BEYOND RETRAINING STABILITY: AUDITING DUPLICATION DEPENDENCE IN VISION SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders learn directions from a chosen activation corpus, making their features sensitive to both model representations and the distribution of training examples. Can a direction recur across retrainings yet depend on repeated exposure to particular images? We investigate this question through controlled near-exact duplication in activation corpora from frozen vision transformers. Our audit compares each reference direction's displacement after corpus intervention with its variability across unchanged-data retrainings. We contrast full removal with copies-only removal, which retains the original image, and include size-matched placebos for both. Directions that are stable under unchanged-data retraining can lose their closest counterparts when repetitions are removed. Above the estimated detection threshold, copies-only removal reproduces most of the full-removal effect and identifies substantially overlapping directions. Wider dictionaries exhibit lower detection thresholds across independent paired experiments, although absolute thresholds vary. These thresholds remain audit-dependent, and calibration diagnostics reveal tail inflation among stable latents, precluding an unconditional false-discovery-rate guarantee. Furthermore, combinations of retrained features recover much of the reference activation signal, distinguishing dedicated-direction loss from information loss. Our findings show that retraining stability alone does not establish robustness to duplication in the SAE training corpus.

1 INTRODUCTION

Sparse autoencoders (SAEs) re-express model activations as sparse combinations of learned directions, often interpreted as features or concepts. Wider dictionaries aim to resolve finer-grained structure. Yet an SAE dictionary is fitted to a finite activation corpus, so its directions can reflect both the frozen representation and input frequency. This creates a provenance problem: *can repeated exposure in the SAE training corpus induce a dedicated direction even when the underlying representation is fixed?*

A direct test would remove repeated examples, retrain the SAE, and ask whether the corresponding direction disappears. However, SAE dictionaries are unstable across retrainings (Paulo & Belrose, 2025). In our reference CLIP ViT-B/16 configuration, the median nearest-direction distance across 40 retrainings on identical data is 0.365. Failure to recover a direction after changing the corpus may therefore reflect ordinary optimisation variability rather than the intervention. Testing corpus dependence requires asking whether a direction changes *more than it changes when the data are left unchanged*. We develop an instability-controlled audit around this comparison. The vision backbone is frozen throughout; planted duplicates are processed once, their activations are cached, and only the SAE is retrained. Forty same-data retrainings estimate each reference latent's ordinary variability. We compare full removal and its size-matched placebo with copies-only removal and its own placebo. The copies-only arm retains the original image, testing dependence on additional exposures. It does not fully separate image identity from repetition, since repetition may enable an image-specific direction to form.

The experiments yield three main findings. First, controlled near-exact repetition is associated with target-associated SAE directions whose removal-induced displacement exceeds measured same-data

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.

retraining variability. At the reference configuration, full removal flags 100–114 directions per replicate while the matched placebo flags none, with an estimated 50% detection dose of $k^* = 5.5$ extra copies. Second, removing only the copies flags 96–102 directions, substantially overlaps the directions identified by full removal, and yields relative displacement ratios of 0.85–1.00 above threshold across the three configurations with all four intervention arms. Third, susceptibility changes strongly with dictionary width. On a fixed 50,762-image corpus, increasing width from 4,096 to 32,768 lowers the estimated detection dose from 23.4 to 1.1 copies. Across five independent paired experiments, the k^*_{4096}/k^*_{16384} ratios range from 3.30 to 6.32. Absolute thresholds vary substantially, so the replicated direction is more reproducible than any single threshold.

These findings concern the provenance and survival of *individual SAE directions*, not disappearance of information from the representation. Combinations of directions in a retrained dictionary recover much of the activation associated with a displaced reference direction. Nor do the experiments demonstrate memorization by the vision backbone, which is never retrained. Our claims therefore concern properties of an SAE dictionary fitted to a particular activation corpus.

Contributions. We **(i)** introduce an instability-controlled audit that compares each SAE direction with its own same-data retraining variability and size-matched placebo interventions while documenting calibration limits; **(ii)** test dependence on additional exposures with a copies-only intervention that retains the original image; **(iii)** establish a replicated width effect, with wider dictionaries requiring fewer repeated exposures for detection across five independent paired experiments; and **(iv)** characterize the audit's scope through recovery, sparsity, natural-duplicate, backbone, dataset, and other configuration analyses.

2 Related Work

Duplication and dedicated features. Carlini et al. (2022b) establish log-linear growth of language-model memorization with duplication, while Kazdan et al. (2026) argue that semantic duplicates increasingly behave like exact duplicates as capability grows. Closest in mechanism, Olah (2025) shows in a one-layer network computing $|x|$ that repeating a datapoint can induce a dedicated transcoder feature without a counterpart in the underlying model, including a no-repeat baseline and a repetition–sparsity scan. Our setting differs in both the object of study and the comparison: we study SAE dictionaries over a frozen ViT, intervene by removing repeated inputs and refitting the SAE, compare against 40 same-data retrainings, and vary dictionary width. The possibility of duplication-induced features itself is therefore not our novelty claim.

Corpus dependence and SAE faithfulness. Bricken et al. (2023) observe apparent memorization of specific training content in large SAE dictionaries without directly measuring its dependence on the corpus. Radhakrishnan et al. (2018) study memorization in overparameterized autoencoders for pixel reconstruction, while Klotz et al. (2026) report reduced perturbation alignment with greater overcompleteness in vision SAEs. Gu et al. (2023) separate configuration and data effects in diffusion-model memorization. More broadly, Heap et al. (2026) and Cho et al. (2025) motivate distinguishing properties of an interpretability method from properties of the underlying model.

Instability as an audit null. Paulo & Belrose (2025) document disagreement between SAEs trained on the same data. Gerasimov et al. (2026) quantify feature recurrence across retrainings using maximum-cosine matching, while Fel et al. (2025) improve stability through an archetypal constraint. We use retraining instability as the empirical null for a controlled corpus intervention rather than proposing a new matching statistic. The closest prior definition to our distance is Wang et al. (2024b)'s per-example encoder score; ours is defined per latent in a dictionary over a frozen encoder. Wang et al. (2024a) localise SSL memorization to units and report a depth effect, while Swann et al. (2026) identify VLA SAE "memorization features" using a generality heuristic rather than controlled duplication counts.

Statistical and membership auditing. We use Benjamini–Hochberg (BH; Benjamini & Hochberg, 1995) with candidate filtering motivated by Bourgon et al. (2010), while explicitly limiting any unconditional FDR interpretation. Enkhbayar (2025) applies FDR control to SAE latents using Model-X knockoffs against a synthetic null; our null instead comes from measured optimisa-

tion variability. Our supplementary membership analysis follows low-FPR reporting in Carlini et al. (2022a) and the evidentiary caution of Ertan et al. (2026). SAE interventions against PII leakage (Frikha et al., 2025) and near-copy generation of patient scans by latent diffusion models (Dar et al., 2026) provide broader context for the consequences of duplication.

3 Experimental Design

Frozen representation, retrained dictionary. The vision backbone is frozen throughout: it processes the corpus once to produce cached activations, and all SAEs within a configuration read the same cache. A reference SAE is trained on the full corpus D. To measure ordinary retraining variability, we retrain the same SAE architecture from scratch on unchanged D with $R = 40$ random seeds. Separate SAEs are then retrained from scratch under each corpus intervention. Thus, all measured changes arise at the dictionary level; the backbone is never retrained.

Controlled duplication and removal arms. Each target group g contains one original image and k near-exact copies generated by a small random crop and mild colour jitter. We use $k \in \{1, 2, 4, 8, 16, 32, 64\}$ with six targets per dose, giving 42 groups. The *dedup* arm removes the original and all of its copies, while the *placebo* arm removes the same number of ordinary images. The *copies* arm removes only the generated copies and retains the original at multiplicity one, while *placebo-c* matches this removal size. Every configuration has five dedup and five placebo replicates, and three configurations additionally have five copies and five placebo-c replicates. Each removal run intervenes on all 42 groups jointly. Target assignment therefore identifies reference specificity, not an isolated causal effect of removing that target's group; targets share the same intervention.

Experimental setup. We use Food-101 (Bossard et al., 2014) corpora containing 8,762–50,762 images and an 8,762-image EuroSAT corpus (Helber et al., 2019), each with 42 planted groups. The reference configuration uses CLIP ViT-B/16 (Radford et al., 2021), block 8, 20,762 images, dictionary width $m = 4,096$, and TopK 32; DINO ViT-B/16 (Caron et al., 2021) provides a second backbone. TopK SAEs with auxiliary loss (Gao et al., 2025) use vision tooling from Stevens et al. (2025) and Joseph et al. (2025), unit-norm decoder columns, Adam with learning rate 3×10^{-4}, batch size 4,096, and 40 epochs. Each image contributes its CLS token and 48 randomly selected patch tokens; activations are rescaled to mean norm $\sqrt{d}$. In the reference configuration, one latent never fires and 4,085 satisfy the ten-firing threshold. Appendix A gives the full configuration inventory.

4 Audit Method and Validation

4.1 Specificity and survival of a dedicated direction

Let

$$M_j(S) = \sum_{v \in S} \max(a_j(v), 0)$$

be latent j's post-activation mass, where $|g|$ and $|D|$ count activation vectors rather than images. We define duplicate-specificity as

$$\rho_{jg} = \frac{M_j(g)/|g|}{M_j(D)/|D|}, \qquad \rho_j = \max_g \rho_{jg}. \tag{1}$$

We first exclude latents firing fewer than ten times. For each target group, among the remaining latents satisfying $\rho_{jg} \geq 2$, we retain the top $N = 10$ by ρ_{jg}; the union across targets forms the candidate set. For attribution, each latent is assigned to the group maximising ρ_{jg}. Because ρ_{jg} is normalised by corpus-wide activation mass, it is invariant to global encoder-output rescaling.

To measure whether a reference direction survives retraining, we match it to its closest decoder direction in the retrained dictionary:

$$\Delta_j = 1 - \max_{1 \leq \ell \leq m} \cos\left(d_j^{\text{with}}, d_\ell^{\text{without}}\right). \tag{2}$$

The maximum makes this criterion permissive: any sufficiently similar surviving direction reduces Δ_j. Thus, Δ_j measures survival of a dedicated direction, not whether the same information is preserved across several directions.

