var[D_ID] grows with k; Var[D_OOD] dominates at every k; the ratio stays > 10 across the paper's k/h $\leq 6\%$ regime; and an OOD set restricted to the fitted support concentrates like ID — the variance explosion is caused by *novel units*, not magnitude.

Real-SAE validation (`outputs/theory/concentration_verification.json`, seed=42):

sparsity	ID var	OOD var	ratio (OOD/ID)
0.05	24.87	126,443	5,085
0.10	9.43	92,807	9,839
0.20	3.70	90,115	24,324
0.30	2.49	38,989	15,649
0.50	2.88	37,945	13,196

The sparsity column is the L0 target, a density: larger values mean more active units. ID variance is largest in the sparsest code (24.9 at 0.05) and falls as the code gets denser (2.49 at 0.30; 2.88 at 0.50), the opposite of prediction (1); OOD variance is 10^4–10^5 throughout and exceeds ID variance by 5×10^3–2.4×10^4 (ratio non-monotonic, peaking at 0.20). Only prediction (2) is supported on real SAEs (Appendix E.5).

E Appendix E. Extended Main-Text Material

Additional details for each main-text section; numbers match the main text.

E.1 Setup details (§3)

Encoders. `vit_base_patch16_224` ((Dosovitskiy et al., 2021); timm `augreg2_in21k_ft_in1k` weights, i.e. ImageNet-21k pretraining with ImageNet-1k fine-tuning; frozen), mean pooling over all 197 post-LayerNorm tokens (CLS + patches) $\rightarrow$ 768-d embeddings; ResNet-50 (`resnet50.tv_in1k`, 2048-d) and self-supervised DINOv2-B/14 ((Oquab et al., 2024); no native head) as cross-architecture checks; ImageNet-1K as the scale ID (§7). The native ViT head was trained on the post-LayerNorm CLS token; the intervention audit applies it to the mean-pooled embedding, which is not the input it was trained on (Appendix B.2 quantifies the gap and reports the CLS-pooled control).

Classifier head. Head-based scorers (MSP/Energy/ViM) require an ID label space. For the detection tables we train a **CIFAR-10 linear probe** (multinomial logistic regression on frozen embeddings, temperature-scaled on a held-out 10% validation split; deterministic; ID test accuracy 0.933, `outputs/results/probe_id_accuracy.json`). This removes the ImageNet-head leakage confound that makes ImageNet-derived sets (DTD, Food-101, STL-10) near-distribution for an ImageNet-pretrained classifier (§4, Finding 3). The intervention audit (§5) uses the frozen **ImageNet-1k head**, whose confidence we perturb.

SAE (strictly ID-only). A single ReLU SAE (768 $\rightarrow$ h=2048) trained only on CIFAR-10 train embeddings (no leakage; the 80% train split of a 20,000-image subset), with an L0-bisection sparsity scheduler targeting 32 active units whose λ ceiling (5.0) stops it short of that target, at achieved $\bar{k} \approx 132$ (6.4% of h). Strict sparsity is what makes unit-level attribution interpretable at all (§6); a denser L1 variant trades AUROC for resolution (Appendix C.1), and a modern **TopK SAE** (Gao et al., 2024) at k=131 (matched to $\bar{k} \approx 132$) reproduces the reconstruction, noise, head-dependence, ID-anchor, and unit-instability findings but not the diffuse-attribution finding (Appendix C.4). ResNet-

50 ($\bar{k} \approx 437$, h=4096), DINOv2 ($\bar{k} \approx 44$), CIFAR-100 ($\bar{k} \approx 151$), and ImageNet-1K scale SAEs ($\bar{k} \approx 629$, not sparsity-matched) are each trained on their own ID embeddings.

Scorers. (a) **SAE-Recon**: reconstruction error $\|z - \hat{z}\|^2$. (b) **SAE-ActMahalanobis**: single Gaussian on ID SAE activations, Mahalanobis score in activation space (empirical covariance with 1e-6 ridge; the zero-inflated covariance is ill-conditioned — a caveat for practitioners). (c) Dense baselines: MSP, Energy, Mahalanobis distance to the nearest class mean under the total ID covariance (1e-4 ridge; a simplification of Lee et al. (2018), who use the tied within-class covariance), KNN (k=50, cosine), and a ViM-style residual score (residual orthogonal to the top principal direction of ID features minus $0.5\times$ max-logit; a simplification of Wang et al. (2022), with α=0.5 chosen on SVHN, so this row is not leakage-free on SVHN); plus **PCA-Recon** (k=256 linear-basis reconstruction), **ReAct-Energy** (99th-percentile clamping + probe), and a **DINOv2-KNN** strong reference. Attribution uses per-unit reconstruction contribution $\delta_i = a_i^2 \|W_{\mathrm{dec}[:,i]}\|^2$ and its normalized entropy H. (The trained SAE's decoder columns are *not* unit-normed, so δ_i genuinely weights by column norm; the unit-norm assumption appears only inside Appendix E.5's idealized proposition, where it makes $\delta_i \to a_i^2$ — the gap between the idealized and deployed attributions is why Appendix E.5 is motivation, not a description of the trained model. $H_{\mathrm{norm}} = H / \log k(x)$, the entropy of the normalized per-sample δ distribution divided by $\log$ of the sample's active-unit count; it varies with active count by construction — the partial correlations in Finding 9(iv) quantify that dependence.)

Regimes. All regimes share the frozen backbone, mean pooling, and the CIFAR-10 SAE; only the *audit classifier* changes (detection: CIFAR-10 probe; intervention audit: ImageNet-1k head, with the probe as its head-dependence control; each additional ID: its own SAE).

The Sparse Reconstruction Audit (§5's reusable artifact). A six-step protocol (`scripts/sparse_reconstruction_audit.py`) characterizing a frozen encoder + trained sparse SAE + audit classifier, producing a single JSON report. The released script ships configurations for the five systems studied here; a new encoder is added as one configuration entry (cache paths, SAE checkpoint, dimensions). Steps: (1) **Reconstruction error** — SAE-Recon AUROC: does reconstruction fidelity separate ID from OOD? (2) **MSP drop under the round-trip** — relative MSP drop of decode(encode(z)) and its drop-as-diagnostic AUROC under the audit classifier. (3) **Attribution concentration** — mean H_{norm} and cross-sample modal-unit concentration. (4) **Noise-matched control** — matched-Gaussian MSP drop: is the effect structured or generic perturbation? (5) **Selectivity report** — ID vs OOD drops and recon-vs-drop AUROC comparison. (6) **PCA linear-basis control** — steps 1–2 with a k=256 PCA round-trip: how much is generic low-rank reconstruction, and what does the sparse code add?

Interpretation criteria combine absolute and within-run relative contrasts (absolute: step-1 AUROC $\geq 0.95 \wedge$ step-2 OOD drop $\geq 50\% \wedge$ step-4 noise magnitude $\leq 15\%$; relative: round-trip OOD-selective if the OOD drop is $\geq 5\times$ the ID drop *and* ≥ 20 points; head-dependent if the wide-head drop is $\geq 5\times$ the probe drop *and* the OOD drop is ≥ 20 points (the bootstrap CIs of the two drops are reported alongside by `scripts/paired_stats.py`); diffuse/localized by H_{norm} gap ≥ 0.10 to the run's own ID value, and modal concentration $\geq 3\times$ the run's ID value). The constants were set from this study's regimes and are not calibrated for false-positive control on unseen systems; each verdict in the JSON report carries its margin to threshold, and the DINOv2/ResNet rows of Appendix A.6 (which fail several arms by design) serve as the out-of-regime negative controls. The scorecard classifies mechanism — it is not a safety certification, and cross-encoder verdicts do not transfer (Appendix A.6).

