Vision SAE Behavior is Backbone-Dominated

ViSAEBench: 60 SAEs across 5 ViT backbones and 7 interpretability metrics

A Controlled benchmark design

Same data. Same layer. Same SAE sweep.
Only the pretrained backbone changes.

5 backbones × 12 configs
= 60 SAEs

7 metrics
across 4 dimensions

Spatial Coherence, Reconstruction, Concept Detection, Disentanglement

B What we found: backbone dominates

Variance decomposition across all 60 configurations

M1 Spatial Coherence	99.6%	variance from backbone
M5 Monosemanticity	91.6%	variance from backbone
M7 Absorption	93.1%	variance from backbone

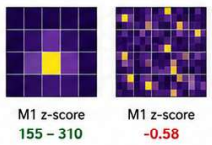
Backbone explains > 60% variance on 5 of 7 metrics (6 of 7 overall)

FVU (reconstruction) is the exception: mostly explained by SAE hyperparameters

C Why it matters

1 Spatial structure is backbone-dependent

SAEs on MAE features show no spatial structure beyond chance, while other backbones produce strong localized features.



2 Low reconstruction error does not imply better downstream preservation.

Across backbones, reconstruction (FVU) and downstream preservation diverge.

Dissociation observed
ρ = -0.70

3 Monosemanticity does not predict fine-grained utility.

Monosemanticity and fine-grained classification show weak positive correlation.

Weak positive relation
ρ = 0.30

NYU

ViSAEBench: Cross-Backbone Evaluation of Vision Sparse Autoencoders Reveals Backbone-Dominated Variance and Metric Dissociations

Vijayraj Gohil*, Diwei Sheng*, Chen Feng

1 TL;DR

MAIN 3 TAKEAWAYS

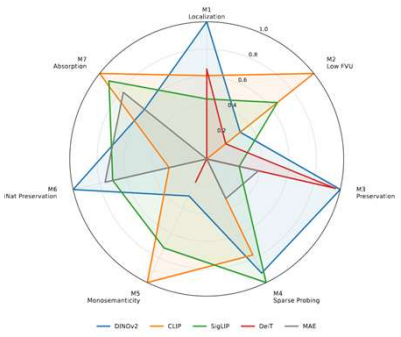
- **SAE behavior is dominated by the underlying pretrained ViT**, not by SAE hyperparameters.
- **Low reconstruction error (FVU) does not guarantee that downstream task accuracy is preserved.**
- **Monosemanticity fails to predict fine-grained classification** across backbones.

2 Background / Motivating Question

- Vision SAEs have **no standardized evaluation protocol** — unlike language SAEs, each paper uses different backbones, metrics, and datasets.
- Existing SAE metrics miss a property unique to vision: **ViT features are spatial**, requiring metrics that capture localization.
- **Key question:** Are single-backbone vision SAE evaluations measuring the SAE — or just the backbone?

3 ViSAEBench: A Unified Evaluation Suite

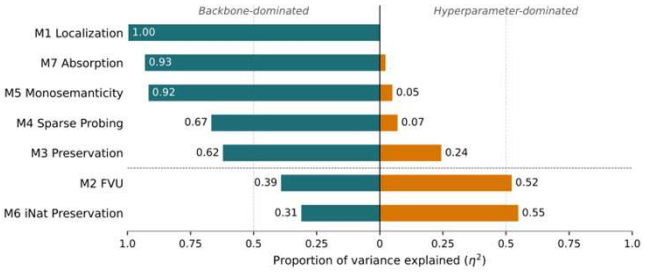
- 60 BatchTopK SAEs trained on **identical ImageNet-1K activations** from 5 ViT-B backbones spanning 4 pretraining paradigms (DINOv2, CLIP, SigLIP, DeiT, MAE).
- **7 metrics across 4 interpretability dimensions:** Spatial Coherence, Reconstruction, Concept Detection, and Disentanglement.

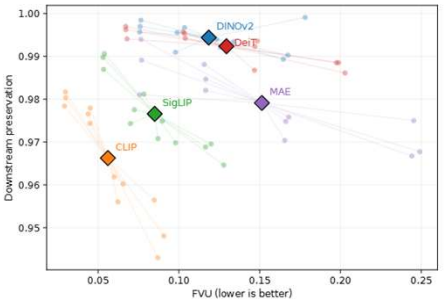


- Novel **spatial coherence metric (M1)** based on Moran's I — quantifies whether SAE features activate in spatially contiguous image regions, a property absent from all prior vision SAE evaluation.

4 Key Findings: Backbone Dominates

- **Variance decomposition** shows backbone explains >90% of variance on 3 of 7 metrics — SAE hyperparameters dominate only FVU.
- **MAE is categorically different:** SAEs on MAE features show spatial activation patterns indistinguishable from random chance.





- **Sparse probing transfers robustly** across domains — ImageNet AUC predicts iNaturalist (ρ = 0.90), CUB-200 (ρ = 0.80), and DTD (ρ = 1.00).
- **No single backbone wins** across all dimensions — DINOv2 leads spatial tasks, SigLIP leads texture, DeiT achieves highest preservation.

5 Limitations and Discussion

- **Only 5 ViT-B backbones** — cross-backbone correlations are directional rather than confirmatory, though findings replicate under JumpReLU (rank correlations ≥ 0.80 on all 7 metrics).
- **SAEs trained on ImageNet-1K only** — features specific to non-natural-image domains (medical imaging, satellite imagery) may be underfit.
- **Backbone standardization is essential** for fair vision SAE comparison — single-backbone evaluations risk measuring backbone properties rather than SAE quality.

PAPER

CODE

Mechanistic Interpretability Workshop, ICML 2026

vg2097@nyu.edu ds5725@nyu.edu