4.2 Conditioning on retraining instability

Latents differ substantially in their baseline stability. Across the 40 same-data retrainings, the 5th and 95th percentiles of each latent's mean null distance are 0.006 and 0.680, while the pooled raw null has mean 0.344. Pooling these distances would therefore compare naturally stable and unstable latents against the same reference distribution.

Let μ_j and σ_j be the mean and standard deviation of decoder-matching distances Δ_j across the 40 same-data null retrainings. We standardise each intervention distance and compare it with a pooled leave-one-out empirical null:

$$\begin{aligned} z_j &= \frac{\Delta_j(\text{removal}) - \mu_j}{\sigma_j + \varepsilon}, \\ Z^0 &= \left\{ \frac{\Delta_j^{(r)} - \mu_j^{(-r)}}{\sigma_j^{(-r)} + \varepsilon} \right\}_{j,r}, \\ p_j &= \frac{1 + \#\{z \in Z^0 : z \geq z_j\}}{1 + |Z^0|}, \qquad \varepsilon = 10^{-6}. \end{aligned} \tag{3}$$

The stabiliser prevents division by zero. The index j in Z^0 ranges over all reference latents firing at least ten times across the corpus, not only the selected candidates. We retain repeated scores, so $|Z^0| = Rm_{\text{live}}$, where m_{live} is the number of eligible reference latents. Matching searches the full retrained dictionary. Standardisation conditions on each latent's scale, although pooling the resulting scores does not guarantee identical tail behaviour across latents. We therefore evaluate calibration empirically rather than assume exchangeability of the pooled standardized null.

4.3 Candidate testing and detection threshold

For the M candidates with ordered p-values $p_{(1)} \leq \cdots \leq p_{(M)}$, Benjamini–Hochberg (BH) rejects ranks $1, \ldots, r^*$, where

$$r^* = \max\left\{ r : p_{(r)} \leq \frac{rq}{M} \right\}, \qquad q = 0.05.$$

No candidates are rejected if this set is empty. The flagged dictionary fraction is

$$\text{MFR} = \frac{1}{m} \sum_j \mathbf{1}\{j \text{ flagged}\}.$$

The per-target rule limits each target's contribution to the candidate set, but the union size and BH denominator can still vary with target count. Candidate selection and the finite empirical null affect detectability; discoveries must therefore be interpreted alongside the control and calibration results below. Because specificity and matching distance share the reference dictionary, candidate filtering is not assumed independent of testing. We evaluate calibration after candidate selection; these empirical checks do not establish formal FDR control.

A target is detected if at least one flagged latent is assigned to it. We fit

$$P(\text{detect} \mid k) = \text{sigmoid}(\alpha \log k + \beta)$$

and report

$$k^* = e^{-\beta/\alpha},$$

the fitted extra-copy dose at which the audit's detection probability reaches 0.5. This threshold depends on candidate selection, matching, and null calibration; it is not an intrinsic memorization threshold. Numerical bootstrap intervals resample targets with all replicate outcomes and remain subject to the fitting limitations in Appendix B.

4.4 Empirical validation and calibration limits

The audit rarely flags unchanged or size-matched control retrainings (Table 1). Across 110 placebo runs spanning all 22 configurations, no latent is flagged. Leave-one-out evaluation of same-data retrainings produces a flag in only 1/40, 2/40, and 0/40 runs across the three tested configurations, with at most one false flag in any run. Ten additional retrainings held entirely outside calibration produce no flags at either of the two configurations tested. These controls characterize observed false flags but do not establish formal FDR control. Calibration varies with latent stability: among the most stable selected candidates, $P(p \leq 0.01) = 0.0220$ under null evaluation (Table 15). Tail inflation remains relevant near the reference configuration's BH thresholds, 0.0125–0.0143, and limits interpretation of discovery counts and fitted detection doses. All-null evaluations also do not establish FDR control when affected and unaffected latents coexist. We therefore assess corpus dependence through placebo-adjusted continuous displacement, copies-only interventions, and independent width repetitions, with flags and detection doses treated as procedure-dependent summaries. Firth refitting assesses sensitivity to logistic estimation; it does not recalibrate the underlying discoveries.

Table 1: False-flag checks for the complete audit. Leave-one-out evaluations reuse overlapping calibration sets; held-out evaluations use fresh retrainings against the original 40-run null. Counts are empirical diagnostics, not error-probability guarantees.

Check	Images	Width	Runs	Runs flagged	Max flags/run
Placebo	All 22 configurations		110	0	0
Leave-one-out	20,762	4,096	40	1	1
	50,762	4,096	40	2	1
	50,762	16,384	40	0	0
Held out	20,762	4,096	10	0	0
	50,762	16,384	10	0	0

5 Main Results

5.1 Repeated exposure supports target-associated SAE directions

At the reference configuration, the removal-minus-placebo displacement gap increases from $+0.059$ at four extra copies to $+0.374$ at eight. The targeted win rate is 0.838 $[0.771, 0.901]$, with 37 of 42 targets favouring removal after within-target aggregation. These comparisons do not use pooled-null calibration or BH, although they retain the target-selection and decoder-matching choices. Separately, the detection procedure flags 100–114 latents under full removal and none under placebo, with an estimated 50% detection dose of $k^* = 5.5$ $[3.4,\ 8.8]$ extra copies. These detector-dependent results remain subject to the calibration limitations in Section 4.4; further statistics appear in Table 2. For each target's most duplicate-specific reference latent, the targeted win rate averages the indicator that dedup displacement strictly exceeds ordinary-placebo displacement over all 5×5 replicate pairs, then averages equally across targets; ties contribute zero. Copies-only placebos are excluded. Separately, under single-group ownership, 57.1% of target-by-replicate outcomes acquire a flagged direction, with counts per target of median 1, mean 2.5, and maximum 10. Because full removal deletes both originals and copies, we next remove only the additional copies while retaining the originals.

Removing only the copies reproduces most of the effect. The copies-only intervention removes the generated copies while retaining the original image at multiplicity one. Despite the original remaining in the SAE training corpus, copies-only removal flags 96–102 latents, compared with 100–114 under full removal, while placebo-c flags zero in every replicate (Figure 1). The same directions are substantially involved. At reference, 110 of the 132 latents in the dedup union also occur in the copies-only union, while 93% of the copies-only union lies inside the dedup union. The union-level Jaccard overlap is 0.79. At the other two four-arm configurations the corresponding Jaccard overlaps are 0.88 and 0.77. Mean pairwise overlaps are 0.87 and 0.80, respectively, so the agreement is not solely a consequence of taking unions across replicates.

Table 2: Reference results using five removal replicates and 40 same-data retrainings. Dedup MFR is reported as mean ± standard deviation across five dedup runs. The win rate and targets favouring removal do not use pooled-null calibration or BH; all flag-based summaries do.

Instability-controlled audit		**Target-level summaries**	
Measure	**Result**	**Measure**	**Result**
Dedup MFR	0.0254 ± 0.0014	Targeted win rate	$\mathbf{0.838}$ $[\mathbf{0.771}, \mathbf{0.901}]$
Dedup flags	**100–114**	Targets favouring dedup	$\mathbf{37/42}$
Placebo MFR	0.0000	Outcomes with a flag	$\mathbf{57.1}\%$
Placebo flags	**0**		
Detection dose k^*	$\mathbf{5.5}$ $[\mathbf{3.4}, \mathbf{8.8}]$	Flags / target	1 median; 2.5 mean; 10 max

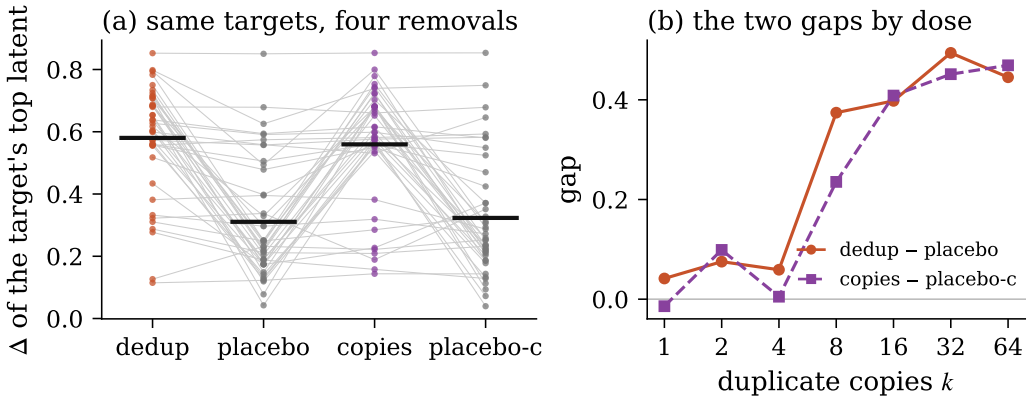


Figure 1: Four removal arms on the same targets. (a) Displacement of each target's most duplicate-specific reference latent, averaged across retrainings within each arm; heavy marks denote means across targets. (b) Removal-minus-placebo displacement by duplication dose. Removing only the additional copies closely tracks full removal while the original image remains in the corpus.

For each target, we select its most duplicate-specific reference latent and average its displacement over retrainings within each arm. Writing $\bar{\Delta}_A$ for the equally weighted target mean of these arm averages, we quantify the relative displacement as

$$S = \frac{\bar{\Delta}_{\text{copies}} - \bar{\Delta}_{\text{placebo-c}}}{\bar{\Delta}_{\text{dedup}} - \bar{\Delta}_{\text{placebo}}}. \tag{4}$$

This is a ratio of mean gaps, not a mean of target-specific ratios. We report S over targets with $k \geq k^*$ to reduce denominator instability. Bootstrap resampling uses this fixed subset and does not propagate uncertainty in the estimated cutoff.

Table 3: Copies-only intervention across the three configurations with all four removal arms. S is the relative displacement ratio in Eq. 4; intervals resample targets.

Images	**Width**	S $(k \geq k^*)$	**Targets favouring copies**
20,762	4,096	$\mathbf{0.914}$ $[\mathbf{0.796}, \mathbf{1.014}]$	$\mathbf{31/42}$
20,762	16,384	$\mathbf{0.854}$ $[\mathbf{0.770}, \mathbf{0.939}]$	$\mathbf{35/42}$
50,762	4,096	$\mathbf{1.001}$ $[\mathbf{0.903}, \mathbf{1.088}]$	$\mathbf{37/42}$

Across the three configurations, the relative displacement ratio is approximately 0.85–1.00 at doses $k \geq k^*$; it is not constrained to $[0, 1]$. At $k = 64$, full removal deletes 65 images from a target group whereas copies-only removes 64, so the informative distinction is that the original remains available at multiplicity one. Together with the flagged-set overlap, these results support dependence on repeated exposure while allowing that repetition may enable an image-specific direction to form.

