Topographic Training Concentrates Causal Circuits Without Improving Neuron Monosemanticity

Gautam Ranka [1] **Shubham Pandere*** [1] **Aiden Ross D'souza*** [1]

Abstract

Mechanistic interpretability of vision transformers seeks to decompose model computation into human-readable units, but learned representations entangle many concepts in each neuron. Feature superposition is widely treated as the central obstacle to this decomposition, yet most mitigations (sparse autoencoders, dictionary learning) are post-hoc and leave the underlying network unchanged. We ask whether a spatial-locality training loss (TopoLoss) can act as a lightweight, training-time prior that improves interpretability of standard mech-interp tools. Training ViT on ImageNet-100 across multiple TopoLoss weights α, we measure causal sufficiency of topographic clusters via activation patching and feature geometry via sparse autoencoders fit to the same residual stream. At $\alpha = 1.0$, topographic clusters are $2.79\times$ more causally sufficient than random unit sets of the same size, with the effect increasing monotonically in α. SAE L0 sparsity decreases by 11% and dead-feature fraction rises 19-fold, yet standard neuron-level monosemanticity scores are unchanged, indicating that topographic pressure acts at circuit level, concentrating causal mass into spatially local structures without disentangling individual neurons. This dissociation suggests current neuron-level monosemanticity metrics are insensitive to a class of real interpretability gains, and positions cheap architectural priors as a viable training-time complement to post-hoc tooling.

[1]IvLabs, VNIT Nagpur, India. Correspondence to: Gautam Ranka <gautam.ranka@ivlabs.in>, Shubham Pandere <shubham.pandere@ivlabs.in>, Aiden Ross D'souza <aiden.dsouza@ivlabs.in>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

The superposition hypothesis (Elhage et al., 2022) posits that neural networks represent more features than they have neurons, encoding multiple unrelated concepts in each unit through approximate orthogonality in high-dimensional space. Superposition makes circuit analysis and in general, mechanistic analysis hard: patching a single neuron contaminates unrelated features, and sparse autoencoders (SAEs) must expend capacity to decompose polysemantic units into monosemantic directions (Bricken et al., 2023; Cunningham et al., 2023). Current solutions are almost entirely *post-hoc* where they learn to decode a fixed representation rather than encouraging the model to form a cleaner one.

A complementary strategy is to impose an inductive bias during training that couples nearby neurons, hoping spatial locality produces a form of *functional organisation* analogous to cortical maps in biological vision (Khosla et al., 2025). Khosla et al. recently showed that a Laplacian weight-smoothness penalty (TopoLoss) applied to vision and language transformers improves brain-similarity scores without sacrificing accuracy. The key open question is whether this reorganisation also improves *mechanistic interpretability metrics*.

We conduct a systematic audit of TopoLoss across three interpretability axes:

- **H1 (monosemanticity):** Do individual neurons become more selective? Measured via entropy-based selectivity M_u.
- **H2 (superposition):** Does the SAE-measured feature complexity decrease? Measured via L0 norm and dead-feature fraction.
- **H3 (causal purity):** Are topographic clusters more causally sufficient for class prediction than random unit sets? Measured via activation-patching logit-delta ratios.

Our main findings are: H3 is strongly supported ($p<0.0001$, $d = 0.95$); H2 is supported at high α with a confound of accuracy; H1 is null. We further observe that H3 effects are more modest in a smaller 4-layer architecture (TinyViT), consistent with a scale or depth requirement for circuit formation, though additional experiments are needed to isolate

this factor.

2. Background and Method

TopoLoss. Following Khosla et al. (2025), we add a weight-smoothness penalty to the standard cross-entropy loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{topo}}, \qquad \mathcal{L}_{\text{topo}} = \alpha \sum_{(i,j)\in\mathcal{E}} \|w_i - w_j\|^2, \quad (1)$$

where $\mathcal{E}$ is the set of adjacent unit pairs on a fixed 2D topographic grid and $\alpha \geq 0$ controls pressure strength. In our ViT-S implementation, TopoLoss is applied to the attention-projection weights (`attn.proj`) of all 12 blocks; units are arranged on a $\sqrt{d} \times \sqrt{d}$ grid matching the head dimension.

