The farther the target domain drifts from ImageNet, the less safe it is to reuse a generic SAE.

Sparse Autoencoders for LoRA-Adapted CLIP Under Domain Shift

Sunayana Samavedam, Deepika Vemuri, Venkat Adithya Amula, P J Narayanan, Vineeth N Balasubramanian

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD · CVIT · भारतीय प्रौद्योगिकी संस्थान हैदराबाद Indian Institute of Technology Hyderabad

OVERVIEW

When CLIP is adapted with LoRA, can an SAE trained on the base model still be trusted to interpret it?

SAEs are reused as fixed dictionaries to interpret CLIP — trained once on base activations, then applied after adaptation. But LoRA rewrites internal weights, **moving the activation distribution** away from the one the dictionary was built on, so the Generic SAE (G-SAE) may explain the adapted model with diffuse, non-selective features.

Contribution. We test whether a generic SAE stays faithful after LoRA adaptation, show it degrades with domain shift, Fine tune it top recover performance and geometry (FT-SAE) and introduce **DAMS**, a coverage-aware monosemanticity metric.

HYPOTHESIS TESTING

1. *baseline reference*

CLIP (Base) → G-SAE

2. *reuse generic dictionary*

CLIP + LoRA → G-SAE

3. *retrain on adapted acts.(key comparison)*

CLIP + LoRA → FT-SAE

SAE : Gated SAE (main); ReLU / Top-k / Matryoshka

Backbone: CLIP ViT-B/16, LoRA on Q & V projections

Datasets: 7 datasets ordered by MMD distance from ImageNet

Caltech · Pets | DTD · UCF101 | EuroSAT · PathMNIST

Domain shift (MMD from ImageNet →)

