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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sparse autoencoders (SAE) are increasingly used to interpret Vision Transformer features, but unlike the language setting, there is no standardized protocol for comparing vision SAEs and no systematic characterization of how SAE quality depends on the pretrained backbone. We introduce ViSAEBench, a unified evaluation suite covering seven metrics across four interpretability dimensions, including a novel spatial coherence metric specific to vision. Using ViSAEBench, we conduct the first controlled cross-backbone study of vision SAEs: 60 SAEs trained on identical ImageNet-1K activations from five ViT-B backbones spanning four pretraining paradigms. Our central finding is that the choice of pretrained backbone dominates vision SAE behavior more than SAE hyperparameters. A variance decomposition shows that backbone explains over 90% of variance on three metrics and over 60% on five of seven, while SAE hyperparameters dominate only reconstruction error. The starkest instance is categorical: across all configurations tested, SAEs trained on Masked Autoencoder features show no spatial structure beyond chance, while the other four backbones produce strongly spatially structured features. Single-backbone vision SAE evaluations are therefore often measuring properties of the backbone more than properties of the SAE. We further identify two metric-level dissociations with practical consequences. First, reconstruction error and downstream task preservation substantially diverge across backbones (Spearman $\rho = -0.70$), so reconstruction error alone cannot be used to compare vision SAEs. Second, monosemanticity, a central SAE quality criterion in language work, does not predict fine-grained classification, indicating that within-feature consistency does not capture the between-class separability downstream tasks require. We release all 60 SAE checkpoints and the ViSAEBench evaluation library.

1 Introduction

Sparse autoencoders (SAEs) have become a standard tool for decomposing neural network representations into interpretable features [Bricken et al., 2023, Templeton et al., 2024]. For language models, this methodology is supported by unified benchmarks such as SAEBench [Karvonen et al., 2025] and large-scale releases such as Gemma Scope [Lieberum et al., 2024]. Vision SAEs are less standardized: recent work shows that SAEs trained on Vision Transformer (ViT) features can recover monosemantic concepts [Pach et al., 2025] and discrete visual primitives [Sharkey et al., 2025], but studies differ in backbone, metrics, and evaluation datasets. Existing vision representation benchmarks such as VTAB [Zhai et al., 2020] and CLIP probing protocols [Radford et al., 2021] evaluate downstream transfer of the raw backbone, not properties of the SAE-decomposed feature space. Moreover, language SAE metrics such as Fraction of Variance Unexplained (FVU), sparsity,

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

and monosemanticity do not capture the spatial structure unique to ViT patch features: a feature firing on a contiguous image region behaves differently from one diffuse across the patch grid.

We introduce **ViSAEBench**, an evaluation suite for SAEs trained on ViT patch features. ViSAEBench spans four dimensions- spatial coherence, reconstruction, concept detection, and disentanglement through seven metrics, including a vision-specific spatial coherence metric (M1) that quantifies whether SAE features activate in spatially contiguous image regions. We apply ViSAEBench to 60 BatchTopK SAEs trained on identical ImageNet-1K activations from five ViT-B backbones spanning four pretraining paradigms: DINOv2 (self-distillation) [Oquab et al., 2024], CLIP [Radford et al., 2021] and SigLIP [Zhai et al., 2023] (vision-language contrastive), MAE (masked reconstruction) [He et al., 2022], and DeiT (supervised distillation) [Touvron et al., 2021].

Our contributions are: (i) ViSAEBench, a seven-metric evaluation suite for vision SAEs spanning spatial coherence, reconstruction, concept detection, and disentanglement; (ii) an evaluation of 60 BatchTopK SAEs trained on identical ImageNet-1K activations from five ViT-B backbones, showing that backbone explains over 90% of variance on three metrics and over 60% on five of seven; (iii) a categorical spatial-coherence separation, with MAE-derived SAEs remaining at the spatial-random null while DINOv2, CLIP, SigLIP, and DeiT produce spatially structured SAE features; and (iv) two metric dissociations, where FVU does not reliably predict downstream preservation across backbones and monosemanticity does not predict CUB-200 fine-grained classification.

