Topographic training restructures circuits but the standard metrics won't notice.

Topographic Training Concentrates Causal Circuits Without Improving Neuron Monosemanticity

Gautam Ranka, Shubham Pandere*, Aiden D'souza*

1 TL;DR

MAIN 3 TAKEAWAYS

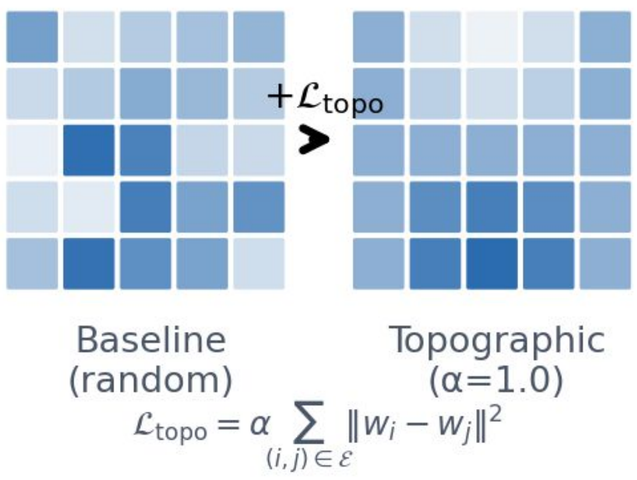
- Topographic training (**TopoLoss**) makes class-relevant neuron clusters **2.8× more causally sufficient** than random unit sets
- **Feature superposition drops** too as SAE L0 sparsity falls 11%, dead features rise 19-fold.
- But neuron-level monosemanticity barely moves; the gain is real but **invisible** to **normal mech-interp metrics**.

2 Background / Motivating Question

- Vision transformers superpose many concepts per neuron.
- Standard fixes (SAEs, dictionary learning) are post-hoc i.e. they decode a frozen model, and not change it.
- We ask a **question**: *can a cheap training-time prior (TopoLoss) make the model itself more interpretable*?

3 Method

- **TopoLoss**: penalize *weight differences* between spatially-adjacent units on a fixed 2D grid, applied to attn.proj across all ViT blocks.
- Audit along: **Neuron selectivity** (entropy-based score per unit), **Feature superposition** (SAE L0 & dead-feature fraction), and **Causal sufficiency** (activation patching: top-16 class neurons).



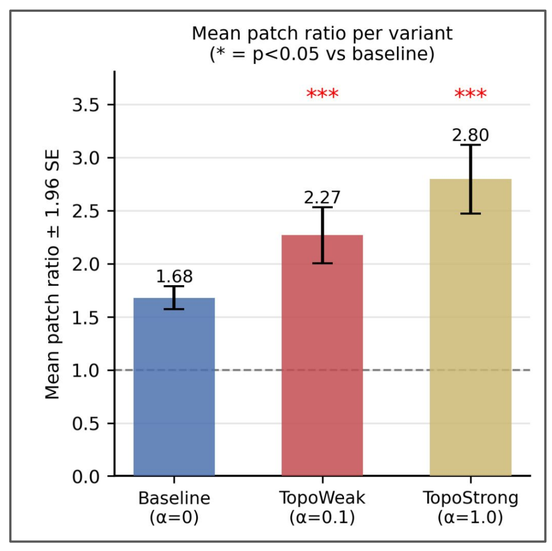
4 Results

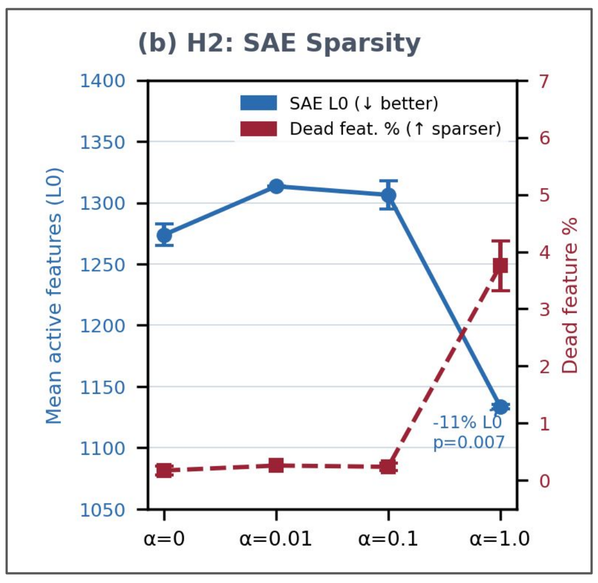
Causal sufficiency (↑)

- Patching a topographic cluster shifts class logits 2.8× more than patching an equivalent random set, with monotone effect in α.

Feature superposition

- SAE L0 falls 11% and dead features rise 19-fold at α=1.0 .



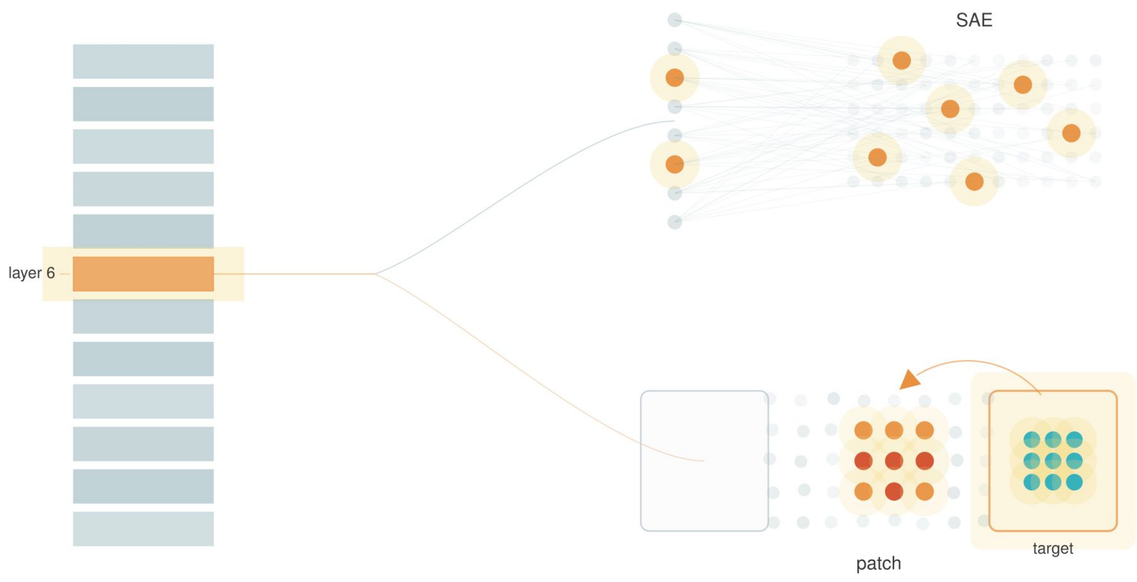


Neuron selectivity (-)

- *Entropy-based selectivity* is **flat** across all α in both the neuron and SAE-feature basis.
- Causal mass **concentrated** into *spatially local circuits* without disentangling individual neurons.

5 Limitations and Discussion

- Single dataset (ImageNet-100), discriminative ViTs only.
- Accuracy-matched ablation left for future work.
- Mu's class-basis may be too coarse to catch mono-semanticity in cross-class features. Future work entails an embedding-based score (*Pach et al., 2025*) to test this independently.



PAPER

CODE

gautam.ranka@ivlabs.in