Models and training. We train ViT-S/16 (Dosovitskiy et al., 2021) on ImageNet-100 (Deng et al., 2009) (126,688 train / 5,000 val, 100 classes) for $\alpha \in \{0.0, 0.1, 1.0\}$, with three seeds $\{42, 123, 456\}$ each, total of nine checkpoints. A broader α-sweep ($\alpha \in \{0, 0.01, 0.1, 1.0\}$, seed 42) provides dose-response curves. We additionally train a lightweight TinyViT (4 layers, $d = 128$, 4 heads) with the same protocol as a preliminary scale probe. All runs use AdamW, cosine schedule, and standard augmentation; full hyperparameters are in Appendix A.

Evaluation metrics. *H1 : Monosemanticity* M_u is the normalised negative entropy of a neuron's class-activation distribution over 5,000 validation images, $M_u = (H_{\max} - H)/H_{\max}$, so $M_u = 1$ is perfectly selective. We also train an SAE (TopK, $k = 32$, 2048 features) on layer-6 residuals and measure per-feature selectivity.

H2 : SAE L0 sparsity. We train one SAE per checkpoint at layer 6 (`attn.proj` output) and report the mean number of active features per token (L0) and the fraction of dead features (active on <0.1% of tokens). Lower L0 and higher dead-feature fraction indicate sparser, more specialised representations. The choice of layer 6 was motivated by it lying in the middle of the effect band (Appendix E)

H3 : Causal purity. For each of 20 randomly selected ImageNet-100 classes and each of the 3 seeds, we identify the top-$k = 16$ neurons most activated by that class, compute the mean absolute logit delta when patching this cluster from a source (target-class image) into a target (other-class image), and normalise by the delta from patching a random size-k set. Ratios > 1 mean topographic clusters carry disproportionate causal weight. The top-k selection rule is by activation magnitude and does not enforce spatial contiguity; we verify post-hoc (Table 2) that these sets are nonetheless spatially co-localised under TopoLoss, justifying the cluster framing.

Table 1. **H3 patching ratios** (cluster / random |Δlogit|), multi-class $n = 60$.

α	Mean ratio	Δ vs base	p	d
0.0	1.68 ± 0.42	—	—	—
0.1	2.27 ± 0.93	$+0.59$	< 0.0001	0.61
1.0	2.79 ± 1.24	$+1.12$	< 0.0001	0.95

Table 2. **Spatial coherence of top-k=16 class units** (mean pairwise grid distance on the 16×24 sheet; lower = more spatially concentrated; random baseline = 10.618). $n = 60$ (20 classes $\times$ 3 seeds).

α	Mean dist	Δ% vs base	p	d
0.0	10.599 ± 0.892	—	—	—
0.1	10.525 ± 0.934	-0.7%	0.661 (n.s.)	0.08
1.0	$\mathbf{9.948 \pm 1.063}$	$\mathbf{-6.1\%}$	**0.0005**	**0.66**

3. Results

3.1. H3: Topographic Clusters Are Causally Sufficient

Table 1 and Figure 2 show the primary result. Topographic clusters at $\alpha = 1.0$ produce a mean patching ratio of 2.79 ± 1.24, versus 1.68 ± 0.42 for the baseline leading to a 66% increase. The effect is monotone in α ($1.68 \rightarrow 2.27 \rightarrow 2.79$) and robust across all 20 classes and all 3 seeds independently (per-seed $\alpha = 1.0$ ratios: 2.53, 2.89, 2.97). A paired t-test on seed-level means yields $p = 0.082$ ($n = 3$), which is underpowered; the multi-class expansion to $n = 60$ gives $p < 0.0001$, $d = 0.95$ (Table 1).

Spatial coherence of top-k unit sets. To verify that the H3 selection rule picks out spatially co-localised units under TopoLoss (and not merely high-activation units distributed arbitrarily across the grid), we compute the mean pairwise Euclidean grid distance among the top-$k = 16$ class units on the 16×24 topographic sheet. Lower values indicate tighter spatial concentration; the random-baseline expected distance for a size-16 subset is 10.618. Results across 20 classes $\times$ 3 seeds ($n = 60$) are shown in Table 2. Top-k units at $\alpha = 1.0$ are 6.1% more spatially concentrated than baseline ($p = 0.0005$, $d = 0.66$); the $\alpha = 0.1$ effect is not significant, mirroring the H3 threshold pattern.