5.2 WIDER DICTIONARIES REQUIRE FEWER REPEATED EXPOSURES

We next ask what controls the duplication dose required for a dedicated direction to become detectable. Holding a byte-identical 50,762-image activation corpus fixed, increasing dictionary width produces a large shift in detection threshold. The estimated k^* falls from 23.4 $[17.0, 33.7]$ at width 4,096 to 6.5 $[4.5, 9.1]$ at 8,192, 2.0 $[1.1, 3.3]$ at 16,384, and 1.1 $[0.4, 1.7]$ at 32,768.

At width $65{,}536$, one extra copy is detected in 25 of 30 target-by-replicate outcomes across six distinct target images. We report this observed frequency, 0.833, rather than extrapolate a 50% detection dose below the tested range. The width trend also appears without the empirical null. On the same 50,762-image corpus, the targeted win rate rises from 0.756 $[0.683, 0.830]$ at width $4{,}096$ to 0.882 $[0.806, 0.950]$ at $16{,}384$ and 0.969 $[0.936, 0.994]$ at $65{,}536$. Narrowing of the matching null with width is therefore unlikely to be the sole explanation, although other measurement effects are not excluded.

The width direction replicates across independent experiments. A single fixed-corpus sweep could depend on its particular target draw or reference dictionary. We therefore run four additional independent paired comparisons between widths 4,096 and 16,384, yielding five experiments including the original. Each experiment independently redraws targets and crops, rebuilds the activation cache, and retrains both reference dictionaries, their 40 same-data nulls, and all removal replicates.

Table 4: Five independent paired width experiments with 20,762 images each. Ratios compare fitted detection doses at widths 4,096 and 16,384. Brackets summarize the central 95% of numerical target-bootstrap ratios, conditional on the trained dictionaries; fitting limitations are documented in Appendix B.

	Detection dose k^*		**Paired width effect**	
Experiment	**Width 4,096**	**Width 16,384**	**Ratio**	**Bootstrap range**
Original	5.54	1.68	$\mathbf{3.30}$	$[\mathbf{1.75}, \mathbf{5.93}]$
Fresh	7.82	1.33	$\mathbf{5.89}$	$[\mathbf{3.68}, \mathbf{10.63}]$
Rep 1	5.18	1.47	$\mathbf{3.54}$	$[\mathbf{2.03}, \mathbf{7.05}]$
Rep 2	8.59	2.35	$\mathbf{3.65}$	$[\mathbf{2.17}, \mathbf{6.07}]$
Rep 3	10.59	1.68	$\mathbf{6.32}$	$[\mathbf{4.12}, \mathbf{9.94}]$

All five paired ratios exceed one (geometric mean 4.4). The wider dictionary also shows greater placebo-adjusted displacement at $k = 2$ and $k = 4$ in every experiment. This comparison avoids pooled-null calibration and BH but retains reference-latent selection and decoder matching. The ordering reverses at $k = 32$, supporting greater low-dose susceptibility rather than uniformly greater displacement. Firth refitting gives similar ratios (3.31–6.32); candidate-count and discrete-crossing checks appear in Appendix H.1. Absolute thresholds vary more: k^*_{4096} ranges from 5.18 to 10.59 across experiments. Thus, $k^* = 5.5$ is not a universal duplication threshold; the more reproducible finding is that wider dictionaries require fewer repeated exposures for detection.

5.3 LOSS OF A DEDICATED DIRECTION DOES NOT IMPLY INFORMATION LOSS

The matching statistic Δ_j asks whether a reference latent retains a close *single-direction* counterpart after retraining. Failure of that match does not establish that the information represented by the direction has disappeared; the signal may instead be redistributed across several directions. We therefore predict each eligible flagged reference latent's activation from the retrained sparse code, fitting on one half of the images in its planted group and evaluating on the other.

With one predictor, the dedup dictionary trails same-data retraining by 0.158 in R^2; with twenty predictors the gap falls to 0.064. Placebo remains close to the same-data null at every K, while copies-only and dedup remain close to one another. Much of the single-direction loss is therefore recoverable from a small combination of retrained directions. A residual gap remains under the tested 20-predictor ridge model, but this does not establish absent information: a larger or differently regularised predictor could recover more. Nor does this experiment establish generalisation to

Table 5: Out-of-sample R^2 for recovering flagged reference latents from retrained sparse codes. Results cover all 121 eligible flagged latents. Values are means with descriptive standard errors across latents; latents share target groups and trained dictionaries.

Arm	$K = 1$	$K = 5$	$K = 20$
Same-data null	0.786 ± 0.016	0.884 ± 0.010	0.920 ± 0.007
Placebo	0.794 ± 0.015	0.885 ± 0.009	0.927 ± 0.006
Copies	0.655 ± 0.017	0.799 ± 0.011	0.866 ± 0.008
Dedup	0.628 ± 0.017	0.785 ± 0.012	0.856 ± 0.012

unrelated images. We therefore interpret the primary audit as measuring survival of a dedicated direction rather than loss of information from the representation. Additional retrainings show similar held-out reconstruction quality across the tested widths ($R^2 = 0.97257$–0.97301; Appendix F.1), but this does not establish functional equivalence. The functional probe covers 11 of 42 targets at width $4{,}096$ and none at $16{,}384$. Within the eligible subset, full removal lowers correspondence descriptively relative to placebo, whereas copies-only removal does not. The functional width comparison therefore remains unresolved; the replicated width finding concerns geometric displacement and procedure-dependent detection.

5.4 Robustness and boundary conditions

The effect is observed across the two tested backbones and datasets and across moderate TopK settings. CLIP and DINO yield similar detection thresholds, as do Food-101 and EuroSAT and TopK 16 versus 32.

These comparisons are robustness checks rather than equivalence tests. Their overlapping intervals therefore should not be interpreted as evidence that backbone, dataset, or sparsity has no effect.

Depth produces a larger descriptive change, with k^* decreasing from 24.0 $[13.6, 63.2]$ at block 4 to 4.5 $[3.1, 6.4]$ at block 11. Because the frozen backbone never trains on the duplicated corpus, this depth trend is descriptive and does not identify a mechanism.

Table 6: Selected configuration checks. Within each comparison, only the named axis changes. Brackets give numerical target-bootstrap ranges; coverage is not established.

Axis	Setting	k^* [range]
Backbone	CLIP	5.5 $[3.4, 8.8]$
	DINO	5.4 $[4.0, 7.1]$
Dataset	Food-101	6.3 $[4.7, 8.4]$
	EuroSAT	6.7 $[4.3, 11.4]$
Depth	Block 4	24.0 $[13.6, 63.2]$
	Block 8	5.5 $[3.4, 8.8]$
	Block 11	4.5 $[3.1, 6.4]$
TopK	16	6.0 $[4.1, 8.5]$
	32	5.5 $[3.4, 8.8]$

Non-detection can reflect reduced detectability.

At TopK 64, no tested dose produces a flagged latent, although the removal-associated displacement remains positive. Over $k \geq 8$, the mean dedup–placebo gap decreases from 0.428 to 0.285, while mean own-null instability increases from 0.235 to 0.393 relative to TopK 32 (Table 7). Their descriptive ratio falls from 1.82 to 0.73.

Table 7: Detection boundary at increased TopK. Gap and instability are averaged over doses $k \geq 8$ on the same corpus and dictionary width.

Quantity	TopK 32	TopK 64
Removal gap	0.428	0.285
Null instability	0.235	0.393
Gap / instability	1.82	0.73
Targeted win rate	0.838	0.704
Latents flagged	100–114	0

The targeted win rate remains 0.704, and the dedup–placebo gap is $+0.385$ at $k = 64$. Thus, a removal-associated displacement persists despite zero discoveries.

The smaller gap and greater retraining variability are consistent with reduced detectability. Their ratio is descriptive, not the audit's decision rule. Zero flags therefore do not establish the absence of corpus dependence.

Naturally occurring near-duplicates produce a weaker signal.

We next test whether the controlled result extends to natural near-duplicates. For 42 Food-101 clusters constructed with cosine-similarity edges at threshold 0.985, no latent is flagged at either tested width. At width $16{,}384$, the dedup–placebo gap remains positive at $k = 2$ and $k = 4$, but the planted effect is $3.1\times$ and $4.2\times$ larger, respectively (Table 8). The targeted win rate is 0.859 for planted copies and 0.572 for natural clusters. The available natural clusters are few and their doses limited. These results limit generalisation to natural duplication; non-detection does not establish the absence of an effect or contradict the controlled finding.

Table 8: Controlled near-exact copies versus natural Food-101 near-duplicate clusters at width $16{,}384$. Gap denotes the dedup–placebo displacement of each target's most duplicate-specific latent.

Quantity	Planted	Natural
Gap at $k = 2$	$\mathbf{+0.162}$	$+0.052$
Gap at $k = 4$	$\mathbf{+0.232}$	$+0.055$
Targeted win rate	$\mathbf{0.859}$	0.572
Latents flagged	**212–220**	0

Corpus size under two training protocols. Wider dictionaries may have more capacity relative to training data. Our corpus-size experiments do not isolate this explanation. With 40 epochs fixed, detection thresholds are 6.3, 5.5, 13.8, and 23.4 at 8,762, 20,762, 35,762, and 50,762 images, respectively. Larger corpora also receive more optimisation steps. Fixing the budget at 12,000 steps gives thresholds of 3.4, 5.9, and 15.1 for the first three sizes, while the largest produces no detection. Under this protocol, larger corpora receive fewer passes through the data. These optimisation confounds prevent identification of a corpus-size scaling law; the replicated width comparisons instead hold corpus size fixed.