2 Related Work

2.1 SAEs for language model interpretability

Sparse autoencoders have emerged as a central tool for decomposing language model representations into interpretable features [Bricken et al., 2023, Templeton et al., 2024, Lieberum et al., 2024]. Large-scale releases such as Gemma Scope have made SAE features a shared substrate for mechanistic interpretability research. Recent work has extended SAEs to cross-modal settings [Gu et al., 2026, Zhao et al., 2026], exploring vision-language alignment through SAE decomposition. SAEBench [Karvonen et al., 2025] introduced a unified evaluation suite for language SAEs, providing the methodological template we adapt here. Monosemanticity, operationalized as within-feature activation consistency, is treated as a central quality criterion across this line of work; whether it tracks downstream task utility has not been systematically tested. ViSAEBench is the vision counterpart, not a replacement: the metric families overlap (sparsity, reconstruction, probing) but the dimensions specific to vision, particularly spatial coherence, have no language analog.

2.2 SAEs for vision

Recent work has applied SAEs to ViT features and shown that monosemantic visual concepts can be recovered from individual SAE features [Pach et al., 2025]. Earlier patch-feature decomposition work and the ViT-Prisma library [Joseph et al., 2025] established the infrastructure for training SAEs on ViT activations. Parallel efforts have explored alternative architectural surrogates for vision interpretability, including cross-layer transcoders [Chatzoudis et al., 2026], which offer dense global decompositions but lack the sparse, localized structure that enables fine-grained concept isolation. Existing vision SAE work has concentrated on a single backbone per paper and a small number of metrics, leaving cross-architectural comparison undone. ViSAEBench's contribution relative to this line of work is a controlled cross-backbone study spanning four pretraining paradigms, a unified protocol covering seven metrics across four interpretability dimensions, and a spatial-coherence metric (M1) absent from prior vision SAE evaluation.

2.3 Vision representation benchmarks and spatial structure

Vision representation benchmarks such as VTAB [Zhai et al., 2020], CLIP probing protocols [Radford et al., 2021], and DINO-style linear-evaluation suites [Oquab et al., 2024] evaluate properties of the raw representation through frozen-backbone probing. ViSAEBench evaluates properties of the SAE-decomposed representation, a different question: a backbone with strong probing accuracy may still produce SAE features that are entangled, spatially diffuse, or poorly disentangled. Spatial structure in vision representations has been studied through object-centric learning [Rubinstein et al., 2025], ViT attention probing [Caron et al., 2021, Darcet et al., 2024], and attribution-based

localization [Han et al., 2025]. M1 imports a different tradition: Moran's I [Moran, 1950] is the standard spatial-autocorrelation statistic in spatial statistics, but to our knowledge has not been used for evaluating SAE features. We adopt it because it provides a principled null model and a single-number per-feature score that aggregates cleanly across SAE configurations.

3 Background and Setup

BatchTopK Sparse Autoencoders. A sparse autoencoder is an encoder-decoder pair (E, D) trained to reconstruct an input activation $\mathbf{a} \in \mathbb{R}^D$ through a sparse intermediate code $\mathbf{c} = E(\mathbf{a}) \in \mathbb{R}^F$, with $F \gg D$. The TopK variant [Makhzani and Frey, 2014, Gao et al., 2024] enforces sparsity by retaining only the K largest activations in $\mathbf{c}$ per sample. BatchTopK [Bussmann et al., 2024] relaxes the per-sample constraint to a per-batch one: in a batch of B activations, the top $B \cdot K$ activations across the entire batch are retained, allowing simpler samples to use fewer features and complex samples to use more. We adopt BatchTopK because visual complexity varies substantially across natural images, and per-sample top-K either underfits complex scenes or wastes capacity on simple ones. The training objective is

$$\mathcal{L} = \mathbb{E}_{\mathbf{a}} \left[\|\mathbf{a} - D(\text{BatchTopK}_K(E(\mathbf{a})))\|_2^2 \right]. \tag{1}$$

To verify our cross-backbone findings are not specific to BatchTopK, we replicate the full evaluation on JumpReLU SAEs across all five backbones; cross-family rank correlations are at least 0.80 on every metric, with six of seven significant at $p < 0.05$ (Appendix Q).