3.2. H2: Reduced Feature Superposition

At $\alpha = 1.0$, SAE L0 decreases by 11% relative to baseline (1133.8 ± 3.2 vs 1273.9 ± 15.0; $p = 0.007$, $d = 7.06$) and the dead-feature fraction rises 19-fold ($3.8 \pm 0.8\%$ vs $0.2 \pm 0.1\%$). The $\alpha = 0.1$ model shows neither effect reliably ($p = 0.054$ for L0, $p = 0.580$ for dead fraction), confirming a threshold between $\alpha = 0.1$ and $\alpha = 1.0$. The L0 reduction is distributed across depth (layers 3–10, peak at layer 10; see Appendix E), not localised to a single block.

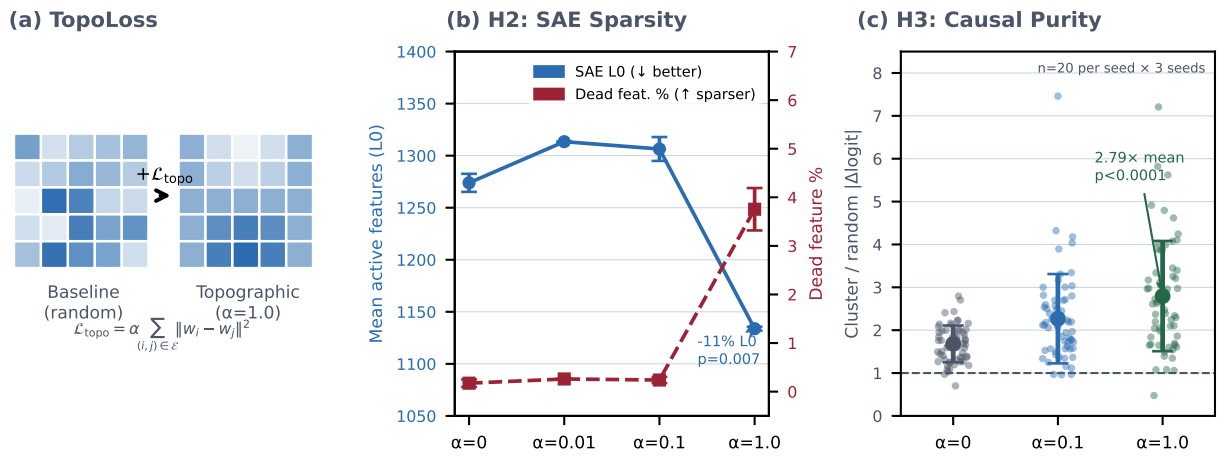

Figure 1. **Overview.** *(a)* TopoLoss (Eq. 1) penalises weight differences between spatially adjacent neurons, pushing functionally similar units to co-localise on the topographic grid. *(b)* H2: SAE L0 (mean active features, blue) decreases monotonically with α, while dead-feature fraction (red, dashed) rises 19-fold at $\alpha = 1.0$ (y-axis truncated; baseline L0 = 1273.9). *(c)* H3: per-class causal patching ratios($n = 60$, 20 classes $\times$ 3 seeds). Points show individual class$\times$seed pairs; filled markers show mean $\pm$ SD. Topographic clusters at $\alpha = 1.0$ are $2.79\times$ more causally sufficient than random unit sets ($p < 0.0001$, $d = 0.95$). The dashed line marks the random-cluster baseline (ratio $= 1$).

Confound. We treat the L0 reduction as suggestive but not confound-free. The $\alpha = 1.0$ model suffers a -1.6 pp accuracy cost ($61.2\% \rightarrow 59.6\%$), and an OLS regression of L0 on validation accuracy across all 9 ViT-S checkpoints yields $R^2 = 0.72$ ($p = 0.004$), indicating accuracy explains substantial L0 variance. Two analyses suggest the effect is not entirely accuracy-driven. First, the PCA d_{95} dissociation at $\alpha = 0.1$, where SAE L0 *increases* ($+32.5$) while d_{95} *decreases* (-8.3), shows SAE L0 captures structure beyond raw effective dimensionality. Second, two of three $\alpha = 1.0$ seeds lie below the accuracy-L0 regression line (sparser than accuracy alone predicts). An accuracy-matched ablation would directly resolve the confound and is left for future work.