Summary of the empirical picture. Taken together, the experiments support a narrower claim than either "duplicates are memorised by the backbone" or "removal destroys information." The backbone is frozen throughout. Instead, repeated exposure in the activation corpus can support dedicated directions in the learned SAE dictionary beyond measured same-data retraining variability. Removing only the repeated exposures reproduces most of the full-removal effect, and wider dictionaries require fewer repeated exposures for detection across five independent paired experiments. At the same time, the boundaries matter. Absolute detection thresholds vary substantially across independent experiments; combinations of retrained directions recover much of the signal associated with a displaced reference direction; TopK 64 demonstrates that the audit can stop producing discoveries while a positive removal-associated displacement remains; and naturally occurring near-duplicates produce a substantially weaker signal than the controlled near-exact intervention. We therefore interpret the results as evidence about the provenance and stability of dedicated SAE directions, rather than as a universal duplication threshold, a claim of backbone memorisation, or evidence that the represented information itself disappears.

6 Conclusion

An SAE is fitted to a corpus as well as a model's activations. Same-data retraining variability supplies the comparison needed to test corpus dependence. Controlled near-exact repetition supports target-associated directions; deleting only repetitions reproduces most of the measured removal effect, and wider dictionaries exhibit lower procedure-dependent detection doses. The width direction replicates across five experiments while absolute thresholds vary. A direction's presence in a wide dictionary is therefore insufficient evidence of robustness to duplication in the SAE corpus.

AI use statement

In this work we used generative AI tools for these tasks requiring disclosure: proposing and refining hypotheses, including predictions subsequently evaluated in our experiments; developing the conceptual framework of the audit; formulating its mathematical claims and the derivation in Appendix L; designing and giving feedback on the methodology; implementing the audit, training

and analysis code; interpreting results; and generating small synthetic data sets, used only to validate statistical constructions before they were applied to real activations. We did not use them for translation, for cleaning or reformatting the image corpora, or for qualitative or thematic analysis; assistance with proofs is not applicable, as this paper contains none. We also used them for tasks with recommended disclosure: the figures, code editing, drafting and readability, literature search, and reference formatting. We have reviewed all AI-assisted work. The authors executed the experimental code on their cluster and checked the reported results against the resulting outputs. We take responsibility for the final text, methodology, analyses, references, and artifacts.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the national academy of sciences*, 107 (21):9546–9551, 2010.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6, 2023.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022a.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022b.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Seonglae Cho, Harryn Oh, Donghyun Lee, Luis Eduardo Rodrigues Vieira, Andrew Bermingham, and Ziad El Sayed. FaithfulSAE: Towards capturing faithful features with sparse autoencoders without external dataset dependencies. *arXiv preprint arXiv:2506.17673*, 2025.

Salman Ul Hassan Dar, Marvin Seyfarth, Isabelle Ayx, Theano Papavassiliu, Stefan O Schoenberg, Robert Malte Siepmann, Fabian Christopher Laqua, Jannik Kahmann, Norbert Frey, Bettina Baeßler, et al. Unconditional latent diffusion models memorize patient imaging data. *Nature biomedical engineering*, 10(3):458–472, 2026.

Tsogt-Ochir Enkhbayar. Which sparse autoencoder features are real? model-x knockoffs for false discovery rate control. *arXiv preprint arXiv:2511.11711*, 2025.

Murat Bilgehan Ertan, Emirhan Böge, Min Chen, Kaleel Mahmood, and Marten van Dijk. On the evidentiary limits of membership inference for copyright auditing. *arXiv preprint arXiv:2601.12937*, 2026.

Thomas Fel, Ekdeep Singh Lubana, Jacob S Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. *arXiv preprint arXiv:2502.12892*, 2025.

Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. PrivacyScalpel: Enhancing LLM privacy via interpretable feature intervention with sparse autoencoders. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 226–238, 2025.

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

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable features, reproducible subspaces: Understanding seed dependence in sparse autoencoders. *arXiv preprint arXiv:2606.12138*, 2026.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. In *International Conference on Learning Representations*, 2026.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Tom Henighan, Shan Carter, Tristan Hume, Nelson Elhage, Robert Lasenby, Stanislav Fort, Nicholas Schiefer, and Christopher Olah. Superposition, memorization, and double descent. *Transformer Circuits Thread*, 2023.

Sonia Joseph, Praneet Suresh, Lorenz Hufe, Edward Stevinson, Robert Graham, Yash Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake Aaron Richards. Prisma: An open source toolkit for mechanistic interpretability in vision and video. *arXiv preprint arXiv:2504.19475*, 2025.

Joshua Kazdan, Noam Levi, Rylan Schaeffer, Jessica Chudnovsky, Abhay Puri, Bo He, Mehmet Donmez, Sanmi Koyejo, and David Donoho. Scale dependent data duplication. *arXiv preprint arXiv:2603.06603*, 2026.

Jonas Klotz, Cassio F Dantas, Pallavi Jain, Diego Marcos, and Begüm Demir. Evaluating the interpretability of sparse autoencoders with concept annotations. *arXiv preprint arXiv:2606.24716*, 2026.

Ioannis Kosmidis and David Firth. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, 108(1):71–82, 2021.

Chris Olah. A toy model of mechanistic (un)faithfulness. *Transformer Circuits Thread*, 2025.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

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

Adityanarayanan Radhakrishnan, Karren Yang, Mikhail Belkin, and Caroline Uhler. Memorization in overparameterized autoencoders. *arXiv preprint arXiv:1810.10333*, 2018.

Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Interpretable and testable vision features via sparse autoencoders. *arXiv preprint arXiv:2502.06755*, 2025.

Aiden Swann, Lachlain McGranahan, Hugo Buurmeijer, Monroe Kennedy, III, and Mac Schwager. Sparse autoencoders reveal interpretable and steerable features in VLA models. *arXiv preprint arXiv:2603.19183*, 2026.

Wenhao Wang, Adam Dziedzic, Michael Backes, and Franziska Boenisch. Localizing memorization in SSL vision encoders. *Advances in Neural Information Processing Systems*, 37:60475–60516, 2024a.

Wenhao Wang, Muhammad Ahmad Kaleem, Adam Dziedzic, Michael Backes, Nicolas Papernot, and Franziska Boenisch. Memorization in self-supervised learning improves downstream generalization. In *International Conference on Learning Representations*, 2024b.

SUPPLEMENTARY MATERIAL

A CORPORA AND CONFIGURATION

Unless an experimental axis is explicitly varied, the reference configuration in Table 10 is used.

Table 9: Corpora and duplicate construction. For controlled groups, k denotes the number of extra copies; each group also contains its original image.

Split	Images	Groups	Duplicate construction
Food-101 controlled	8,762–50,762	42, $k \in \{1, 2, 4, 8, 16, 32, 64\}$	crop 90–100%, colour $\pm 6\%$
EuroSAT controlled	8,762	42, $k \in \{1, 2, 4, 8, 16, 32, 64\}$	identical construction
Food-101 *found*	20,762	42, $k \in \{2, 3, 4, 5, 6, 18\}$	naturally present, cosine ≥ 0.985
Food-101 held-out	20,000	–	never sampled into any corpus

Table 10: Experimental configuration. Bold entries denote the reference setting; remaining entries identify experimentally varied settings.

Setting	Configuration
Backbone	**CLIP ViT-B/16** (Radford et al., 2021); DINO ViT-B/16 (Caron et al., 2021)
Read-out	residual stream after block **8**; also blocks 4 and 11
Tokens per image	CLS + 48 randomly selected patch tokens (49 total)
SAE	TopK with auxiliary loss (Gao et al., 2025); unit-norm decoder columns
Width m	2,048 / 4,096 / 8,192 / 16,384 / 32,768 / 65,536
Active latents	16 / **32** / 64
Optimiser	Adam, learning rate 3×10^{-4}, batch size 4,096, 40 epochs
Input scaling	activations rescaled so that the mean norm is $\sqrt{d}$
Null retrainings R	**40** same-data seeds
Replicates	**5** dedup and **5** placebo per configuration
Nominal BH level q	**0.05**; top $N = 10$ candidate latents per target

Robustness experiments vary one principal axis at a time where possible, including backbone, read-out depth, dictionary width, sparsity, dataset, and corpus size. All configurations retain the same basic audit structure: a reference dictionary, same-data retrainings for estimating ordinary instability, and matched corpus interventions for measuring removal-associated displacement.

B UNCERTAINTY ESTIMATION, IN FULL

Target-bootstrap intervals resample targets together with their replicate outcomes. Other uncertainty summaries are identified separately.

Resampling unit. The target image. A single training run removes all 42 planted groups at once, so targets are measurements within a shared intervention rather than independent experiments (Section 3); resampling targets is the closest available approximation to an independent unit, and it is not a substitute for repeating the experiment. When a target is drawn, *all* of its replicate outcomes are carried with it, so the within-target correlation across the five replicates is preserved rather than broken up.

Draws, seed and interval. Original detection-threshold and paired-width analyses use 4,000 bootstrap draws; relative-displacement and targeted-win-rate analyses use 10,000, all with seed 0. The retrospective Firth analysis uses 4,000 draws with seed 20260925. Numerical ranges use the 2.5th and 97.5th percentiles without bias correction; discrete-crossing envelopes are described separately in Appendix H.1.

Stratification. The four detection thresholds and the two paired width ratios of Table 12 were computed two ways; the relative-displacement-ratio and overlap intervals elsewhere in the paper use the uniform scheme only. The uniform scheme resamples targets over all 42, so the number contributing at each dose varies between draws. The stratified scheme resamples six within each dose, holding the dose composition at the value the design fixed, which is arguably the closer match to how the experiment was built. Table 12 gives both. They agree closely, and the stratified intervals are equal or slightly narrower, as expected when a source of between-draw variation is removed. *We report the uniform intervals throughout the paper*: they are the wider of the two, and having computed both we prefer not to headline the tighter one.

Paired width ratio. Both widths of a pair are trained on the same corpus and the same targets. Each bootstrap draw therefore resamples target identities *once* and refits the emergence curve at both widths on that same resampled set, so the ratio is paired, preserving the dependence between the two widths rather than treating them as independent samples. The code refuses to pair two runs whose dose vectors differ.

Logistic-fit diagnostics and limitations.

We retrospectively classify the original saved paired target-bootstrap fits by outcome degeneracy, separation, slope, and crossing range. At $N = 10$, all 4,000 draws per experiment return finite numerical ratios, but some involve separation or extrapolated crossings (Table 11). Both widths receive the in-range, no-separation-detected classification in 3,496–3,996 draws per experiment. Optimizer convergence was not recorded and is not established by these classifications. The original experiment has 3,999 ratios above one; the other four have 4,000 each. These counts are not calibrated probabilities of a population width effect. All draws remain in the diagnostic accounting. Numerical bootstrap intervals are conditional on the trained dictionaries and subject to these fitting limitations. Appendix H.1 reports separate Firth-refitting and censored dose-crossing checks; neither validates the original intervals.