E.2 Mechanism details (§5)

This section concerns the MSP drop, the mechanism-facing quantity; it says nothing about detection AUROC, which is §4's object. ImageNet-O is adversarially constructed against ImageNet classifiers, so the wide-head regime measured here is the leakage-retaining regime by design: the leakage-controlled evidence is the SVHN/CIFAR-100 selectivity (round-trip OOD drops 3.9–4.3$\times$ the ID drop, below the audit's $5\times$ threshold; Table 4) and the probe-matched 4.5% bound of §5.1.

If a sparse subset of units drove OOD detection, ablating those units should reproduce the effect. We test this causally with a **control battery** that separates (a) the lossy SAE encode→decode *round-trip* from (b) the ablation of specific units. Two protocol details matter and are stated up front: the "random active units" control must select genuinely *active* units (a common masking scheme silently

selects inactive units, making the control a no-op — Appendix B.1), and cached embeddings are post-LayerNorm, so the frozen head takes them without re-normalization; because the head was trained on the CLS token, the mean-pooled audit is itself a head-input mismatch, and the CLS-pooled control is reported in Appendix B.2. **Sample-set convention:** the battery table (Appendix C.5) uses the *held-out* half of ImageNet-O (n=1,000, the second half of the cache in file order; fixed unit sets are selected on the disjoint first half), while Tables 4b/A.6/9a–9b use the *full* 2,000-sample ImageNet-O — the two protocols give 70.3% (full) vs 73.3% (held-out) for the same round-trip; both are reported. All numbers use mean pooling and the strict sparse SAE ($\bar{k} \approx 132$).

Protocol. For each sample: encode $z \to$ activations $a \to$ (optionally ablate a unit subset) $\to$ reconstruct $\hat{z} = W_{\text{dec}}a' + b_{\text{dec}} \to$ pass $\hat{z}$ through the frozen ViT head $\to$ measure max-softmax (MSP) confidence. We report relative drop on mean MSP against the original-embedding baseline, including the reconstruction-only control $\hat{z}_{\text{full}} = \text{decode(encode(z))}$ with no ablation.

Finding 4 — the collapse is dominated by the round-trip, not unit identity, and is not generic noise. Reconstruction-only — *no units ablated* — drops OOD MSP by 70.3% (full) / 73.3% (held-out) versus 10.7% on ID; the OOD drop is 3.9× (SVHN), 4.3× (CIFAR-100), and 6.6× (ImageNet-O) the ID drop (Appendix C.5, Table 4). Every unit-ablation protocol that leaves ID intact reproduces the round-trip drop to within ~1 point at k = 5; at larger k the OOD-selected fixed set adds more (k = 10/20: +2.9/+4.7 points on held-out ImageNet-O and +8.9/+10.7 on SVHN, with ID changes $\leq$ 0.5 points; `intervention_controls_mean_s42_imagenet.json`), and a **Gaussian perturbation matched to the reconstruction MSE** drops ImageNet-O MSP by only 7.2% (ID 1.0%), so the structured reconstruction, not generic noise of the same magnitude, produces the drop. With the head's native input — CLS pooling and a CLS-trained SAE — the round-trip is smaller but still selective (47.0% vs 11.2% ID, 4.2×; matched noise 14.1%), so the effect does not depend on the pooling mismatch (Appendix B.2).

Base-confidence control. Because ImageNet-O's baseline MSP (mean base MSP 0.49 on ImageNet-O vs 0.38 on ID; the case-study exemplar starts at 0.913) leaves more drop headroom than the ID's, the relative-drop comparison could in principle be a headroom artifact. Stratifying by *ID-fitted* baseline-MSP quintiles — ID and OOD samples compared at matched base confidence — largely removes the concern: in four of five bins the OOD relative drop exceeds the ID drop by 3.7–11× (≈0.20: 14.1% vs 52.4%; ≈0.32: 14.6% vs 67.7%; ≈0.51: 13.4% vs 73.9%; ≈0.79: 6.8% vs 72.6%), but in the lowest-confidence bin (≈0.10) the two are close (12.2% vs 14.8%). The *absolute* drop shows the same pattern (bin 5: 0.051 ID vs 0.580 OOD; bin 1: 0.013 vs 0.030), so the selectivity is not a base-confidence artifact except at very low confidence. The leakage-free sets stratify the same way (SVHN: 17.6% at base ≈0.09 rising to 52.7% at base ≈0.72; CIFAR-100: 17.3–54.0%; against the ID's 6.8–14.6% bin band) (`outputs/results/base_msp_stratified_drops.json`; `scripts/stratify_msp_drops.py`). Additional controls (Appendix C.5): the drop is direction-driven, not norm-driven (norm-preserving reconstruction: 73.8%); a PCA round-trip gives a comparable 73.0%; a random-orthonormal projection and an untrained SAE drop everything (96.6%/99.6%) — confirming that *structured, ID-fit* reconstruction, not generic perturbation, carries the selectivity. All core comparisons are per-sample significant (paired Wilcoxon + bootstrap: round-trip vs matched noise +0.61 [0.58, 0.64], p < 1e-200; Appendix C.3). Scope: because a k=256 PCA round-trip reproduces the collapse (73.0%) and matches the raw detection signal, the mechanism this section accounts for is a property of the *reconstruction-based scoring pipeline* (encoder–code–head), not an effect unique to sparse codes; what the sparse code specifically contributes is the attribution description of §6 and the ID-anchor decomposition of Finding 5. The mechanism is a property of the pipeline; what is specific to sparse codes is the descriptive layer.

Finding 5 — the largest-contribution units are ID anchors. The only protocol adding over the round-trip is per-sample top-δ ablation, and a magnitude-matched random control reproduces most of it (magnitude-matched vs top-δ: ID 72.6% vs 91.3%; OOD 83.0% vs 87.6%): the effect scales with removed contribution *mass*, not unit identity. Removing the single largest-contribution unit drops ID MSP by 61.8% (k=1) while adding only +7.5 points on ImageNet-O; scaling k∈{1,2,5,10,20}, ID drops saturate near 100% (Fig. 1b). The largest units gate *in-distribution* confidence — ID-anchor features.

Finding 6 — the diagnostic is an explanatory lens, not SAE-unique. Scoring by MSP drop $S = \text{MSP}_{base} - \text{MSP}_{recon}$ achieves AUROC 0.865 on held-out ImageNet-O (0.859 on the full set). Its specificity is bounded by three controls: the effect is **direction-driven** (norm-preserving

reconstruction: 73.8% drop), **not SAE-unique** (a k=256 PCA round-trip gives a comparable drop, 73.0%, and a stronger drop-AUROC, 0.950 vs 0.859), and **not a better detector than reconstruction error itself** (plain SAE-Recon: 0.997). The sparse SAE's distinctive value is therefore not raw reconstruction power but the attribution geometry it enables (§6) and the ID-anchor decomposition (Finding 5).