ViT backbones evaluated. 'We evaluated five ViT-Base backbones that span four pretraining paradigms: DINOv2 (self-distillation [Oquab et al., 2024]), MAE (masked patch reconstruction [He et al., 2022]), CLIP and SigLIP (vision-language contrastive [Radford et al., 2021, Zhai et al., 2023]), and DeiT (supervised classification with distillation [Touvron et al., 2021]). All five share the ViT-B architecture (12 transformer blocks, 768 hidden dimension) and use 16×16 patches, except DINOv2, which uses 14×14 patches. We tap activations from block 11 of each backbone, near the top of the residual stream where features have been observed to be most semantically discriminative. Sensitivity to layer choice is reported in Appendix D.

SAE training protocol. We train 12 SAE configurations per backbone, sweeping expansion factors $\{8, 16, 32\}$ and sparsity levels $K \in \{64, 128, 192, 256\}$ on 1.28M ImageNet-1K training-split activations cached from each backbone's block 11. All SAEs are trained with identical optimizer, learning rate, and step counts across backbones; full hyperparameters are reported in Appendix A. The full sweep yields 60 SAEs (5 backbones $\times$ 3 expansion factors $\times$ 4 K values), all released on HuggingFace under the `visaebench` organization.

4 The ViSAEBench Evaluation Suite

4.1 Spatial Coherence: Feature Localization Score (M1)

Motivation. A vision SAE feature is interpretable in the localization sense if it activates on a spatially contiguous region of the input image. Existing SAE metrics do not capture this: FVU measures activation-space reconstruction without reference to spatial layout, sparse probing measures concept presence without regard to where the concept activates, and monosemanticity is invariant to spatial structure within any single image.

Definition. Let $\mathbf{f}_i \in \mathbb{R}^P$ denote the activation vector of SAE feature i across the P patches of one image, where P=196 for patch-16 backbones at 224^2 input and P=256 for DINOv2's patch-14 backbone at 224^2 input. We define the per-feature, per-image Moran's I [Moran, 1950] as

$$I(\mathbf{f}_i) = \frac{P}{\sum_{p,q} w_{pq}} \cdot \frac{\sum_{p,q} w_{pq}(f_{i,p} - \bar{f}_i)(f_{i,q} - \bar{f}_i)}{\sum_p (f_{i,p} - \bar{f}_i)^2}, \tag{2}$$

with w_{pq}=1 for 8-connected (queen-contiguity) neighbours on the patch grid and 0 otherwise, and $\bar{f}_i$ the mean activation across patches. A feature is *evaluable* on an image if it activates on at least 5 patches with activation variance above ε=10^{-8}, and *evaluable in aggregate* if it is evaluable on

at least 50 validation images. The per-feature score $\bar{I}_i$ is the mean of $I(\mathbf{f}_i)$ over the feature's valid images. The SAE-level statistic is the mean across evaluable features:

$$\overline{I}_{\text{SAE}} = \frac{1}{N_{\text{eval}}} \sum_{i \in \mathcal{E}} \bar{I}_i, \tag{3}$$

where $\mathcal{E}$ is the set of evaluable features and $N_{\text{eval}} = |\mathcal{E}|$. We report $\overline{I}_{\text{SAE}}$ as the primary M1 score, alongside the median, interquartile range, and $N_{\text{eval}}/N_{\text{dict}}$ as the fraction of evaluable features. Higher values indicate more spatially coherent features.