Table 11: Diagnostics of 4,000 original bootstrap draws per experiment at $N = 10$. Sep.: complete or quasi-complete separation; Out.: outside-range crossing; Const.: constant outcomes. Categories are mutually exclusive under the classification order. Both in range requires in-range crossings with no separation detected at either width; it does not certify convergence.

	4,096		16,384			Both in range
Experiment	Sep.	Out.	Sep.	Out.	Const.	
Original	1	0	25	12	0	3,963
Fresh	62	0	75	373	1	3,496
Rep 1	0	0	24	217	0	3,759
Rep 2	0	0	1	3	0	3,996
Rep 3	0	0	123	0	0	3,877

The above-threshold subset. The relative displacement ratio is computed over doses $k \geq k^*$ using the point estimate of k^* for that configuration. The subset is therefore fixed across bootstrap draws rather than recomputed within each draw, so its interval does not propagate uncertainty in k^* itself.

Scope of the bootstrap summaries. These summaries describe target-resampling variation conditional on the trained dictionaries. They do not propagate variation from reference training, null estimation, or independently repeated corpus interventions, and their coverage is not established. Appendix H examines between-experiment variation by changing the targets and reference dictionaries.

C Calibration of the complete procedure

The placebo gives observed false-flag counts on the arms we ran, but a finite set of controls cannot establish calibration in general. We therefore add an *internal cross-validation diagnostic*, at no training cost. The R same-data nulls are retrainings on an unchanged corpus, so a latent flagged in one of them is a false discovery by construction. Holding out each null r in turn and running the complete pipeline on it, including the per-target top-N candidate rule and Benjamini–Hochberg at q, gives R pseudo-experiments in which every flag is known to be false.

Table 12: Detection thresholds and paired width ratios under uniform and dose-stratified bootstrap resampling.

Configuration	k^*	**Uniform range**	**Stratified range**
20,762 images, $m = 4{,}096$	$\mathbf{5.5}$	$[\mathbf{3.4}, \mathbf{8.8}]$	$[3.8, 8.4]$
20,762 images, $m = 16{,}384$	$\mathbf{1.7}$	$[\mathbf{1.2}, \mathbf{2.4}]$	$[1.2, 2.4]$
50,762 images, $m = 4{,}096$	$\mathbf{23.4}$	$[\mathbf{17.0}, \mathbf{33.7}]$	$[17.6, 30.2]$
50,762 images, $m = 16{,}384$	$\mathbf{2.0}$	$[\mathbf{1.1}, \mathbf{3.3}]$	$[1.1, 3.1]$
Paired width ratio: $k^*(4{,}096)/k^*(16{,}384)$			
Original targets, 20,762 images	$\mathbf{3.3}$	$[\mathbf{1.8}, \mathbf{5.9}]$	$[1.9, 5.7]$
Fresh targets, 20,762 images	$\mathbf{5.9}$	$[\mathbf{3.7}, \mathbf{10.6}]$	$[3.9, 9.7]$

The exclusion is nested. Run r is dropped from the per-latent μ_j, σ_j *and* from the pooled standardised distribution Z^0 against which its p-values are read; within that construction each remaining run s is standardised with s itself also excluded, so no run contributes to its own reference. What this cannot do is substitute for fresh experiments: the R pseudo-experiments share one reference dictionary and heavily overlapping calibration sets, so they are not R independent repetitions, and 40 evaluations locate a 5% error probability only loosely. Fresh held-out retrainings provide a separate diagnostic, reported below.

Table 13: Null-vs-null calibration. Each held-out null is run through the complete pipeline as if it were a dedup arm, so every flag is a false discovery.

Configuration	**Runs flagged**	**Mean false flags**	**Max false flags**
20,762 images, $m = 4{,}096$	$\mathbf{1/40}$ $(\mathbf{0.025})$	$\mathbf{0.03}$	$\mathbf{1}$
50,762 images, $m = 4{,}096$	$\mathbf{2/40}$ $(\mathbf{0.050})$	$\mathbf{0.05}$	$\mathbf{1}$
50,762 images, $m = 16{,}384$	$\mathbf{0/40}$ $(\mathbf{0.000})$	$\mathbf{0.00}$	$\mathbf{0}$

Table 14: Sensitivity of the calibration diagnostic to candidate count N and null count R, reported as pseudo-experiments with at least one false flag. Left: N varies at $R = 40$; right: R varies at $N = 10$, so each denominator is R. Reduced null sets use the first R of 40 runs, not a random subsample; the reference settings are $N = 10$ and $R = 40$.

Candidate-count sensitivity ($R = 40$)				**Null-count sensitivity** ($N = 10$)			
N	20,762/4k	50,762/4k	50,762/16k	R	20,762/4k	50,762/4k	50,762/16k
5	$1/40$	$0/40$	$0/40$	10	$1/10$	$1/10$	$0/10$
10	$\mathbf{1/40}$	$\mathbf{2/40}$	$\mathbf{0/40}$	20	$1/20$	$1/20$	$0/20$
20	$2/40$	$3/40$	$1/40$	30	$2/30$	$2/30$	$0/30$
50	$3/40$	$2/40$	$3/40$	40	$\mathbf{1/40}$	$\mathbf{2/40}$	$\mathbf{0/40}$

Where the procedure actually decides. Counting flags on null evaluations says little on its own, because almost no latent is flagged in these null evaluations. We therefore examined the p-values of the candidates the procedure actually selects, over the 40 leave-one-out evaluations, and compared them against the uniform distribution they are treated as. The candidate rule here is the same top-N-per-target rule, at the same N and the same liveness threshold, as the one Table 17 reports; the two analyses share the code path that builds the set.

Two things follow, and they pull in opposite directions. Pooled over all candidates the p-values track the uniform closely at moderate thresholds and are *conservative* in the far left tail, by roughly an order of magnitude at $u = 0.001$. Split by stability they are not uniform at all: among the most stable third, p-values fall below 0.05 about 1.8 times as often as nominal and below 0.01 about 2.2 times as often, while the least stable third is markedly conservative. Standardising each latent by its own mean and standard deviation evidently does not make the tails exchangeable, which is a direct limitation of pooling them into one reference distribution.

Table 15: Empirical $P(p \leq u)$ among selected candidates under null evaluations, where any rejection is false. Left: pooled candidate calibration; right: calibration by latent stability tercile. Results use the reference configuration (20,762 images, $m = 4{,}096$) across 40 evaluations.

Pooled calibration			**Calibration by latent stability**		
u	Observed	Nominal	Stability group	$P(p \leq 0.05)$	$P(p \leq 0.01)$
0.001	0.0001	0.001	Most stable	**0.0882**	**0.0220**
0.010	0.0081	0.010	Middle	0.0498	0.0019
0.050	0.0538	0.050	Least stable	0.0237	0.0004
0.250	0.2617	0.250	*Held out, all*	0.0597	0.0115

This matters because flagged latents concentrate at the stable end (Appendix P), where null-tail inflation is strongest. Benjamini–Hochberg is a step-up procedure, so the threshold a discovery must clear depends on how many are made, and we extract that threshold rather than assume it. Over the five dedup replicates of the reference configuration the procedure rejected 100 to 114 of the $M = 400$ candidates, putting the realised bar rq/M between 0.0125 and 0.0143, mean 0.0130, with the largest rejected p-value in each run between 1.21×10^{-2} and 1.40×10^{-2}: two orders of magnitude from the rank-one bar of 1.25×10^{-4}. At that operating point the pooled calibration is close to nominal (0.0081 observed at $u = 0.01$, 0.0264 at $u = 0.025$), while the most stable tercile is inflated by about a factor of two. The inflation is live where the audit actually decides, and we do not claim otherwise.

The available empirical checks report the following counts: the size-matched placebo flagged nothing in 22 configurations, the leave-one-out diagnostic produced $1/40$, $2/40$ and $0/40$, and ten fresh null retrainings per configuration produced no flags in either of two configurations. These are counts on all-null evaluations, where every null hypothesis is true and every rejection is a false discovery. They are observed frequencies, not an established bound on the error probability, and a setting containing both true and false hypotheses can behave differently.

Held-out retrainings. The diagnostic above reuses the runs that define the reference. We therefore trained ten further same-data nulls per configuration, on fresh seeds and used for nothing else, and tested them against a calibration distribution built from the original 40 nulls alone. The reference dictionary $\mathrm{SAE}^{\mathrm{with}}$ is a separate training from all of them. No run appears on both sides.

Table 16: Calibration on retrainings held out entirely from the audit. Each configuration uses its original 40 nulls for calibration and 10 fresh held-out retrainings for evaluation. Configurations are reported separately; ten runs per configuration provide limited evidence about false-flag probability.

Corpus	**Width** m	**Calibration runs**	**Held-out runs**	**Runs flagged**	**Max flags/run**
20,762	4,096	40	10	**0/10**	**0**
50,762	16,384	40	10	**0/10**	**0**

No flags occurred in either held-out set. We report the configurations separately; these evaluations use fresh retrainings rather than leave-one-out reuse of calibration runs. In the candidate-count sensitivity check, runs with flags reach $3/40$ at $N = 50$ in two configurations. Counts also vary with null-set size, but the few events and overlapping calibration sets limit comparisons. These checks neither establish calibration at $N = 10$ and $R = 40$ nor demonstrate that alternative settings are miscalibrated.

D THE AUDIT CONSTRUCTION

The two flat-ρ rows apply the same construction to a pilot with eight planted targets and to the full design with 42 targets. Because ρ_j is defined as a maximum over groups, adding targets admits more latents: the candidate count increases from 759 to 3,208, which correspondingly lowers the rank-one BH threshold from 6.6×10^{-5} to 1.6×10^{-5}. The full 42-target construction consequently

produces no flagged latents despite using the same trained SAEs, $q = 0.05$, and null protocol. These two rows are therefore not a like-for-like comparison between distinct candidate rules. Rather, they demonstrate that a flat specificity threshold can make detectability depend on the number of planted targets. The per-target construction used in the main audit instead retains the top $N = 10$ candidates per target, yielding 400 tested candidates and 100–114 flagged latents in the reference configuration. Placebo MFR remains zero under all constructions shown.