Metrics: accuracy after reconstruction, NMSE, cos-sim, L0, DAMS

KEY OBSERVATIONS

92% accuracy recovered on EuroSAT

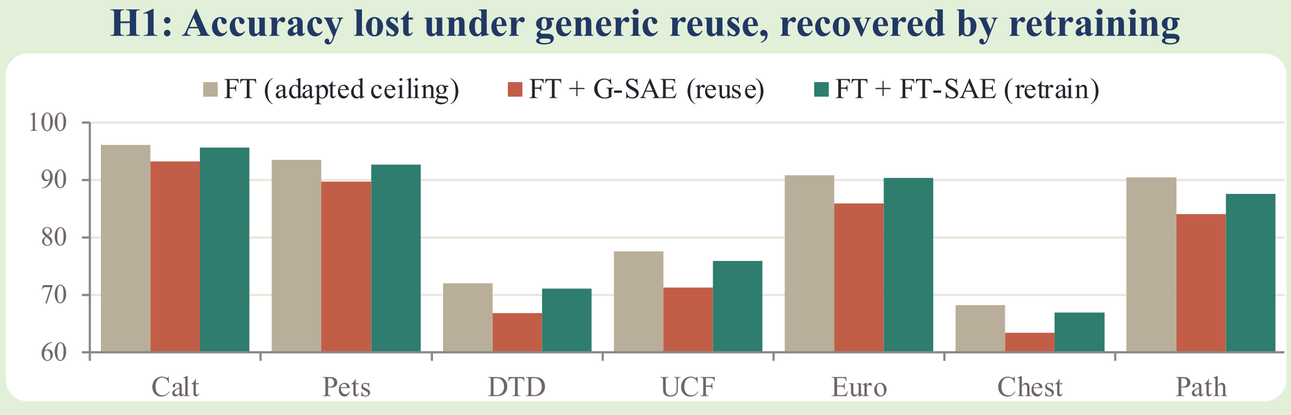
25% NMSE drop on DTD

+0.298 DAMS gain on PathMNIST

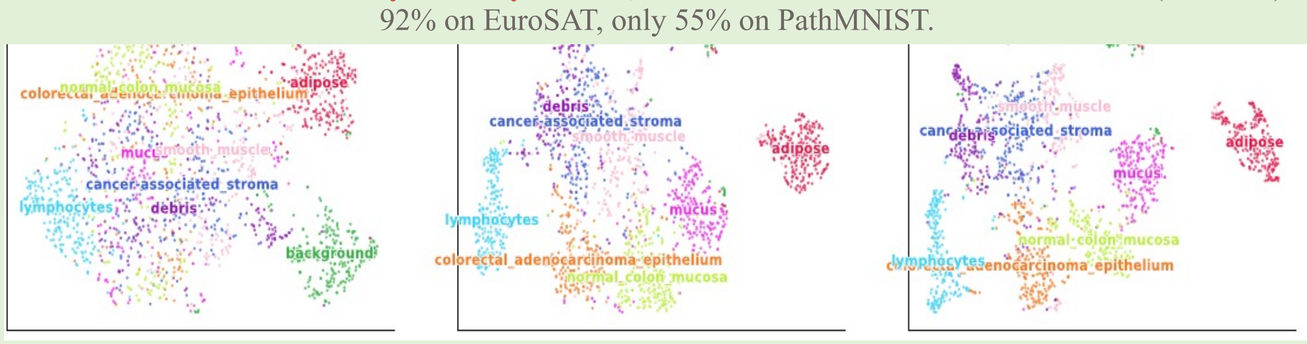
+30% active latents (L0) on Oxford-IIIT Pets

RESULTS

H1: Accuracy lost under generic reuse, recovered by retraining

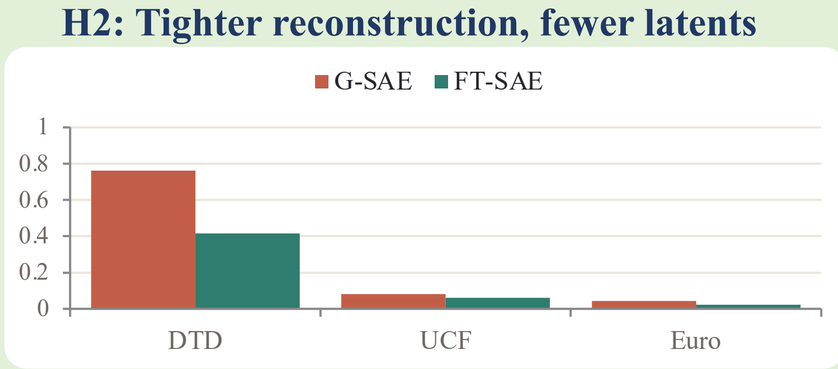


Generic reuse costs accuracy on every dataset; domain-matched SAEs recover 79–83% (near/mid), 92% on EuroSAT, only 55% on PathMNIST.



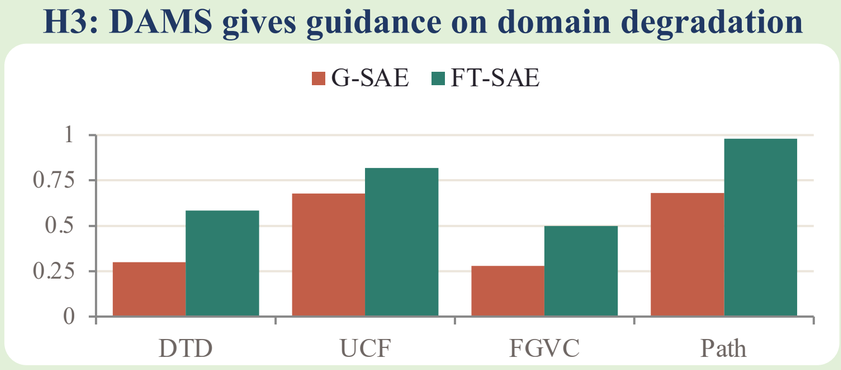
ZS CLIP+GSAE→ entangled, *FT CLIP+GSAE,* *FT-SAE → class-separable clusters*

H2: Tighter reconstruction, fewer latents



NMSE ↓ — drops on 6/7 datasets; L0 falls sharply on shifted domains.

H3: DAMS gives guidance on domain degradation



DAMS ↑ — jumps +0.285 / +0.298 on DTD & PathMNIST

CONCLUSION

The farther the target domain drifts from ImageNet, the less safe it is to reuse a generic SAE.

Treat SAE reuse as a testable assumption — retrain on adapted activations when accuracy drops, NMSE rises, L0 becomes diffuse, or DAMS flags weak target selectivity. MMD is an early-warning signal, not a decision rule on its own.

Project Website

Mechanistic Interpretability Workshop, ICML 2026

ICML International Conference On Machine Learning Seoul, South Korea