3.3. H1: No Change in Neuron Monosemanticity

Entropy-based M_u is flat across α (0.234 ± 0.002 at $\alpha = 0$ vs 0.236 ± 0.002 at $\alpha = 1.0$; $\Delta = 0.0019$, $n.s.$). SAE-feature-level M_u is also null. Full tables are in Appendix B. The H1 null alongside the H3 positive indicates that topographic pressure reorganises the circuit-level feature geometry without altering the per-neuron class-selectivity distribution.

3.4. Scale Dependence: TinyViT

As a preliminary probe, we trained TinyViT (4-layer, $d = 128$) with the same protocol. The H2 trend replicates: L0 decreases monotonically with α (-2.1% at $\alpha = 1.0$), consistent with the weight-norm scaling rule $\alpha_{\text{eff}} = \alpha \cdot (\text{rms}_{\text{ref}}/\text{rms}_{\text{tgt}})^2$ (see Appendix F). H3, however, saturates: the maximum ratio across all α is $1.42\times$ ($\alpha = 0.1$), far below the ViT-S result of $2.79\times$. We caution that TinyViT and ViT-S differ simultaneously in depth, width, accuracy, and training budget; with a single seed per α we cannot isolate the responsible factor. Depth is one plausible hypothesis as a 4-layer network may lack sufficient representational stages for topographic pressure to form coherent causal circuits, but this requires a controlled depth-sweep to confirm. We treat the TinyViT result as a useful negative: the simple weight-norm calibration rule predicts H2 transfer (which we observe) but not H3 transfer (which we do not), suggesting H3 depends on architectural factors beyond effective topographic pressure.

4. Discussion

Why H3 but not H1? Monosemanticity (M_u) measures the distribution of a single neuron over classes, a marginal statistic insensitive to the *joint* arrangement of multiple neurons. Activation patching measures whether a specific set of neurons is jointly sufficient for a prediction, a multivariate, *circuit-level* probe. Our results suggest that TopoLoss reorganises neurons into functionally coherent spatial clusters (H3) without changing how selectivity is spread across individual neurons (H1). This is consistent with Elhage et al. (2022): superposition redistributes information across neurons, so spatial co-location (not per-neuron selectivity) is the relevant unit of circuit purity.

TopoLoss as a training-time interpretability prior. Standard mech-interp interventions are post-hoc: SAEs learn