Table 17: Candidate-construction sensitivity under the same trained SAEs, $q = 0.05$, and null protocol. Placebo MFR remains zero throughout.

Candidate construction	**Candidates**	**Rank-one BH threshold**	**Latents flagged**
Pooled null, gate after FDR	4,085	1.2×10^{-5}	0
Per-latent, flat $\rho \geq 5$, 8 targets	759	6.6×10^{-5}	103
Per-latent, flat $\rho \geq 5$, 42 targets	3,208	1.6×10^{-5}	0
Per-latent, top-10 per target	**400**	$\mathbf{1.25 \times 10^{-4}}$	**100–114**

E FOUR-ARM COMPARISONS ACROSS CONFIGURATIONS

The table below reports the relative displacement ratio for each four-arm configuration. We emphasize doses $k \geq k^*$ because small denominators can destabilize the ratio below the detection threshold. The final column counts targets whose most duplicate-specific latent moves farther under copies-only removal than under its size-matched placebo. We do not attach a binomial p-value because each training run removes all 42 groups jointly; targets are measurements within a shared intervention rather than independent experiments.

Table 18: Copies-only intervention with the original image retained. S is the relative displacement ratio in Eq. 4; $S = 1$ denotes equal placebo-adjusted displacement for copies-only and full removal. Intervals use target-clustered bootstraps; target counts are descriptive because all 42 groups share an intervention.

Images / width	S**, all**	S**,** $k \geq k^*$	**Favour copies**
20,762 / 4,096	0.876 [0.752, 0.980]	$\mathbf{0.914}$ $[\mathbf{0.796}, \mathbf{1.014}]$	$\mathbf{31/42}$
20,762 / 16,384	0.823 [0.712, 0.924]	$\mathbf{0.854}$ $[\mathbf{0.770}, \mathbf{0.939}]$	$\mathbf{35/42}$
50,762 / 4,096	1.006 [0.887, 1.135]	$\mathbf{1.001}$ $[\mathbf{0.903}, \mathbf{1.088}]$	$\mathbf{37/42}$

F RECOVERY: DIRECTION LOSS VERSUS INFORMATION LOSS

The matching distance Δ_j asks whether any single retrained direction resembles a reference latent. A latent can fail this criterion even if its activation is recoverable from several retrained directions. We therefore predict each flagged reference latent's activation $a_j(x)$ from the retrained dictionary's sparse code on its own planted group. Images are split equally into fit and evaluation sets, so no image contributes to both. Predictors are the K retrained latents with the largest absolute fit-half correlation with the response; selection, centring, and scaling use fit-half data only. We fit ridge regression on standardized predictors with fixed penalty $10^{-3}n$ and report out-of-sample R^2. Thus, $K = 1$ tests recovery by a single direction, whereas $K = 20$ permits recovery from a small combination of directions.

The analysis covers the union of 132 flagged reference latents. A latent is eligible when its planted group contains at least four images, each split retains at least eight activation rows, and the reference response varies on those rows; 121 latents satisfy these criteria. Observations are token-level activation rows, but splitting occurs by image. The implementation floors R^2 at -1, although no score reaches this floor, so all reported values are ordinary out-of-sample R^2. Because each planted group consists of one image and its near-duplicate variants, this analysis measures within-group recovery rather than generalisation to unrelated images.

The results show a consistent pattern. Same-data null and placebo retrainings have similar recovery across K, as do copies-only and dedup retrainings. More importantly, the recovery gap between the null and dedup arms decreases from 0.158 at $K = 1$ to 0.064 at $K = 20$. Thus, much of the activation associated with a displaced reference direction is recoverable from a small combination of retrained directions, although a residual gap remains under the tested predictor. We therefore interpret the matching result as evidence about the survival of a *dedicated direction*, not the presence or absence of information. Failure of the restricted recovery model does not establish information loss, since a larger or differently regularised predictor could recover additional signal. All quantities here concern the retrained SAE dictionaries; the backbone remains frozen throughout.

F.1 Reconstruction quality and functional correspondence

We complement the within-group recovery analysis with a held-out functional-matching check on the original corpus draw at widths $4{,}096$ and $16{,}384$. At each width, we train five additional SAEs for each of five conditions: same-data null, full removal, matched placebo, copies-only removal, and its matched placebo. These 50 retrainings provide a complementary check rather than additional independent corpus draws.

Probe and matching procedure. We use 1,024 held-out images, excluding images present in any of the five experimental training corpora. We split the probe into 512 images for matching and 512 for evaluation, with eight cached activation rows per image. For each target, we select its most duplicate-specific reference latent. Eligibility requires positive activation on at least eight matching images and nonzero activation variance. We select the retrained latent with the highest Pearson correlation on the matching split, then evaluate that fixed match on the evaluation split. Undefined evaluation correlations are recorded as missing. This analysis uses neither the empirical null nor FDR correction.

Reconstruction metric. Reconstruction is evaluated in raw activation units on the evaluation split, using each checkpoint's stored scaling factor. For evaluation vectors $x_i \in \mathbb{R}^d$ and reconstructions $\hat{x}_i$, we compute

$$R^2_{\mathrm{rec}} = 1 - \frac{\sum_i \|x_i - \hat{x}_i\|_2^2}{\sum_i \|x_i - \bar{x}\|_2^2}, \qquad \bar{x} = \frac{1}{n}\sum_i x_i.$$

Activation rows receive equal weight, and centring uses the evaluation-set mean separately in each activation dimension. Achieved sparsity is the mean number of strictly positive latent activations per evaluation row.

Results and scope. Held-out reconstruction quality remains similar across the reference dictionaries and additional retrainings ($R^2 = 0.97257$–0.97301), with achieved sparsity of 32 active latents per token. This argues against a gross reconstruction-quality difference in this comparison, but does not establish equivalent convergence or eliminate other measurement explanations. Functional matching is feasible for 11 of 42 targets at width $4{,}096$ and none at width $16{,}384$ (Table 19). Among eligible targets at width $4{,}096$, the descriptive mean correlation is lower after full removal than after its placebo (0.321 versus 0.368). The copies-only summaries do not show a corresponding decrease (0.375 versus 0.365); these arms also contain different missing observations. These descriptive comparisons do not establish statistical significance or equivalence. The absence of eligible targets at width $16{,}384$ reflects insufficient activation coverage on this probe, not zero functional correspondence or evidence of information loss. Consequently, this analysis does not validate the width effect under functional matching. It reinforces the need to distinguish geometric displacement of a dedicated SAE direction from loss of represented information.

G Overlap of the flagged sets

We compare the latents identified by full removal and copies-only removal. For each arm, the union combines flags across its five replicates. To check whether union-level agreement masks inconsistent individual runs, we also compute Jaccard overlap for all 25 dedup-by-copies replicate pairs within each configuration. Because these comparisons reuse the same five trained dictionaries per arm, we report their mean and range descriptively rather than treating them as independent experiments.

Table 19: Reconstruction quality and held-out functional correspondence. Panel A reports reference-dictionary quality and probe eligibility. Panel B reports the average of five retraining-level mean correlations among eligible targets at width 4,096. Higher correlation indicates greater correspondence. These correlations differ from the within-group recovery R^2 reported in the preceding analysis.

A. Reference quality and probe coverage		
Quantity	Width 4,096	Width 16,384
Reconstruction R^2_{rec}	0.972714	0.972925
Active latents per token	32	32
Never-firing latents, full corpus	1/4,096	98/16,384
Eligible targets	11/42	0/42
B. Functional correspondence at width 4,096		
Arm	Mean correlation	Defined pairs
Same-data null	0.419	55/55
Full removal	0.321	55/55
Matched placebo	0.368	55/55
Copies-only removal	0.375	54/55
Copies-only placebo	0.365	54/55

Table 20: Overlap between latents flagged by full removal and copies-only removal. Set counts and Jaccard use the union across five replicates per arm; pairwise values summarize the 25 replicate-pair comparisons.

Images / width	**Dedup**	**Copies**	$\cap$	$\cup$	**Jaccard**	**Pairwise Jaccard**	
						Mean	**Range**
20,762 / 4,096	132	118	110	140	**0.786**	**0.841**	[0.746, 0.888]
20,762 / 16,384	247	241	228	260	**0.877**	**0.872**	[0.846, 0.906]
50,762 / 4,096	28	27	24	31	**0.774**	**0.804**	[0.692, 0.958]

Overlap remains substantial at the individual-replicate level: mean pairwise Jaccard is 0.841, 0.872, and 0.804 across the three configurations, with a minimum of 0.692 over all 75 pairwise comparisons. Thus, the union-level agreement is not solely an artifact of combining flags across replicates. Pairwise overlap is not uniformly higher than union-level overlap, however; each configuration contains individual pairs below its union Jaccard.

H Independent repetitions of the width comparison

We repeat the full width comparison across five independent experiments: the original draw, a fresh draw, and three additional repetitions. Each repetition redraws the 42 targets and near-duplicate crops, rebuilds the activation cache, and independently trains the reference dictionary, 40 same-data nulls, five dedup replicates, and five placebo replicates at each width. All other experimental settings are fixed. The repetition, rather than the target, is therefore the independent unit. Because widths 4,096 and 16,384 are evaluated on the same targets within each repetition, uncertainty in their ratio is estimated by a paired target bootstrap that resamples targets jointly across widths.

Replicated width effect. The ratio

$$\frac{k^*(4{,}096)}{k^*(16{,}384)}$$

exceeds one in all five experiments. Across repetitions, its geometric mean is 4.4, with a descriptive log-scale t interval of $[3.0, 6.4]$. With only five independent repetitions, this interval provides a limited summary of between-experiment variation rather than a precise population estimate. Absolute thresholds are less stable. At width 4,096, k^* ranges from 5.18 to 10.59 across repetitions, and the largest value lies outside the reference experiment's target-bootstrap interval $[3.4, 8.8]$. Thus,

within-experiment intervals do not capture experiment-to-experiment variation. We therefore emphasize the replicated direction of the paired width effect rather than a universal value of k^*.