Mismatched Wide-Head Reconstruction Sensitivity. The collapse is a property of the *head+code system*, not of the sparse code alone. Re-running the reconstruction-only control under different audit classifiers yields a four-way structure: under the CIFAR-10 probe (10 classes) the round-trip is nearly inert (4.5% OOD drop, 0.5% ID); under the wide 1,000-class ImageNet head on the narrow CIFAR-10 ID it reaches 70.3% OOD vs 10.7% ID (47.0% vs 11.2% when the head receives its native CLS input; Appendix B.2); a 100-class probe on CIFAR-100-as-ID stays nearly inert (0.7% OOD, 4.5% ID); and the same wide head on its native ImageNet-1K ID is small (3.2% native-ID vs 6.4% on ImageNet-O; Appendix A.5). The full cross-backbone control battery is in Appendix A.6.

We name this **Mismatched Wide-Head Reconstruction Sensitivity**: a wide pretrained head's max-softmax confidence is sharply sensitive to a reconstruction deficit only when the head's label space does not match the audited ID distribution. In the four-cell grid above, head width and ID dataset co-vary in three cells; the same-ID head-width sweep (§5.1 and the table at the end of this subsection) separates them: at a fixed CIFAR-10 ID, restricting the ImageNet head to W random classes raises the ImageNet-O drop monotonically with W while the ID drop stays flat, and heads trained on the ID label space are inert at comparable widths. Width is therefore the lever for a mismatched head, and the mismatch is what makes it bite. This setting is meant to model a common deployment pattern — a broad pretrained classifier applied to a narrower target domain — with the caveat that here the head also receives mean-pooled input it was not trained on (Appendix B.2). In exactly that regime, confidence drops under reconstruction are (a) real and OOD-selective, but (b) a property of the head–code system, not evidence that the SAE "detects OOD units"; they warn against reading softmax-collapse as intrinsic to the encoder. Width alone does not explain it (a matched 100-class probe on a different ID stays nearly inert, 0.7%), and it is not a logit-scale artifact (on ImageNet-O it coincides with a margin collapse, top-1 vs top-2 logit 2.14 $\rightarrow$ 0.99, while the logit norm does not shrink; offset-invariant MSP gives the identical 70.3% drop; fixed temperatures keep the OOD drop 4.8–6.6$\times$ the ID drop at T $\in \{1, 2, 5\}$; `outputs/results/calibration_analysis.json`). Two further checks (`scripts/head_subspace_margin.py`, `outputs/results/head_subspace_margin.json`) address candidate explanations of the width dependence. *Softmax arithmetic:* a fixed shrinkage of the logit margin lowers MSP more when more classes share the softmax, but the margin collapse is not fixed — the relative top-1 minus top-2 margin drop on ImageNet-O is 0.7%, 28.7%, 40.6%, 51.8%, and 53.9% at W = 10, 30, 100, 300, and 1,000 (ID: 10.6%, 10.9%, 7.3%, 3.6%, -0.3%), so width changes which logits compete, not only how the softmax scales them. *Head / ID-subspace misalignment:* projecting the class-centered head weights onto the ID reconstruction subspace (top 290 principal directions, 99% of the round-trip outputs' variance), the ImageNet head keeps 53–54% of its energy outside it at every W, and a CIFAR-10 logistic-regression probe even more (59%) although its round-trip drop is negligible (ID 0.2%, ImageNet-O 1.5%); this simple misalignment measure therefore does not explain which heads are sensitive. The diagnostic is therefore informative for auditing a wide pretrained head on a new (narrow) deployment ID, and weak when the head is at home — including on a self-supervised ViT with no native head (DINOv2: round-trip inert under a probe, ImageNet-O -16.3%, a confidence *increase* — the direction of the confidence change is itself head-idiosyncratic; Appendix A.6).

Steps 1 (recon AUROC) are encoder-dependent but head-independent ($\geq$0.997 for natural-image semantic shift on every backbone); steps 2–4 are head-dependent. Step 6 quantifies "why not just PCA": PCA matches the SAE on everything that does not involve units (recon AUROC, raw drop) — the SAE's advantage is the sparse attribution geometry (§6) — while on ResNet the SAE round-trip exceeds PCA (14.1% vs 4.8%) and on DINOv2 their signs differ, so the two bases are not redundant everywhere.

One Example Through the Audit. To make the protocol concrete, Appendix C.5 (Table 4c) walks one example per set (deterministic index 0 of each test cache) through all six steps under the Ima-

geNet head. The pattern at single-sample resolution is the paper's mechanism, with one caveat: the ImageNet-O example has a 4.9× higher reconstruction deficit than the ID example, more diffuse attribution, and a 96% relative MSP drop under reconstruction (0.913 → 0.040) versus 13% under matched noise; but the ID example is not uniformly modest — starting from a low base MSP (0.28), its round-trip drop is 45%. Single-sample drops are noisy; the selectivity claim is about aggregates.

Intervention stability across SAE seeds. The full battery on seeds 42/1337/2024 confirms the aggregate structure (mean, with the range over the three seeds in brackets): round-trip ID/held-out IN-O drops 9.7% [9.2, 10.7] / 73.9% [73.3, 74.4]; MSP-drop AUROC 0.870 [0.865, 0.872]; even the single-unit ID-anchor effect is stable (61.9% [61.3, 62.6]). The aggregate causal structure is reproducible even though unit identities are not (§6) — which is why our claims rest on aggregate statistics.

CLS-pooled control battery. Re-running the full battery with CLS pooling (CLS-trained SAE; the head receives its native input; `intervention_controls_cls_s42_imagenet.json`) preserves the structure: the round-trip drops ID MSP by 11.2% versus 47.0% on full ImageNet-O (53.6% held-out; drop-AUROC 0.774); top-δ ablation removes 91.4% of ID confidence vs 67.6% on held-out ImageNet-O, magnitude-matched random ablation 73.9% vs 61.2%; the single largest unit alone removes 61.8% of ID confidence (k=1); and fixed or random unit sets add at most 1.1 points over the round-trip. ID anchors and round-trip dominance therefore do not depend on the pooling mismatch; the round-trip magnitude does.

Same-ID head-width sweep. ID = CIFAR-10 test, OOD = full ImageNet-O, fixed SAE and head weights; the ImageNet head's softmax is restricted to W random classes (10 draws per W; mean ± s.d. over draws; W = 1,000 is the full head); relative MSP drop under the round-trip and under matched-Gaussian noise (`scripts/head_width_sweep.py`, `outputs/results/head_width_sweep.json`). Restricting a pretrained head to a random class subset is not the same as training a narrower head; the matched probes of §5.1 cover that case.

Pooling	Quantity	W = 10	W = 30	W = 100	W = 300	W = 1000
mean	ID drop	8.8 ± 2.7%	13.7 ± 1.5%	13.9 ± 2.7%	13.1 ± 2.5%	10.7%
mean	IN-O drop	−3.0 ± 23.1%	14.3 ± 9.1%	35.1 ± 9.0%	57.0 ± 2.9%	70.3%
mean	IN-O noise	−15.4 ± 1.6%	−12.9 ± 3.1%	−3.1 ± 1.4%	3.8 ± 0.7%	7.2%
CLS	ID drop	7.3 ± 2.5%	10.4 ± 1.9%	11.7 ± 2.1%	13.4 ± 2.0%	11.2%
CLS	IN-O drop	−40.0 ± 36.8%	−23.4 ± 11.9%	−6.7 ± 19.4%	24.3 ± 7.9%	47.0%
CLS	IN-O noise	−31.8 ± 3.6%	−25.0 ± 5.9%	−6.0 ± 2.2%	6.5 ± 0.7%	14.1%

Trained heads of varying width. Linear softmax heads trained (AdamW, 30 epochs, 300 images per class) on frozen mean-pooled ImageNet-1K train embeddings of W random classes (five draws for W ≤ 100, three for W = 300, all classes for W = 1,000), so the head's input pooling matches its training; ID = CIFAR-10 test, OOD = full ImageNet-O, CIFAR-10 SAE (`scripts/trained_head_width.py`, `outputs/results/trained_head_width.json`).