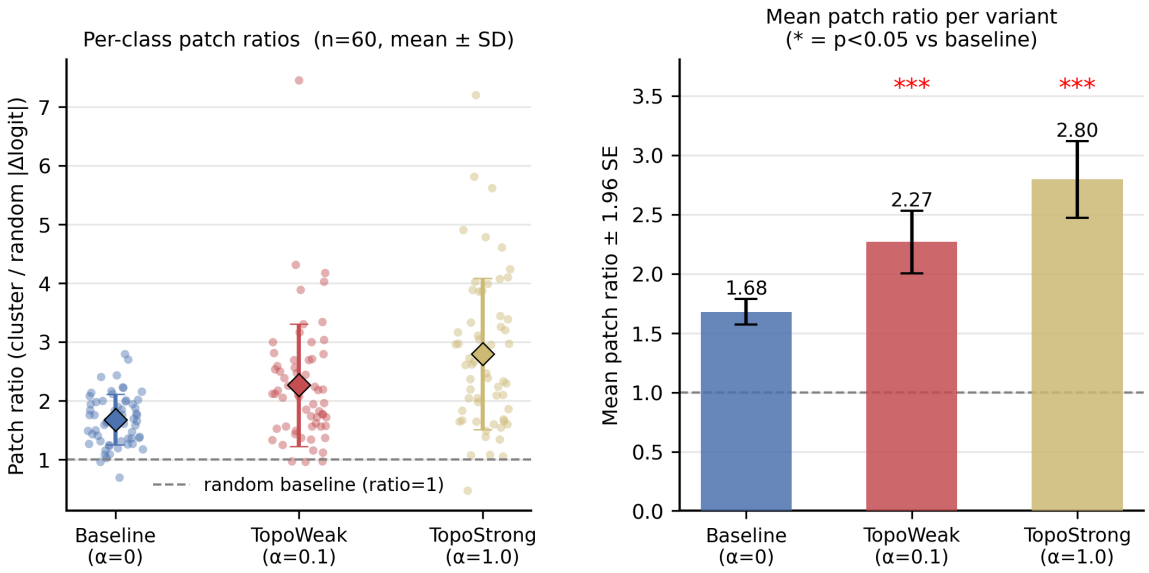

Figure 2. **H3: Causal patching ratios across 20 classes $\times$ 3 seeds ($n = 60$).** Each point is the mean cluster/random logit-delta ratio for one class$\times$seed combination. The ratio at $\alpha = 1.0$ (2.79 ± 1.24) is significantly above baseline (1.68 ± 0.42), and above $\alpha = 0.1$ (2.27 ± 0.93). Both comparisons: $p < 0.0001$, Cohen's $d \geq 0.61$.

to decode a representation after training; patching reveals circuits but does not change them. Our results show that a <0.1k-parameter regularisation term, added to the training loss, systematically shifts circuit geometry in a way measurable by existing tools. The effect is modest in cost (-1.6 pp accuracy at $\alpha = 1.0$) and strong in signal ($d = 0.95$ for H3).

Related work. Khosla et al. (2025) demonstrated that TopoLoss improves brain similarity without accuracy loss; we evaluate it on mechanistic interpretability axes for the first time. Bricken et al. (2023) and Cunningham et al. (2023) use SAEs as diagnostic tools; we show that TopoLoss *shifts* the SAE's diagnosis. Conmy et al. (2023) use activation patching to localise circuits within a single model; we use the same protocol to *compare* models trained with different priors. Topographic constraints in deep networks have a longer lineage (Lee et al., 2020; Doshi & Konkle, 2023; Margalit et al., 2024); we focus on TopoLoss as a recent, lightweight instantiation amenable to standard mech-interp tooling. Work on topographic self-organisation in networks (Zeiler & Fergus, 2014; Olah et al., 2020) has noted functional clustering informally; we quantify it with controlled patching experiments.

Limitations.

- Single dataset (ImageNet-100).
- H2 has a partial accuracy confound, addressed but not fully resolved.
- Explored only on discriminatory vision transformer models.
- The M_u metric uses ImageNet-100 classes as the concept basis, which may be too coarse to detect monosemanticity in features that cut across classes (e.g. textures, parts). The 25k-image SAE-feature M_u evaluation (Appendix B) addresses the sample-size concern but inherits the same class-basis limitation.

Future work.

- A depth-controlled sweep (TinyViT $\rightarrow$ ViT-S $\rightarrow$ ViT-B at fixed width) would test the scale hypothesis.
- Extension to other conventional vision architectures like CNNs and ResNets.
- An embedding-based monosemanticity score that sidesteps the class basis Pach et al. (2025), would provide an independent test of whether the null holds, applied either to raw neurons or to the SAE-recovered features.
- Brain-similarity alignment (NSD fMRI), SOM post-hoc control, and language-model replication are natural extensions.

Impact Statement

This paper presents work whose goal is to advance the field of mechanistic interpretability by studying training-time interventions that reduce feature superposition. Improved interpretability tools may help identify and mitigate harmful capabilities in AI systems. There are no direct negative societal consequences we foresee from this methodological

contribution.

5. Code

The code repository can be accessed at https://github.com/IvLabs/toposae.

References

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic Blog*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Doshi, F. R. and Konkle, T. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, 9(25):eade8187, 2023. doi: 10.1126/sciadv.ade8187.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Khosla, M., Murty, N., Tuckute, G., and Kanwisher, N. TopoNets: High performing vision and language models with brain-like topography. In *International Conference on Learning Representations*, 2025.

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L. K., and DiCarlo, J. J. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*, 2020. doi: 10.1101/2020.07.09.185116.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., and Yamins, D. L. K. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451, 2024. doi: 10.1016/j.neuron.2024.04.018.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020.

Pach, M., Karthik, S., Bouniot, Q., Belongie, S., and Akata, Z. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2014.