H.1 Sensitivity of the width comparison

Candidate count and continuous displacement. Using the existing trained dictionaries from all five paired experiments, we repeat the audit with $N \in \{5, 10, 20, 50\}$ candidates per target while holding the other audit settings fixed. For $N = 5, 10, 20$, the observed fitted threshold ratio $k^*(4{,}096)/k^*(16{,}384)$ exceeds one in every experiment (Table 21). Its magnitude nevertheless depends on candidate count. At $N = 50$, the narrower dictionary does not reach 50% detection within the tested dose range in any experiment, so we do not report a finite threshold ratio. In four experiments, all 4,000 narrow-dictionary bootstrap samples have constant outcomes. In Rep 2, 1,457 have constant outcomes, 2,541 receive the nonpositive-slope classification, and two receive the outside-range classification. Numerical ratios returned in this regime are therefore not interpreted as evidence for a reversed width effect. A complementary analysis uses each target's fixed reference-selected latent and its mean removal-minus-placebo displacement, without empirical-null calibration or BH. The wider dictionary has greater displacement at $k = 2$ and $k = 4$ in all five experiments. This ordering is not uniform across doses: at $k = 32$, displacement is greater in the narrower dictionary in all five experiments. The result therefore supports a shift toward lower-dose sensitivity, rather than uniformly larger displacement at every dose. These comparisons retain reference-latent selection and decoder matching.

Table 21: Observed fitted width ratios under candidate-count sensitivity. Ratios are $k^*(4{,}096)/k^*(16{,}384)$. These are descriptive point estimates, not uncertainty intervals. At $N = 50$, the narrower dictionary does not reach 50% detection within the tested doses, so a finite ratio is not reported.

Experiment	$N = 5$	$N = 10$	$N = 20$
Original	3.19	3.30	6.21
Fresh	5.53	5.89	6.52
Rep 1	2.54	3.54	4.79
Rep 2	2.91	3.65	5.29
Rep 3	5.45	6.32	7.25

Sensitivity to logistic estimation. We retrospectively refit the saved detection outcomes using Firth-penalized logistic regression (Kosmidis & Firth, 2021). At $N = 10$, the five paired width ratios remain close to the original estimates (Table 22). All 4,000 paired target-bootstrap draws per experiment yield finite ratios above one. Fits satisfy the implemented numerical criterion of a maximum absolute penalized-score component below 10^{-5}, although optimizer success flags are not uniformly positive. Some fitted crossings remain outside the tested dose range and are retained in the reported summaries. These descriptive bootstrap ranges are conditional on the trained dictionaries; their coverage is not established.

Table 22: Retrospective Firth sensitivity at $N = 10$. Ranges give the 2.5th and 97.5th percentiles of all 4,000 paired target-bootstrap ratios, including extrapolated crossings. In-range counts require both fitted thresholds to lie within the tested dose range.

Experiment	Ratio	Bootstrap range	Both in range
Original	3.31	[1.75, 6.09]	3,989/4,000
Fresh	5.91	[3.73, 10.91]	3,617/4,000
Rep 1	3.54	[1.98, 6.94]	3,779/4,000
Rep 2	3.65	[2.20, 6.25]	3,998/4,000
Rep 3	6.32	[4.06, 10.00]	3,996/4,000

Across $N = 5, 10, 20$, all fitted paired ratios exceed one. At $N = 50$, narrow-width outcomes are constant in four experiments and the fitted slope is nonpositive in the remaining experiment, so an increasing detection-dose ratio is not reported. The refit supports stability of the paired width direction under a different estimator, but does not establish detector calibration.

Discrete dose-crossing sensitivity.

To assess dependence on the logistic curve, we fit monotone isotonic detection rates over the tested doses and bracket the 50% crossing between adjacent doses. Crossings outside the tested range remain censored rather than receiving extrapolated point estimates. At $N = 10$, the wider dictionary crosses earlier in every experiment (Table 23). We retain all 4,000 paired target-bootstrap draws, including censored crossings. The outer percentile envelopes are descriptive summaries, not confidence intervals with established coverage; some include one or have no finite upper bound. This supports the observed direction with substantial uncertainty, but does not replace the logistic estimator or validate its bootstrap intervals.

Table 23: Discrete 50% detection crossings at $N = 10$. Brackets are in extra copies. Ratio envelopes describe the narrow-to-wide crossing ratio; ∞ denotes an unbounded upper endpoint.

Experiment	Width 4,096	Width 16,384	Ratio envelope
Original	$[4, 8]$	$[1, 2]$	$[1, 16]$
Fresh	$[8, 16]$	$[1, 2]$	$[2, \infty)$
Rep 1	$[4, 8]$	$[1, 2]$	$[0.5, \infty)$
Rep 2	$[8, 16]$	$[2, 4]$	$[1, 16]$
Rep 3	$[8, 16]$	$[1, 2]$	$[2, 32]$

I CORPUS SIZE UNDER TWO TRAINING PROTOCOLS

Corpus size is not independently identified by either training protocol. With 40 epochs fixed, larger corpora receive more gradient updates; with a fixed 12,000-step budget, the largest corpus receives fewer effective passes through the data and produces no detection. We therefore report both protocols as sensitivity analyses rather than estimate a corpus-size scaling law. At 50,762 images under the matched 12,000-step protocol, no latent is flagged at any tested dose, so a finite k^* cannot be estimated.

Table 24: Detection thresholds across corpus sizes under fixed epochs and a matched 12,000-step training budget. The protocols introduce different training confounds and therefore do not isolate a corpus-size effect.

Corpus	**k^*, 40 epochs**	**Steps**	**k^*, 12k steps**	**MFR, 12k steps**
8,762	6.3	4,160	3.4	0.0450
20,762	5.5	9,920	5.9	0.0255
35,762	13.8	17,080	15.1	0.0145
50,762	23.4	24,280	*Not estimable*	**0.0000**

J FOUND DUPLICATES

Naturally occurring Food-101 near-duplicates produce a weaker signal than controlled near-exact copies. At width 16,384, the dedup-minus-placebo displacement is positive at both matched doses, but the controlled effect is $3.1\times$ larger at $k = 2$ and $4.2\times$ larger at $k = 4$. The targeted win rate is 0.859 for planted copies and 0.572 for natural clusters. No latent is flagged for the natural clusters under the tested configuration. These results limit generalisation to natural duplication: the available clusters have limited doses and weaker displacement than controlled copies. Non-detection does not establish the absence of an effect.

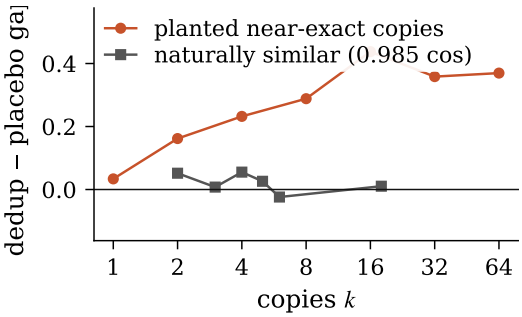


Figure 2: Dedup-minus-placebo displacement by duplication dose for controlled near-exact copies and naturally occurring Food-101 clusters at cosine similarity ≥ 0.985.

Natural groups are connected components of a cosine-similarity graph over normalized cached CLS activations, excluding planted groups from the search. We retain the 42 largest eligible components, with eligible sizes of 2–25 images. The threshold 0.985 applies to edges, not every pair within a component. For natural groups, $k = |g| - 1$, and full removal deletes the entire group.

K MEMBERSHIP INFERENCE IN DETAIL

Threat model. We test whether flagged SAE latents carry information about membership in the SAE training corpus. Given an image and the trained SAE, the score is the strongest firing among flagged latents, normalized by each latent's typical corpus firing. The flagged set is identified using knowledge of the planted duplicate groups, so this measures membership information in discovered latents rather than an unaided membership-inference attack. The backbone is frozen and never trained on the planted duplicates.

Membership signal. The normalized flagged-latent score achieves higher AUC and TPR at 1% FPR than reconstruction error, but lower TPR at 0.1% FPR (Table 25). This reversal motivates reporting operating points alongside AUC (Carlini et al., 2022a). Non-duplicated members yield AUC 0.506, while scoring over all latents yields 0.467. Non-members are ordinary Food-101 images rather than hard lookalikes, and the aggregate is not dose-balanced.

Table 25: Membership inference in the reference configuration using 132 flagged latents pooled across five dedup replicates and 20,000 non-members. Operating points report TPR at fixed false-positive rates.

Score	AUC	TPR @ 1% FPR	TPR @ 0.1% FPR
Flagged latents, normalized	**0.938**	**0.669**	0.063
Flagged latents, raw	0.932	0.373	0.012
Reconstruction error	0.902	0.609	**0.398**
All latents	0.467	0.014	0.001
Non-duplicated members	0.506	0.011	0.001

L MULTIPLE TESTING: DERIVATIONS

Empirical resolution and the rank-one BH threshold. With R null retrainings and m_{live} eligible reference latents in the pool, the minimum attainable empirical p-value is $1/(1+Rm_{\text{live}})$. Requiring this floor to clear the rank-one Benjamini–Hochberg (BH) threshold gives

$$\frac{1}{1 + Rm_{\text{live}}} \leq \frac{q}{M} \qquad \Longleftrightarrow \qquad R \geq \frac{M/q - 1}{m_{\text{live}}}. \tag{5}$$

This is only a rank-one resolution condition, not a necessary condition for BH rejection. Under the step-up rule, multiple sufficiently small p-values can satisfy a higher-rank threshold even when none clears q/M. For example, with $M = 100$ and $q = 0.05$, two p-values of 8×10^{-4} satisfy the rank-two threshold $2q/M = 10^{-3}$, although neither satisfies the rank-one threshold $q/M = 5 \times 10^{-4}$.

If all m latents are eligible and tested ($m_{\text{live}} = M = m$), Eq. 5 becomes

$$R \geq \frac{1}{q} - \frac{1}{m}.$$

For $q = 0.05$ and $m = 4{,}096$, the smallest integer satisfying this condition is $R = 20$. Restricting the test to $M = 400$ candidates reduces the corresponding requirement to $R \geq 2$. Our audit uses the full BH step-up procedure; the choice $R = 40$ is motivated by estimation of each latent's null variability rather than by this resolution bound.

Effect of candidate restriction. Testing all $m = 4{,}096$ latents gives a rank-one BH threshold of approximately 1.2×10^{-5}. For illustration, restricting testing to $M = 400$ candidates increases

this threshold to 1.25×10^{-4}. Candidate restriction therefore materially changes detectability by reducing the multiple-testing burden, which is why candidate construction is evaluated as part of the audit rather than treated as innocuous preprocessing.