Sanity check against the spatial-randomness null. Under the null hypothesis of no spatial autocorrelation, Moran's I has analytic expectation $\mathbb{E}[I] = -1/(P-1)$ [Moran, 1950]. All 60 SAEs reject this null at $|z| > 2$; we therefore use z only as a sanity check and report mean Moran's I, rather than z, as the M1 effect-size statistic because N_{eval} mechanically inflates z through the standard error. The exact z-statistic is given in Appendix F.

4.2 Reconstruction: FVU (M2) and Downstream Preservation (M3)

Motivation. Reconstruction quality is the most commonly reported SAE metric, almost always as Fraction of Variance Unexplained (FVU). The implicit assumption is that low FVU implies preservation of task-relevant information. We argue this assumption fails empirically across backbones (§5) and therefore report FVU as M2 alongside a task-conditional preservation metric M3.

Definition (M2: FVU). Let $\mathbf{a} \in \mathbb{R}^D$ denote a backbone activation vector and $\hat{\mathbf{a}}$ its SAE reconstruction. FVU is the fraction of activation variance unexplained:

$$\text{FVU} = \frac{\mathbb{E}_{\mathbf{a}}\left[\|\mathbf{a} - \hat{\mathbf{a}}\|_2^2\right]}{\mathbb{E}_{\mathbf{a}}\left[\|\mathbf{a} - \bar{\mathbf{a}}\|_2^2\right]}, \tag{4}$$

with $\bar{\mathbf{a}}$ the mean activation over the evaluation set. We compute FVU on the ImageNet-1K validation split using activations from the SAE training layer (§3). Lower is better.

Definition (M3: Downstream Preservation). Let ϕ denote a linear classifier trained on raw backbone activations, and let a and $\hat{a}$ denote raw and SAE-reconstructed activations from a held-out evaluation split. We define

$$M3 = \frac{\text{acc}(\phi(\hat{a}))}{\text{acc}(\phi(a))}.$$

Unlike M2, which penalizes all activation-space reconstruction error, M3 measures the fraction of downstream-relevant information preserved after SAE reconstruction. All probes use matched architectures, splits, and hyperparameters; details are in Appendix A.

4.3 Concept Detection: Sparse Probing (M4), Monosemanticity (M5), Cross-Domain (M6)

Motivation. Spatial coherence (M1) and reconstruction (M2, M3) say nothing about whether SAE features are useful units for reading concepts out of the representation. M4 asks whether features support concept readout under sparsity constraints, M5 asks whether individual features correspond to consistent semantic content, and M6 asks whether these properties hold beyond the SAE's training distribution. Concept detection is the dimension on which prior vision SAE work has concentrated [Pach et al., 2025] and the dimension along which H2 and H4 are tested (§4.5).

Sparse Probing (M4). Given an SAE with feature dimension $F \in \{8D, 16D, 32D\}$, let $c \in \mathbb{R}^F$ denote the pooled SAE code for an image. For each of the 1000 ImageNet-1K classes, we rank SAE features by a univariate class-discrimination score on the probe-training split, train a logistic-regression classifier on the top-k ranked features, and evaluate AUC on held-out examples. M4 is reported as mean AUC at $k = 128$; the full sparsity–accuracy curve appears in Figure 6. Implementation details, including the ANOVA F-statistic ranking, split construction, and probe hyperparameters, are reported in Appendix A.

A non-sparse linear probe trained on all F features can recover most of the information in raw activations regardless of how features are organized, since the probe can find a linear combination that approximates the raw representation. The interesting question is whether individual SAE features, used in small numbers, support concept readout. The shape of the sparse-probing curve across feature budgets is therefore as informative as any k value, and we use $k = 128$ as the main summary point.