A. Training Details

All models trained with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), cosine learning-rate schedule, weight decay 0.05, warm-up 10 epochs. ViT-S: batch 256, lr 5×10^{-4}, up to 100 epochs with early stopping (patience 15). TinyViT: same schedule, batch 128, lr 3×10^{-4}, 50 epochs. SAE training: TopK sparse autoencoder, $k = 32$, 2048 features, Adam, lr 10^{-3}, 50 epochs. All experiments run on a single NVIDIA RTX 3050 (4 GB) for TinyViT and analysis; ViT-S checkpoints trained on a remote A100.

Table 3. Training results summary (ViT-S/16, multi-seed).

α	Label	Val acc (mean±std)	Epochs
0.0	Baseline	0.612 ± 0.005	~ 87
0.1	TopoWeak	0.614 ± 0.002	~ 67
1.0	TopoStrong	0.596 ± 0.005	~ 79

B. H1: Full Monosemanticity Tables

Table 4. Neuron-level entropy monosemanticity M_u (ViT-S/16, multi-seed; higher = more selective).

α	Mean M_u	Frac. > 0.5
0.0	0.2340 ± 0.0019	0.26%
0.1	0.2348 ± 0.0025	0.52%
1.0	0.2359 ± 0.0016	0.78%

No α-level comparison of neuron-level M_u is statistically significant.

SAE-feature M_u at 25k images (EXP_012b). The original SAE-feature M_u evaluation (EXP_012, 5k validation images) was undersampled. We re-ran the analysis on 25,000 stratified training images (250 per class, expansion factor 4, 1536 features, $L_1 = 0.005$, 50 epochs); see Table 5. Mean per-feature M_u remains flat across α (0.021, 0.021, 0.026), and the fraction of features with $M_u > 0.5$ is zero throughout. The H1 null therefore holds in both the neuron basis and the SAE-feature basis at the larger sample size, strengthening the H1/H3 dissociation.

Table 5. SAE-feature M_u on 25k stratified train images (expansion factor 4, 1536 features, $n = 3$ seeds per α).

α	SAE L0	Mean M_u	Frac. > 0.3	Frac. > 0.5
0.0	1304	0.021	0.000	0.000
0.1	1316	0.021	0.000	0.000
1.0	1189	0.026	0.000	0.000

C. H2: Confound Analyses

EXP_010 — Accuracy-L0 OLS. OLS regression of SAE L0 on validation accuracy across all 9 ViT-S checkpoints: $R^2 = 0.724$, $p = 0.004$. Residuals for the three $\alpha = 1.0$ seeds: -54, -37, $+21$ (two below the regression line — sparser than accuracy alone predicts, one above). The pattern is suggestive but not conclusive across all seeds.

EXP_011 — PCA dimensionality control. The dissociation at $\alpha = 0.1$ (PCA d_{95} decreases while L0 *increases*) shows that SAE L0 captures representational geometry beyond effective dimensionality.

D. H3: Per-Seed Robustness

All three seeds individually exceed $2\times$; the effect is not driven by a single outlier seed.

Table 6. PCA 95%-variance dimension (d_{95}) vs SAE L0 (ViT-S/16).

α	PCA d_{95}	Δd_{95}	SAE L0	ΔL0
0.0	166.7 ± 3.3	—	1273.9	—
0.1	158.3 ± 3.8	-8.3	1306.4	$+32.5$
1.0	142.3 ± 3.4	-24.3	1133.8	-140.1

Table 7. Per-seed H3 ratios at $\alpha = 1.0$ (single-class, seed-level).

Seed	Cluster $\lvert\Delta\text{logit}\rvert$	Random	Ratio
42	1.282	0.340	3.77×
123	0.911	0.386	2.36×
456	0.892	0.401	2.22×

E. Layer-Wise SAE Analysis

H2 effects are distributed across depth rather than concentrated in a single layer (Figure 3). The L0 gap between $\alpha = 1.0$ and baseline first appears at layer 3 and widens through layers 5–10 (peak gap at layer 10). The $\alpha = 0.1$ curve tracks baseline throughout. The choice of layer 6 for the main-text H2/H3 analyses sits in the middle of the effect band and is therefore representative.