M THE AUDIT, END TO END

Algorithm 1 summarizes the audit for one configuration. All arms use the same reference dictionary, candidate set, and same-data calibration ensemble; each intervention dictionary is trained from scratch. Flags and detection rates are procedure-dependent summaries, interpreted alongside placebo-adjusted displacement and the calibration diagnostics in Appendix C.

Algorithm 1 Instability-controlled duplication audit

Input: frozen backbone f, corpus D, groups $\{g_1, \ldots, g_G\}$, width m, level q, retrainings R, candidates per target N
Cache $f(x)$ once for all $x \in D$ ▷ the backbone is never retrained
Train $\mathrm{SAE}^{\mathrm{with}}$ on D; record ρ_{jg} from Eq. 1
for $r = 1, \ldots, R$ **do** ▷ the null: same data, different seed
    Train $\mathrm{SAE}^{(r)}$ on D; set $\Delta_j^{(r)} \leftarrow 1 - \max_\ell \cos(d_j^{\mathrm{with}}, d_\ell^{(r)})$
end for
Drop latents firing fewer than ten times across the corpus ▷ liveness filter
$\mu_j, \sigma_j \leftarrow$ mean and sd of $\{\Delta_j^{(r)}\}_{r=1}^R$; build Z^0 leave-one-out (Eq. 3) ▷ each latent standardised by its own σ_j
$C \leftarrow \bigcup_g \{\text{top-}N \text{ latents by } \rho_{jg} \text{ with } \rho_{jg} \geq 2\}$ ▷ candidate filter, per target
$\mathrm{own}(j) \leftarrow \arg\max_g \rho_{jg}$ ▷ one owning group per latent, for the catch rate
for each replicate of each arm: dedup, placebo, and (three configurations only) copies, placebo-c **do**
    Retrain from scratch on $D \setminus$ (removed set); compute Δ_j (Eq. 2) and z_j (Eq. 3)
    $\mathrm{flag}_j \leftarrow \mathbb{1}[j \in C] \cdot \mathbb{1}[\mathrm{BH}(p_j; q) \text{ rejects over } C]$
    $\mathrm{MFR} \leftarrow \frac{1}{m} \sum_j \mathrm{flag}_j$
end for
Target g counts as detected when at least one flagged latent has $\mathrm{own}(j) = g$
Return MFR on dedup and on placebo, and per-target catch rates

N COST MODEL

The audit requires $R + AR_{\mathrm{rep}} + 1$ SAE trainings per configuration, where R is the number of same-data null retrainings, A is the number of removal arms, and R_{rep} is the number of replicates per arm. With $R = 40$ and $R_{\mathrm{rep}} = 5$, the two-arm configurations ($A = 2$) require 51 trainings each, while the three configurations with copies and placebo-c arms ($A = 4$) require 61 trainings each. The backbone forward pass is amortised because activations are cached once and reused by all SAE trainings. At $m = 4{,}096$ with approximately 10^6 cached activation vectors, one SAE training takes roughly 2.5 GPU-minutes on a V100, making a 51-training configuration approximately two GPU-hours. At $m = 65{,}536$, one training takes roughly 2.5 GPU-hours. The null ensemble dominates the audit cost, but reducing R also reduces empirical p-value resolution and the precision of the per-latent instability estimates. The reference, null, and intervention trainings in the stated 22-configuration design comprise

$$19 \times 51 + 3 \times 61 = 1{,}152$$

SAE trainings. This is a subtotal for that design, rather than a reconciled total for all analyses.

O PREDICTION CHECKS

We summarize six prediction checks and their evaluation criteria. Three predictions were supported and three were falsified. The falsified predictions were: **(i)** a fixed corpus-share account motivated

by Henighan et al. (2023), which predicted $k^* = 2.2$ at 8,762 images but yielded 6.3, exceeding the evaluation threshold of 4; **(ii)** corpus size rather than dictionary width as the primary driver of the width effect, which predicted $k^* \approx 19$ at width 16,384 but yielded 2.0; and **(iii)** greater susceptibility on the visually simpler EuroSAT dataset, which predicted a threshold below Food-101's 6.3 but yielded 6.7. The supported predictions were: **(i)** $k^* = 6.2$ at width 8,192, with 6.5 observed; **(ii)** high single-copy detection at width 65,536, with $P(\text{detect} \mid k = 1) = 0.833$ against an evaluation threshold of 0.6; and **(iii)** a targeted win rate in $[0.45, 0.60]$ for naturally occurring duplicates at width 16,384, with 0.572 observed. These checks are consistent with the observed empirical regularity while rejecting several proposed explanations for it.

P Additional diagnostics

Configuration sensitivity. Depth produces the largest descriptive change outside dictionary width: k^* decreases from 24.0 $[13.6, 63.2]$ at block 4 to 4.5 $[3.1, 6.4]$ at block 11. Because the backbone is frozen and never trained on the duplicates, this trend does not establish encoder memorization or a shared mechanism with prior depth effects. Brackets are numerical target-bootstrap intervals conditional on the trained dictionaries; their coverage is not established. The width trend is nonlinear. On the 50,762-image corpus, descriptive log-log slopes are -1.850 and -1.698 over the lower-width intervals and approximately -0.862 from 16,384 to 32,768. These are local descriptive slopes rather than a universal scaling exponent.

Could the width trend arise from the null?

At width 65,536, 1,153 of 65,536 latents never fire. More generally, nearest-direction matching could change its null behaviour as dictionary width increases. The targeted win rate provides a complementary diagnostic that uses neither the empirical null nor FDR correction: on the same 50,762-image corpus it rises from 0.756 at width 4,096 to 0.969 at 65,536. This does not rule out every measurement explanation, but the width trend is therefore not confined to the null-based detection statistic.

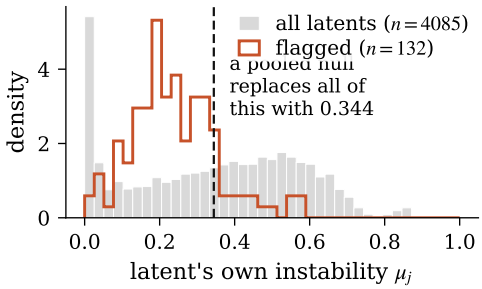


Figure 3: Per-latent retraining instability. Latents differ substantially in their null variability, motivating latent-specific standardization rather than a single pooled raw null.

At TopK 64, the mean removal-minus-placebo gap over $k \geq 8$ falls to 0.67× its TopK 32 value, while mean own-null instability rises to 1.67×. The gap-to-instability ratio consequently falls from 1.82 to 0.73. No latent is flagged, although the targeted win rate remains 0.704. Thus, non-detection does not imply disappearance of the measured displacement.

Why condition on latent stability? Per-latent retraining instability spans a wide range. A pooled raw null would therefore compare naturally stable and unstable directions against the same reference distribution. The flagged directions are concentrated toward the more stable part of this distribution, where losing latent-specific conditioning is particularly consequential. This motivates the standardized per-latent null used by the audit.

The targeted win rate asks whether each target's most duplicate-specific latent moves farther under dedup than under its size-matched placebo. It uses no empirical null or FDR correction. Because targets within a training run share the same intervention, the reported proportions are descriptive and are not assigned binomial p-values. For width 2,048, no tested dose yields a fitted detection threshold, so k^* is reported as not estimable. At width 65,536, detection occurs in 25 of 30 target-by-replicate outcomes at $k = 1$, across six target images. We report this observed frequency, 0.833, rather than extrapolate a 50% detection dose below the tested range. At TopK 64, no latent is flagged despite a positive dedup-minus-placebo displacement.

Taken together, these diagnostics distinguish measured displacement from the audit's ability to flag it. Latent-specific standardisation accounts for differences in baseline variability, but does not guar-

antee comparable null tails across latents or widths. The targeted win rate avoids null calibration while retaining reference-latent selection and decoder matching. Agreement between these summaries therefore provides complementary evidence, rather than independent validation of the measurement procedure.

Table 26: Detection thresholds across configuration axes. Within each block, only the named axis changes and the corpus is fixed. Brackets give central 95% numerical target-bootstrap ranges conditional on trained dictionaries; coverage is not established.

Axis	Setting	k^*	Bootstrap range	Detection rate at $k = 1$
Width	2,048		*Not estimable*	
	4,096	23.4	$[17.0, 33.7]$	0.000
	8,192	6.5	$[4.5, 9.1]$	0.033
	16,384	2.0	$[1.1, 3.3]$	0.267
	32,768	1.1	$[0.4, 1.7]$	0.367
	65,536	*Not reported*	–	**0.833**
Block	4	24.0	$[13.6, 63.2]$	0.100
	8	5.5	$[3.4, 8.8]$	0.267
	11	4.5	$[3.1, 6.4]$	0.100
TopK	16	6.0	$[4.1, 8.5]$	0.100
	32	5.5	$[3.4, 8.8]$	0.267
	64	*No flags; positive gap*		
Backbone	CLIP ViT-B/16	5.5	$[3.4, 8.8]$	0.267
	DINO ViT-B/16	5.4	$[4.0, 7.1]$	0.100
Dataset	Food-101	6.3	$[4.7, 8.4]$	0.100
	EuroSAT	6.7	$[4.3, 11.4]$	0.200

Table 27: Targeted win rates across tested configurations. Where available, intervals are descriptive 95% target-bootstrap percentile intervals conditional on the trained dictionaries.

Configuration	Win rate	Bootstrap interval
CLIP, 20,762 / 4,096	0.838	$[0.771, 0.901]$
CLIP, 50,762 / 4,096	0.756	$[0.683, 0.830]$
CLIP, 50,762 / 16,384	0.882	$[0.806, 0.950]$
CLIP, 50,762 / 65,536	**0.969**	$[0.936, 0.994]$
CLIP, 20,762 / 16,384	0.859	$[0.777, 0.933]$
DINO, 20,762 / 4,096	0.816	$[0.741, 0.887]$
EuroSAT, 8,762 / 4,096	0.790	$[0.709, 0.865]$
Block 4, 20,762 / 4,096	0.709	$[0.635, 0.780]$
Block 11, 20,762 / 4,096	0.850	$[0.772, 0.919]$
TopK 16, 20,762 / 4,096	0.785	$[0.701, 0.866]$
TopK 64, 20,762 / 4,096	0.704	$[0.610, 0.794]$
Natural duplicates, $m = 4{,}096$	0.516	$[0.461, 0.570]$
Natural duplicates, $m = 16{,}384$	0.572	–