W	Head val. acc.	ID drop	IN-O drop	IN-O noise
10	0.980	8.7%	3.0% ± 4.7	−14.2%
30	0.972	9.6%	12.6% ± 7.2	−14.6%
100	0.933	9.4%	28.8% ± 3.9	−4.8%
300	0.885	8.5%	38.0% ± 2.4	0.8%
1000	0.780	6.0%	40.3%	6.4%

ImageNet-100 as a narrow in-domain ID. ID = 100 random ImageNet-1K classes (all train images for the SAE; 5,000 validation images as ID test), with its own ReLU SAE trained by the CIFAR recipe (the λ ceiling stops it at $\bar{k} \approx 282$, 13.8% of h, denser than the CIFAR SAE); OOD = full ImageNet-O; heads trained as above on the listed classes.

W	Head classes	ID drop	IN-O drop	IN-O noise
100	ID classes only (matched)	0.2%	1.0%	−1.0%
300	ID + 200 other classes	0.0%	10.3%	0.1%
1000	ID + 900 other classes	−0.4%	10.4%	1.6%

Second backbone: DeiT-B/16. `deit_base_patch16_224` (supervised, native ImageNet-1k head on the CLS token; the head reproduces the true-forward logits to 4e-06 on CLS features), with the ViT pipeline unchanged: same CIFAR-10 subsets and transforms, all of ImageNet-O, and a ReLU SAE per pooling trained by the CIFAR recipe ($\bar{k} \approx 64$ CLS, 79 mean). Same-ID head-width sweep (ten random class subsets per W):

Pooling	Quantity	W = 10	W = 30	W = 100	W = 300	W = 1000
CLS	ID drop	10.7% ± 3.4	17.5% ± 3.9	21.3% ± 4.1	23.7% ± 3.8	21.9%
CLS	IN-O drop	−56.5% ± 40.6	−33.4% ± 33.0	13.0% ± 15.7	44.6% ± 8.5	71.3%
CLS	IN-O noise	−53.1% ± 6.1	−45.0% ± 8.9	−14.5% ± 3.1	5.2% ± 2.1	18.7%
mean	ID drop	13.5% ± 2.8	21.4% ± 2.4	27.6% ± 3.1	28.4% ± 5.9	24.6%
mean	IN-O drop	11.3% ± 9.9	38.5% ± 7.8	69.3% ± 3.9	85.5% ± 2.1	93.9%
mean	IN-O noise	−27.0% ± 2.4	−23.5% ± 4.0	−8.4% ± 2.3	1.3% ± 1.3	5.8%

A matched 10-class CIFAR-10 probe is inert with CLS pooling (ImageNet-O −16.1% vs ID 0.8%, i.e. confidence rises) but not with mean pooling (16.3% vs 1.3%). The width dependence therefore replicates across the two supervised backbones, while its selectivity (OOD/ID ratio) and the matched-probe behaviour under mean pooling are backbone-specific (`scripts/second_backbone_deit.py`, `outputs/results/deit_backbone.json`).

ID anchors beyond the ViT/CIFAR-10 setting. Control battery (k = 5; relative mean-MSP drop; full ImageNet-O) on DeiT-B/16 with CLS pooling and its native head, and with ImageNet-100 (second class draw, SAE seed 42) under the broad trained head (`scripts/id_anchor_generalization.py`, `outputs/results/id_anchor_generalization.json`). Single largest unit (k = 1): DeiT ID 72.9%, ImageNet-100 ID 54.0%. In ImageNet-100 the magnitude-matched random control reproduces much less of the top-δ effect than elsewhere; with this denser SAE the control's 10k-unit cap (Appendix B.4) may bind, so we do not read this as evidence for unit identity.

Protocol	DeiT ID	DeiT IN-O	IN-100 ID	IN-100 IN-O
Round-trip only	21.9%	71.3%	−0.5%	9.0%
Top-δ (per-sample)	95.5%	82.6%	80.4%	31.7%
Random 5 active units	35.4%	71.8%	0.8%	10.0%
Magnitude-matched random	93.6%	76.2%	13.0%	19.2%
Fixed ID-anchor set	26.6%	71.5%	−0.5%	9.7%

ImageNet-100: second class draw, three SAE seeds. Matched 100-class head vs broad head (ID + 900 classes), both trained once on frozen embeddings; round-trip drops per SAE seed.

Seed	$\bar{k}$	Matched ID	Matched IN-O	Broad ID	Broad IN-O	Broad IN-O noise
42	276	0.0%	0.4%	−0.5%	9.0%	2.2%
1337	277	0.1%	2.2%	−0.2%	10.3%	2.2%
2024	278	0.0%	1.2%	−0.4%	9.8%	2.2%

Third-party SAE audit (Prisma CLIP ViT-B/32). Artifact `Prisma-Multimodal/sparse-autoencoder-clip-b-32-sae-vanilla-x64-layer-11-hook_resid_post` on open_clip `laion/CLIP-ViT-B-32-DataComp.XL-s13B-b90K`; the reconstruction replaces the output of block 11 for all 50 tokens (the SAE's own hook point), and CLS-token ablations use $\delta_i = a_i^2 \|W_{\text{dec}}[i]\|^2$. Heads are zero-shot text classifiers (seven prompt templates). Mean active latents on the CLS token: 77 (CIFAR-10), 105 (ImageNet-O), 74 (ImageNet-100); reconstruction-error AUROC (CLS) 0.494 for CIFAR-10 vs ImageNet-O and 0.774 for ImageNet-100 vs ImageNet-O (`scripts/external_sae_audit.py`, `outputs/results/external_sae_audit.json`). Relative MSP drop, ID / ImageNet-O:

Setting	Round-trip	Matched noise	Top-δ (CLS, k=5)	Magnitude-matched random
CIFAR-10 ID, matched 10-class head	1.4% / 0.8%	1.5% / 1.1%	15.1% / 8.8%	19.5% / 13.2%
CIFAR-10 ID, 1,000 ImageNet classes	9.5% / 5.6%	9.4% / 12.0%	47.9% / 30.1%	41.1% / 30.7%
ImageNet-100 ID, matched 100-class head	5.3% / −1.0%	3.3% / 8.6%	28.8% / 15.7%	39.6% / 17.7%
ImageNet-100 ID, 1,000 classes	16.0% / 5.6%	8.2% / 12.0%	48.5% / 30.1%	54.2% / 30.7%

Same-ID width sweep (CIFAR-10 ID, W random ImageNet classes, ten draws; round-trip drop ID / ImageNet-O) for W = 10, 30, 100, 300, 1,000: 0.5% / −0.1%; −2.6% / −1.1%; −4.0% / −0.8%; 1.7% / 1.4%; 9.5% / 5.6%.