Monosemanticity (M5). For each alive SAE feature i, we identify its top-N maximally activating validation images (N=16), embed them with a *separate* backbone, and define the feature's monosemanticity score as the mean pairwise cosine similarity of the N embeddings. Aggregated to a per-SAE score, raw monosemanticity is normalized against a random baseline:

$$\mathrm{M5} = \frac{s_{\mathrm{obs}} - s_{\mathrm{baseline}}}{1 - s_{\mathrm{baseline}}}, \tag{5}$$

where s_{baseline} is the mean pairwise cosine similarity of N randomly sampled validation images under the cross-model embedding. $\mathrm{M5}{=}0$ means the feature's top images are no more similar than random; $\mathrm{M5}{=}1$ means maximally similar. To avoid circularity from scoring features in the same embedding space used to train the SAE, each SAE is evaluated with a different backbone, and the score is normalized by a random-image baseline; details are in Appendix K.

Cross-Domain Generalization (M6). SAEs are trained on ImageNet-1K activations and may overfit to ImageNet-distribution structure. M6 evaluates whether SAE representations preserve task-relevant information under distribution shift. We evaluate three non-ImageNet classification datasets: iNaturalist, CUB-200, and DTD, covering fine-grained species recognition, fine-grained bird classification, and texture recognition. For each dataset, matched linear probes are trained on raw backbone features, SAE-reconstructed features, and SAE code features. We report preservation as the reconstructed-feature probe accuracy divided by the raw-feature probe accuracy, and also evaluate sparse SAE-code probing at $k = 128$. This ratio factors out the raw OOD capacity of the backbone and isolates SAE-induced information loss. Full probe details are given in Appendix L. Full raw, reconstructed, and preserved accuracies appear in Appendix M.

4.4 Disentanglement: Probe-Based Feature Absorption Rate (M7)

Motivation. A well-disentangled SAE should expose class-relevant concepts through small numbers of individual features rather than requiring several related features to recover the same distinction. Feature absorption [Chanin et al., 2024] describes the failure mode in which information that could ideally be represented by a single feature is instead distributed across related features. We adapt this idea to vision through a probe-based criterion: a class distinction is treated as absorbed when the top individual feature is insufficient, but a small set of top-ranked features recovers the distinction.

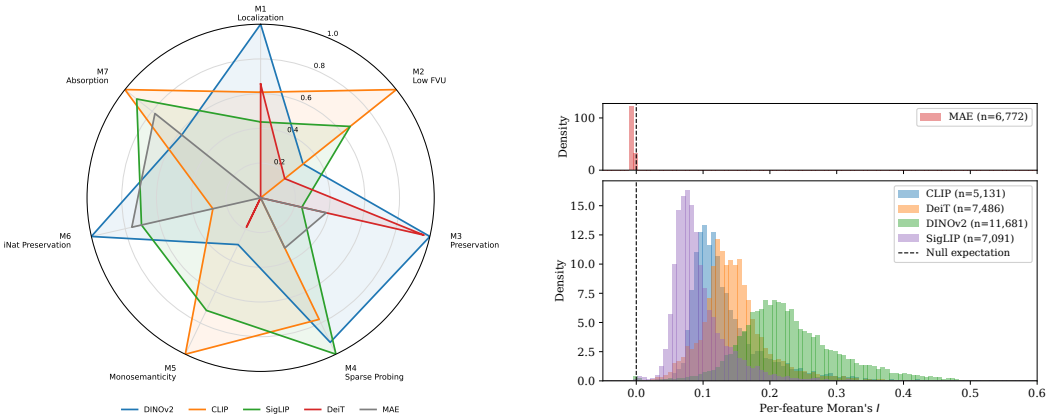
Definition. For each ImageNet hierarchy sibling group with at least three classes, we construct class-versus-sibling probing tasks. SAE codes are pooled to image-level features, and features are ranked by a univariate class-discrimination score on the probe-training split. For each task, we train sparse logistic-regression probes using the top $k = 1$ and top $k = 4$ ranked features and compute their F1 scores. A task is counted as absorbed when

$$F1(k = 1) < \alpha \quad \text{and} \quad F1(k = 4) > \beta,$$

with $\alpha = 0.8$ and $\beta = 0.5$ by default. M7 is the fraction of eligible class-versus-sibling tasks satisfying this criterion. Lower values indicate fewer cases where a class distinction is unrecoverable from the top individual feature but recoverable from a small set of related features. Sensitivity to α and β is reported in Appendix N.