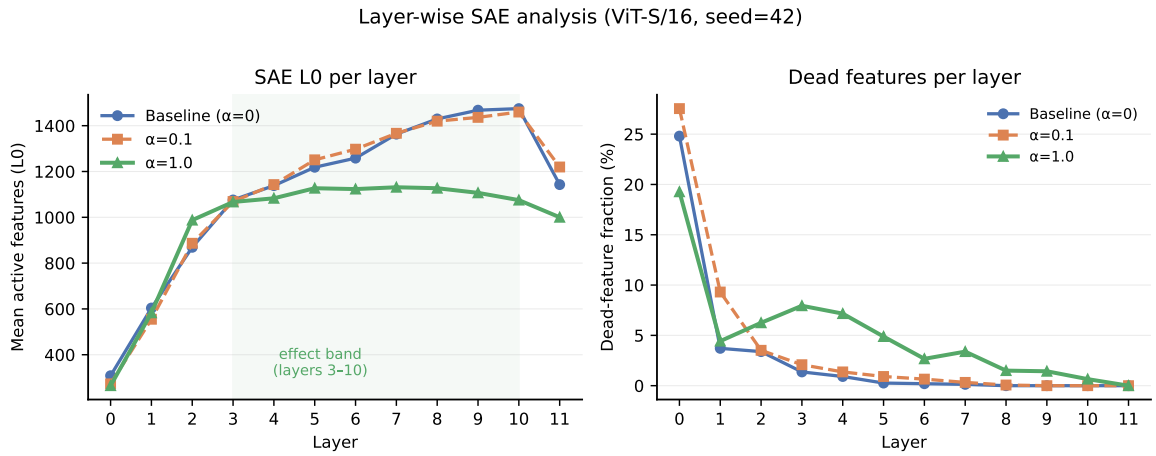

Figure 3. Layer-wise SAE L0 (seed 42). The $\alpha = 1.0$ curve diverges from baseline at layer 3 and remains below through layer 10; $\alpha = 0.1$ tracks baseline. Layer 6 (used for main-text H2/H3 analyses) lies in the middle of the effect band.

F. Alpha-Scaling Rule for H2

TopoLoss gradient magnitude scales as $\nabla_w \mathcal{L}_{\text{topo}} \propto \alpha \cdot w$, so the effective topographic pressure on weight magnitudes scales as $\alpha_{\text{eff}} \propto \alpha \cdot \text{rms}(w)^2$. Calibrating to a reference model (ViT-S, $\alpha = 1.0$):

$$\alpha_{\text{eff}} = \alpha \cdot \left(\frac{\text{rms}_{\text{ref}}}{\text{rms}_{\text{tgt}}} \right)^2 .$$

At $\alpha = 1.0$, TinyViT undergoes weight-norm explosion (RMS grows 3.4× from baseline) while ViT-S remains stable (1.27×). Using the calibration rule, TinyViT at $\alpha = 0.38$ should match ViT-S at $\alpha = 1.0$ in effective pressure. We ran this configuration and found H2 (L0) behaved as predicted (L0 intermediate between $\alpha = 0.1$ and $\alpha = 1.0$) but H3 did not improve over $\alpha = 0.1$ (1.11× vs 1.42×). This falsification suggests H3 depends on factors beyond pressure calibration — model scale, depth, or training budget are candidate explanations requiring a controlled sweep to isolate.

Table 8. Weight-norm scaling and effective α.

Model	α	attn.proj RMS	α_{eff}
TinyViT	0.0	0.0695	—
TinyViT	1.0	0.2346	2.6 (explosion)
ViT-S	0.0	0.0432	—
ViT-S	1.0	0.0548	1.12 (stable)

G. TinyViT Full Results

Table 9. TinyViT (4L/128D) results, seed=42, 50 epochs.

α	Val acc	M_u	SAE L0	Dead%	H3 ratio
0.0	0.472	0.253	444.9	2.0	1.19×
0.1	0.462	0.260	443.7	2.5	**1.42**×
0.38	0.470	0.262	437.5	1.4	1.11×
1.0	0.457	0.249	435.5	2.9	1.12×

H3 peak is at $\alpha = 0.1$, not $\alpha = 1.0$ as in ViT-S. Single-seed; interpret with caution.