Controlled comparison at the same hook. To separate the SAE's training distribution from the other differences between the Prisma audit and our main setting, we hold the CLIP model, the hook (block-11 output, all 50 tokens), the zero-shot heads, and the evaluation sets fixed, and train two ReLU SAEs with the paper's recipe (768→2048, L0-bisection scheduler with its λ ceiling, 15 epochs, seed 42) on 250,000 block-11 token activations each, from 5,000 CIFAR-10 train images or 5,000 ImageNet-1k train images (five per class) (`scripts/controlled_clip_sae.py`, `outputs/results/controlled_clip_sae.json`). Both stop at the λ ceiling but at different densities (CLS L0 in the table); the third arm below, fit on Food-101 at the CIFAR-10 SAE's density, controls for this. FVE is ID / ImageNet-O; the zero-shot column is CIFAR-10 accuracy after the round-trip (clean 95.8%); drops are ID / ImageNet-O:

SAE fit on	FVE	CLS L0	Recon AUROC	Zero-shot	Matched 10-class	1,000 classes	Noise (OOD, 1,000)
CIFAR-10 (audited ID)	0.81 / 0.66	197	0.999	96.1%	−0.6% / −18.1%	7.7% / 60.0%	18.0%
ImageNet-1k	0.81 / 0.78	323	0.496	93.9%	2.3% / −7.0%	22.8% / 31.1%	12.0%

Same-ID width sweep (ten draws; round-trip ID / ImageNet-O, matched noise on ImageNet-O in parentheses) for W = 10, 30, 100, 300, 1,000 — CIFAR-10-fit SAE: 0.9% / 10.8% (2.8%); 1.7% / 21.1% (6.2%); 2.0% / 34.9% (10.9%); 2.1% / 47.0% (14.4%); 7.7% / 60.0% (18.0%). ImageNet-fit SAE: 2.0% / 2.3% (2.0%); 5.3% / 6.6% (4.3%); 11.3% / 14.5% (7.7%); 16.4% / 22.9% (9.8%); 22.8% / 31.1% (12.0%). Only the SAE fit on the audited ID yields an OOD-selective, width-dependent round-trip well above matched noise; under the matched 10-class head its round-trip raises OOD confidence rather than collapsing it, as with the DINOv2 probe (Appendix A.6).

Third training distribution and seeds. Because the ImageNet-fit SAE is also fit on the OOD set's source domain, the two-arm comparison above cannot separate "fit on the audited ID" from "not fit on the OOD domain". We therefore add a third arm fit on neither — Food-101 train images — and repeat all three arms with seeds 42, 1337, and 2024, with the same recipe and 250,000 training tokens each (`scripts/controlled_clip_sae_seeds.py`, `outputs/results/controlled_clip_sae_seeds.json`; block-11 outputs of the evaluation sets are cached, and the remaining forward pass matches the full model exactly). Ranges over the three seeds; drops are ID / ImageNet-O under the 1,000-class head; the last column is the ImageNet-O drop at W = 10 and 1,000:

SAE fit on	FVE (ID / OOD)	CLS L0	Recon AUROC	Zero-shot after	1,000-class drop	Noise (OOD)	Width (OOD)
CIFAR-10 (audited ID)	0.81 / 0.66	181–197	0.999	95.8–96.3%	5.2–10.1% / 60.0–63.6%	18.2–18.8%	−10.6 to 3.6% → 60.0–63.6%
ImageNet-1k (OOD source domain)	0.81 / 0.78	313–323	0.460–0.496	93.8–94.0%	22.8–27.7% / 30.5–31.1%	12.5–13.0%	1.3–2.3% → 30.5–31.1%
Food-101 (neither)	0.65–0.66 / 0.68–0.69	192–203	0.352–0.365	86.2–88.3%	51.9–59.0% / 41.0–44.2%	16.9–17.8%	−13.2 to −6.0% → 41.0–44.2%

Only the SAE fit on the audited ID yields an OOD-selective round-trip well above matched noise. The Food-101 SAE has the CIFAR-10 SAE's density, so density does not explain the difference; it reconstructs both CIFAR-10 and ImageNet-O less faithfully (FVE 0.65–0.66), its reconstruction error

ranks ImageNet-O as *more* in-distribution than CIFAR-10 (AUROC 0.352–0.365), and its round-trip lowers ID confidence more than OOD confidence.

Stronger noise nulls. The matched-Gaussian control adds isotropic noise with the set-level reconstruction MSE, which puts little energy into the directions the head reads. Three per-sample nulls, five draws each, match the round-trip more closely (`scripts/stronger_noise_nulls.py`, `outputs/results/stronger_noise_nulls.json`; full ImageNet head, mean pooling): (a) *logit-matched* noise, an isotropic direction scaled so that its displacement of the class-centered logits equals the round-trip's for that sample; (b) *residual-covariance* noise, drawn from the covariance of ID reconstruction residuals (CIFAR-10 train) and scaled to the sample's own residual norm; (c) a *residual swap*, adding the residual of another random sample of the same set. Relative MSP drop (mean, ± s.d. over draws):

Set	Round-trip	Logit-matched	Residual-covariance	Residual swap
ID (CIFAR-10)	10.7%	0.8% (±0.1)	1.0% (±0.3)	1.6% (±0.1)
ImageNet-O	70.3%	6.3% (±1.0)	7.6% (±0.6)	8.6% (±0.5)
SVHN	41.7%	−0.5% (±0.3)	−0.8% (±0.3)	10.5% (±0.4)
CIFAR-100	46.5%	1.5% (±0.4)	1.9% (±0.3)	0.9% (±0.2)

Even noise that displaces the logits exactly as far as the round-trip leaves confidence nearly unchanged, and so does the set's own residual distribution applied to the wrong sample; the collapse therefore depends on each sample's own reconstruction residual, not on its magnitude, its logit-space footprint, or the residual distribution of the set.

Width sweep on every OOD set. The same-ID sweep above uses ImageNet-O, which was built adversarially against ImageNet classifiers. Repeating it with the identical protocol (mean pooling, seed-42 SAE, the same class subsets and ten draws) on all 13 OOD sets of Table 1 (`scripts/width_sweep_all_ood.py`, `outputs/results/width_sweep_all_ood.json`) gives the round-trip MSP drop by W, the excess over matched noise at W = 1,000, and the OOD/ID ratio at W = 1,000:

Set	W = 10	30	100	300	1,000	Excess over noise (1,000)	OOD/ID (1,000)
ID (CIFAR-10)	8.8%	13.7%	13.9%	13.1%	10.7%	—	—
SVHN	8.7%	18.5%	17.6%	31.6%	41.7%	42.0	3.9×
LSUN-r	4.9%	10.9%	12.1%	21.5%	25.5%	25.7	2.4×
LSUN-c	5.2%	13.2%	21.4%	35.7%	48.6%	49.4	4.5×
TinyImageNet	9.7%	21.0%	31.8%	46.7%	61.6%	58.5	5.7×
Places365	−0.3%	13.6%	24.4%	43.4%	52.8%	48.8	4.9×
CIFAR-100	9.2%	17.9%	25.3%	36.4%	46.5%	44.6	4.3×
STL-10	5.5%	9.0%	12.9%	19.1%	27.8%	25.7	2.6×
MNIST	−17.9%	−19.4%	0.9%	29.3%	28.1%	26.9	2.6×
DTD	−9.0%	12.2%	22.8%	47.1%	65.2%	58.9	6.1×
Food-101	−2.6%	33.1%	46.3%	71.0%	77.4%	73.3	7.2×
ImageNet-O	−3.0%	14.3%	35.1%	57.0%	70.3%	63.2	6.6×
ImageNet-R	0.1%	13.2%	29.2%	44.8%	63.6%	59.0	5.9×
ImageNet-Sketch	−5.1%	1.9%	12.6%	28.7%	45.4%	42.6	4.2×

The drop rises monotonically with W on 11 of 13 sets (SVHN dips once between W = 30 and 100; MNIST is negative at narrow widths and flat from W = 300), exceeds matched noise at W = 1,000 on every set, and reaches the audit's 5× OOD/ID ratio on 5 of 13, so the width dependence generalizes beyond ImageNet-O while its selectivity is set-dependent.

E.3 Attribution details (§6)

We attribute each sample's OOD signal to individual SAE units via $\delta_i = a_i^2 \|W_{\mathrm{dec}[:,i]}\|^2$ and compute the Shannon entropy H of the normalized contribution distribution: low $H \Rightarrow$ localized shift, high $H \Rightarrow$ diffuse shift.

Finding 8 — the localized↔diffuse spectrum is real and metric-consistent. ID is most localized (0.636) and ImageNet-R the most diffuse (0.771), with ImageNet-O (0.764) and ImageNet-Sketch

(0.757) at the same diffuse end; MNIST (0.645) sits next to ID. The ordering holds across metrics (top-1 mass 0.250 → 0.105; PR 14.0 → 48.6; effective units 27.1 → 89.5); the ID-vs-ImageNet-O gap is strong (Δ = +0.128 [0.123, 0.132], Mann–Whitney p < 0.001, d = 1.38). Cross-sample modal-unit concentration separates *extreme* domain shift (MNIST 87.4%, ImageNet-Sketch 29.4%, SVHN 15.1% of samples fire one modal unit) from natural-image semantic shift (8.5–12.8%) and ID (4.5%) — a *cross-sample* property, not a within-sample one (spectrum visualization: Figure 5; per-sample distributions: Figure 6).

Finding 9 — the spectrum is not an artifact; the SAE attribution is sharp and non-redundant. (i) The dense decision attribution is flat (0.878–0.888) and does not separate ID from OOD, while the SAE spans 0.636–0.771; per-sample Spearman correlations are weak (-0.20 to +0.10). (ii) A k=256 PCA attribution is uniformly more diffuse (0.84–0.93): the sparse dictionary concentrates attribution far more than a linear basis. (iii) Null models confirm dictionary-dependence (column-norm permutation lowers H_{norm} by 0.022–0.063). (iv) Confound checks: H_{norm} correlates with reconstruction error (+0.43 to +0.67) and active count (+0.28 to +0.69); the partial correlation with OOD-ness controlling for reconstruction error remains positive but small (pooled ρ = 0.18, residual AUROC 0.63).

Concept labeling. Frequent units (>1% firing, n=1,170) labeled by dominant CIFAR-10 class: mean purity 0.273, 2.4× the label-shuffle null — the sparse code is *partially* class-aligned, not a clean class decomposition. CLIP zero-shot mapping was degenerate (it collapsed to a single token across units; `outputs/concept_labeling.txt`), so class purity is the labeling basis.

Unit stability across SAE seeds (negative result). Across three mean-SAE replicates (split fixed; only SAE training varies), only ~8% of units have a >0.7-correlated partner in another seed's dictionary, and after mapping each unit to that best-correlated partner the per-sample top-10 overlap is Jaccard 0.23–0.24 (`scripts/matched_topk_overlap.py`): the specific dictionary is not a canonical decomposition. The exact-k TopK SAE family — the modern training recipe, which removes the λ-schedule confound — shows the same qualitative result: matched top-10 Jaccard 0.17–0.18, with 9.7–10.4% of units having strongly correlated counterparts (Appendix C.4): exact-k training slightly raises the functional-correspondence rate (vs 8.2–8.8% for ReLU) but still yields no stable unit-level vocabulary. This quantifies, for a vision-SAE OOD pipeline, the dictionary non-identifiability that superposition theory predicts (Elhage et al., 2022) and that Paulo and Belrose (2026) and Fel et al. (2025) report for language- and vision-model SAEs; we scope the claim to our regime (768-d mean-pooled ViT embeddings, h=2048, matched k, three seeds) and, at fixed training data, wider dictionaries are *less* stable (dictionary-size sweep below). But the *aggregate endpoints are stable* (ID H_{norm} 0.636/0.619/0.624; ImageNet-O 0.764/0.752/0.766; partial ρ remains positive in every seed, 0.12–0.23, residual AUROC 0.59–0.68): our interpretability claims rest on aggregate statistics, never on individual units (unit visualizations: Figure 7).

Failure-mode structure. HDBSCAN (McInnes et al., 2017) on 4D per-sample features (computed with the denser comparison SAE, `sae_h2048_l1030`) recovers 2 clusters per dataset, driven almost entirely by reconstruction error; OOD is a *continuous* reconstruction-error continuum with a novel-concept sub-population (the SVHN high-recon cluster is the only group with non-zero novel-unit count), not discrete mechanistic failure modes (UMAP; (McInnes et al., 2018); visualization: Figure 4).

Dictionary-size sweep. TopK SAEs with k/h fixed at 6.4%, three seeds per size (42/1337/2024), CIFAR-10 recipe; ranges over the three seed pairs (`scripts/dictionary_size_sweep.py`, `outputs/interpretability/dictionary_size_sweep.json`). The h = 2,048 row reproduces the TopK partner fractions of Appendix C.4. The Jaccard column compares raw unit indices, which carry no cross-seed order, and so only tracks its chance level (~5/h); correlation-matched overlaps are in the text.

h	k	Units with corr. partner	Raw-index top-10 Jaccard	SAE-Recon AUROC (IN-O)
1,024	66	12.4–14.3%	0.0037–0.0061	0.993–0.994
2,048	131	9.7–10.4%	0.0022–0.0040	0.995
4,096	262	7.3%	0.0015–0.0020	0.996
8,192	524	6.4–6.8%	0.0002–0.0011	0.997

Training-data size. The three ImageNet-1K TopK scale SAEs (Appendix A.5; h = 4,096, k = 262, seeds 42/1337/2024) match the h = 4,096 row above in dictionary size and sparsity but are trained on the 1.15M-embedding ImageNet-1K cache rather than the 16k-embedding CIFAR-10 train split. Measured the same way on 10,000 random ImageNet validation embeddings (the size of the CIFAR-10 test set), 37.3% of units have a >0.7-correlated partner (37.9–38.6% on all 50,000), against 7.3% at CIFAR scale, while the correlation-matched per-sample top-10 overlap is similar to CIFAR scale (Jaccard 0.22–0.23 vs 0.17–0.18 for the h = 2,048 CIFAR-10 TopK SAEs; `outputs/interpretability/matched_topk_overlap.json`). The higher partner rate, however, comes from rarely firing units (`scripts/unit_stability_by_firing.py`, `outputs/interpretability/unit_stability_by_firing.json`): 41.0–41.2% of the ImageNet-scale units never fire on the evaluation set, only 19.4–19.8% fire on $\geq$1% of samples, and among those the partner rate is 14.0–15.6%, against 7.1–8.1% (ReLU) and 15.6–17.9% (TopK) at CIFAR scale (dead units: 15.4–16.6% and 41.7–42.5%). Units that fire regularly are therefore no more stable with ~70$\times$ more training data; the ID also differs between the two settings (`scripts/scale_unit_stability.py`, `outputs/interpretability/scale_unit_stability.json`).

Fixed unit sets as detectors. The unit-level controls of §5 and §6 test causal effects on head confidence and seed stability; this paragraph tests detection directly (`scripts/fixed_unit_detection.py`, `outputs/interpretability/fixed_unit_detection.json`). With the paper's selection rule for the OOD-discriminative fixed set (units ranked by the mean-activation gap between the ImageNet-O selection half and CIFAR-10 train), the summed activation of the selected units scores CIFAR-10 test against the held-out ImageNet-O half, and, without re-selection, SVHN and CIFAR-100. For the ReLU SAEs (three seeds), one, five, and twenty units reach held-out ImageNet-O AUROC 0.77–0.85, 0.90–0.94, and 0.97–0.98, against 0.998–0.998 for SAE-Recon and 0.994–0.994 for the active-unit count alone; five units transfer to SVHN at 0.74–0.84 (SAE-Recon 0.812–0.821) and to CIFAR-100 at 0.61–0.66 (SAE-Recon 0.833–0.835). The TopK SAEs, whose active count is fixed by construction, give 0.67–0.83 with five units and 0.86–0.94 with twenty. The selected units are seed-specific: of each seed's five units, 20–20% (ReLU) and 0–40% (TopK) have a >0.7-correlated partner in another seed, and the set mapped through best-correlated partners overlaps the other seed's own selection with Jaccard 0.00–0.43 (ReLU) and 0.00–0.50 (TopK). Small fixed unit sets therefore flag ImageNet-O within a given model, as Karmacharya et al. (2026) report for their setting, but less well than reconstruction error or activation density, with weak transfer across shifts and without a seed-stable identity.

E.4 Detection and robustness details (§4, §7)

Rankings against all baselines, including the published ViM and Mahalanobis, are stated in the main-text Findings 1, 10, and 12; this subsection adds details. On CIFAR-10, SAE-Recon exceeds the stronger-backbone DINOv2-KNN reference on Places365 (0.992 vs 0.990) and ImageNet-Sketch (0.999 vs 0.994) and ties it on ImageNet-R; ReAct-Energy is markedly weaker on every semantic-shift set (ImageNet-O 0.931 vs 0.997); and SAE-Recon's seed standard deviation is $\leq$ 0.009 per set (Appendix C.2).

CIFAR-10-C. Under the Table-1 protocol (CIFAR-10 probe; strict sparse SAE, three seeds), the published ViM is best overall (0.746) and on noise and weather (0.837, 0.706); Energy and MSP lead blur; SAE-Recon averages 0.660 (weather 0.677) and is best only on saturate, spatter. The DINOv2-KNN reference averages 0.732. With the ImageNet head and a denser SAE instead, SAE-Recon ranks first on noise and weather; that ranking does not hold under the Table-1 protocol (Appendix A.2).

CIFAR-100 as ID. (KNN's 0.294 on MNIST is a manifold-density artifact — MNIST forms a compact cluster inside a diverse 100-class ID manifold; `scripts/diagnose_knn_mnist_cifar100.py` — the same artifact recurs on iNaturalist at scale. Full table: Appendix A.3.)

Cross-architecture. The control battery on ResNet-50 (overcomplete SAE, $\bar{k} \approx 437$; overcompleteness 2.0$\times$ vs the ViT's 2.7$\times$) differs sharply from the ViT: the round-trip is small (15.7% vs 73.3% on held-out ImageNet-O), and magnitude-matched random ablation reproduces most of the top-δ

effect (31.3% vs 36.5%; at k=20 it exceeds top-δ). **Finding 11:** the causal structure is architecture-dependent; in all architectures the largest-contribution units are ID anchors (ResNet 17.3%, ViT 91.3% at k=5), and the reconstruction-based MSP-drop diagnostic is at least as strong as any unit-targeted intervention (Appendix A.6).

DINOv2 (self-supervised, no native head). Detection transfers: the published ViM is best on SVHN (0.998 vs SAE-Recon 0.997), and SAE-ActMahalanobis is best or tied-best on STL-10 (0.934), Places365 (0.999), ImageNet-O (0.999), ImageNet-R (0.997, tied with KNN), ImageNet-Sketch (0.994, tied) — DINOv2 features near-saturate most benchmarks. But under a probe audit the round-trip collapse is absent (ImageNet-O -16.3%, Appendix A.6): the large collapse is specific to the ViT + wide-ImageNet-head combination, consistent with §5.1. Full table: Appendix A.4.

ImageNet-1K scale validation. Same scorers, native 1,000-class head, seven OOD sets (Places365/DTD/SUN397/iNaturalist + ImageNet-O/R/Sketch); a single ReLU scale SAE (h=4096, achieved $\bar{k} \approx 629$, 15.4% of h) plus a sparsity-matched TopK variant below.

Finding 12 — limitations at both ends. *(i) Detection does not transfer at the denser protocol $\bar{k}$:* the ReLU scale SAE ($\bar{k} \approx 629$) is best on **none** of the seven sets (SAE-Recon 0.61–0.83); dense Mahalanobis/ViM lead six of the seven columns (MSP leads iNaturalist); KNN collapses on iNaturalist (0.354) — the manifold-density artifact of Appendix A.3. *(ii) The round-trip effect largely disappears:* under the ReLU scale SAE the round-trip drop is still larger on OOD than on ID and larger than matched noise (OOD 4.0–7.8% vs ID 3.2%; noise $\leq$0.7%), but at 1.2–2.4$\times$ the ID drop it fails the audit's OOD-selectivity criterion, and ImageNet-O falls 70% $\rightarrow$ 6.4% — completing the Mismatched Wide-Head story (§5.1) — and the SAE round-trip is far *smaller* than the PCA round-trip, so the MSP-drop diagnostic is weak (drop-AUROC 0.47–0.56), consistent with Finding 6. *(iii) Attribution partially transfers:* the H_{norm} gap narrows (ID 0.804 vs OOD 0.809–0.860), but extreme domain shift localizes more sharply (ImageNet-Sketch modal concentration 0.454 vs ID 0.029). *(iv) Sparsity-matched validation improves detection but does not reach the best dense baseline.* Because the λ-ceiling protocol cannot reach the CIFAR SAE's 6.4% at scale, we retrain the scale SAE as an **exact-k TopK SAE** (k = 262 = 6.4% of h, same 1.15M-embedding cache): it beats the denser ReLU scale SAE on **all seven** OOD sets on SAE-Recon (a SAE-vs-SAE comparison, not a detection-value claim; the TopK SAE-ActMahalanobis score is weaker than its ReLU counterpart on every set, 0.48–0.69), and it trails the strongest dense baseline (including the published ViM) on all seven: by 0.006–0.007 on ImageNet-O and ImageNet-R and 0.030–0.051 on DTD, Places365, SUN397, iNaturalist, and ImageNet-Sketch (seed 42; across three seeds 0.005–0.007 and 0.029–0.056). Family and sparsity change together in this comparison; because TopK and ReLU were roughly equivalent at matched k at CIFAR scale (Appendix C.4), we attribute the improvement mainly to sparsity matching, but a ReLU SAE at $\bar{k} \approx 262$ would be needed to separate the two. Its audit keeps the qualitative ordering: round-trip larger than matched Gaussian (IN-O 5.5% vs 1.5%; ID 1.8%), far below the PCA round-trip (53.9%), and extreme-shift modal localization (ImageNet-Sketch 0.218 vs ID 0.059) (`outputs/results/imagenet1k_{detection_topk,audit_topk}.json`; Appendix A.5). Note the primary scale-audit numbers (i)–(iii) come from the denser ReLU SAE ($\bar{k} \approx 629$, 15.4%); the TopK audit rows are the sparsity-matched (6.4%) counterpart.

Attribution entropy is architecture-specific. With the sparse ViT SAE, the natural-image semantic-shift sets are the *most* diffuse (ImageNet-R 0.771, ImageNet-O 0.764); ResNet-50 shows the *opposite* (0.808 vs ID 0.886) — global self-attention scatters semantic OOD mass, CNN local receptive fields trigger localized texture-novelty detection. The causal structure and the spectrum are both architecture-dependent; the ID-anchor effect is architecture-independent.

E.5 Geometric motivation: variance separation in sparse codes

We present this as geometric motivation, not a formal theorem; assumptions are explicit (Appendix D).

Proposition (variance concentration under sparsity). For the whitened Mahalanobis distance D of SAE activations ($a(x) \in \mathbb{R}^h$, $k(x)$ active units, $\bar{k} \ll h$; Σ estimated by Ledoit–Wolf shrinkage; unit-normed decoder columns), exactly k-sparse activations on a fixed ID support S satisfy: (1) ID variance of D grows with k and stays small for $k \ll h$; (2) OOD inputs activating units outside S acquire large, non-concentrating terms ($\mu_j \approx 0$, small σ_j), so OOD variance does not concentrate; (3) magnitude drift *within* S does not break concentration.

We check this two ways. **(a) Controlled synthetic data** with a known k-sparse dictionary (Ledoit–Wolf covariance): all four predictions hold (Appendix D). **(b) Real SAEs** trained at five L0 targets (0.05–0.50 of h, i.e. $\bar{k} \approx 129$–$1{,}101$ active units; empirical covariance with the deployed 1e-6 ridge). Here only prediction (2) is supported: OOD variance exceeds ID variance by 5×10^3–2.4×10^4 at every target. Prediction (1) is *not* supported: ID variance is largest in the sparsest code (24.9 at $\bar{k} \approx 129$) and falls as the code gets denser (2.49 at $\bar{k} \approx 680$, 2.88 at $\bar{k} \approx 1{,}101$). The real SAEs violate the proposition's idealizations (fixed support, independent active coordinates, unit-normed decoder; Appendix D), so the proposition motivates the OOD-side behaviour but does not describe the ID side of the deployed model. **Scope:** the synthetic check covers $k/h \leq 6.2\%$; every real SAE in this paper, including the deployed one ($\bar{k}/h = 6.4\%$) and the ImageNet-1K scale SAE (15.4%), lies just outside or well outside it.

Observations (empirical, no formal claim). The ID–OOD mean-distance gap is largest in the sparsest code and narrows as the code gets denser (294.6 at $\bar{k} \approx 129 \rightarrow$ 167.0 at $\bar{k} \approx 1{,}101$): in this range sparsity widens the separation. Decoder geometry does not distinguish the SAE from a random basis (pairwise |cos| 0.026 vs random 0.029 vs PCA 0.000), so we make no claim about semantic versus geometric coherence from it.

E.6 Extended discussion, practitioner guide, and limitations (§8)

What sparsity buys, and what it does not buy (full decision table: Appendix C.7). The contribution is a tested boundary: raw OOD AUROC — PCA often suffices; semantic-shift detection — SAE leads several sets (with two of the wins being ceiling ties); unit-level OOD localization — **no** (the main negative result, tested in both SAE families); aggregate attribution geometry and explaining confidence collapse — the *decomposition* is a contribution (round-trip vs ablation mass; ID anchors), but the entropy quantity itself is substantially a re-encoding of reconstruction error and active count (Finding 9(iv)) and is not established as detection-relevant; a deployment-ready detector — not necessarily.

Scope of the mechanistic account. Three different scopes must be kept apart. *SAE-family-robust* (ReLU and exact-k TopK at matched k): reconstruction-dominance, noise-specificity, ID-anchor structure, and unit-level seed-idiosyncrasy (the diffuse-attribution spectrum is ReLU-specific). *Architecture-dependent* (round-trip magnitude and attribution spectrum: 70–73% round-trip on ViT vs 14–16% on ResNet; ViT diffuse vs ResNet localized entropy ordering; §7, Appendix A.6). *Head-regime-specific*: the Mismatched Wide-Head Sensitivity magnitude itself, which appears only in the wide-head-on-narrow-ID configuration. The account therefore generalizes as a method (the audit + controls) and as the negative unit-level claim, not as a claim that every backbone/head system shows a 70% round-trip collapse. One further scope limit: the six-step audit has been validated only on this paper's own encoder+SAE+head stack; a first third-party audit (the Prisma CLIP SAE, §5.1) shows the round-trip effect is absent for a general-purpose SAE, and a controlled comparison at the same hook traces this to the SAE's training data; auditing the SAEs released by Olson et al. (2025) and Oh et al. (2026) remains future work.

Practitioner guide. Use sparse SAEs when you need an interpretable attribution *description* or want to audit *why* confidence changes; do not expect them to beat KNN/Mahalanobis/ViM for raw AUROC; test PCA reconstruction before concluding sparse features are necessary; and run the head-dependence audit (§5.1) before interpreting confidence drops when a wide pretrained head is deployed on a narrow ID. The audit's decision criteria (reconstruction-dominant / head-dependent / diffuse-vs-localized / noise-robust) act as a mechanism scorecard for a new encoder+SAE+head system; verdicts carry their threshold margins and do not transfer across encoder families (Appendix A.6) — they characterize, and are not a safety certification. Steps 1–2 and 5 require labeled OOD data and are offline-only; steps 3–4 are computable on ID plus an unlabeled stream. If raw detection is the goal, start from KNN/ViM/Mahalanobis on the strongest available embedding; if drift monitoring is the goal, log reconstruction error and attribution entropy against ID-calibrated thresholds, remembering that recon error is matched by PCA and is not evidence the sparse code is doing anything specific.

Limitations. (a) Backbone coverage: ViT-B/16 primary; DeiT-B/16, ResNet-50, and DINOv2 as checks, plus one third-party CLIP SAE. (b) Concept vocabulary limited to CIFAR-10 class purity; CLIP-based labeling was degenerate (collapsing to a single token). (c) Semantic-similarity failures

persist. (d) Appendix E.5's observations are empirical, not proven. (e) Scale: a sparsity-matched exact-k TopK SAE addresses the denser-primary-SAE caveat (§7), but scale results remain a fixed-model validation; scale margins come from three TopK SAE seeds without formal significance testing. (f) Individual units are seed-idiosyncratic in both SAE families studied, so all claims rest on stable aggregates; wider dictionaries are less stable at fixed training data, and far more training data raises correspondence only among rarely firing units (Appendix E.3); small fixed unit sets detect OOD within a model but are seed-specific (Appendix E.3), and unit ablations at ImageNet scale are untested. (g) Head-dependence bounds the round-trip's deployment scope to wide-head settings; the same-ID width sweeps use ImageNet-derived heads (restricted pretrained heads and heads trained at each width) on two IDs, CIFAR-10 and ImageNet-100. (h) Dense-baseline hyperparameters were not swept (KNN k fixed at 50), so some SAE-lead margins may shrink under tuned baselines. (i) The audit's thresholds were set on the studied regimes and are not false-positive-calibrated for unseen systems. (j) The battery lacks an ID-ward matched-norm displacement null (e.g., shrinkage toward the ID centroid); the PCA round-trip is the generic low-rank displacement control, and the base-MSP stratification has been run (§5, base-confidence control), but an ID-ward null could further narrow what the reconstruction-specific component is. (k) Only one external SAE (Prisma CLIP ViT-B/32) has been audited; the Olson et al. (2025) and Oh et al. (2026) artifacts are future work. (l) The ImageNet head is applied to mean-pooled features, not the CLS token it was trained on; with its native CLS input the round-trip drop on ImageNet-O is 47% rather than 70% (Appendix B.2), so the headline magnitude partly reflects this input mismatch.