4.5 Hypotheses

The seven metrics measure properties internal to the SAE; whether those properties predict downstream task utility is a separate empirical question. We pre-registered four hypotheses linking SAE-internal metrics to downstream tasks before computing cross-task correlations, each tested cross-backbone with $n = 5$ by aggregating the SAE-internal metric across the 12 configurations per backbone and correlating against a per-backbone downstream number. Spearman's ρ is the test statistic, with confirmation thresholds fixed in advance: **H1** M1 $\rightarrow$ ADE20K segmentation mIoU



(a) Backbone profiles.

(b) Moran's-I distributions.

Figure 1: **Cross-backbone SAE behavior.** (a) Normalized ViSAEBench scores show distinct metric profiles across backbones. (b) Per-feature Moran's I separates MAE, which concentrates near the spatial-random null, from the other four spatially structured backbones.

($\rho \geq 0.7$); **H2** normalized M5 $\rightarrow$ CUB-200 fine-grained accuracy ($\rho \geq 0.6$); **H3** M3 $\rightarrow$ image-image retrieval R@1 on a held-out ImageNet subset ($\rho \geq 0.6$); **H4** M4 (ImageNet) $\rightarrow$ sparse probing AUC on iNaturalist at k=128 ($\rho \geq 0.7$), with CUB-200 and DTD as additional cross-domain checks. The mechanisms are: spatially coherent features carry the structure a segmentation head exploits (H1); within-feature consistency under an independent embedding should track semantically coherent visual content (H2); retrieval is a near-direct test of representation similarity (H3); SAE features that are concept-discriminative on ImageNet should remain so on related natural-image tasks (H4). All four results are reported in §5 regardless of outcome. No single scalar adequately captures vision SAE quality across the four dimensions, and the framework's value lies in characterizing trade-offs rather than producing a leaderboard.

5 Experiments and Findings

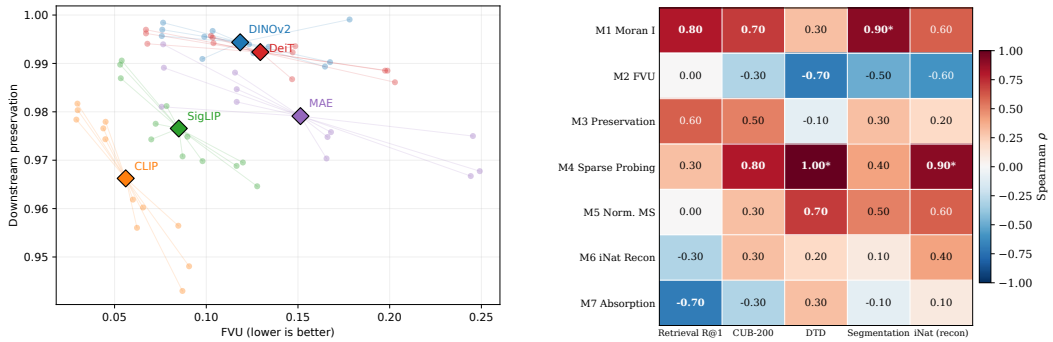
We trained 60 BatchTopK SAEs on identical 1.28M-image ImageNet-1K activation caches (5 backbones, 3 expansion factors, 4 sparsity levels) and evaluated each along the seven metrics defined in §4. Figures 1a through 3 summarize the empirical findings.

Cross-backbone profile Figure 1a shows normalized scores across the seven metrics for each backbone. The five backbones produce qualitatively different SAE feature profiles: DINOv2 dominates the spatial dimension (M1) and ranks first or second on concept-detection metrics; CLIP and SigLIP cluster together on monosemanticity and probing; DeiT trails on concept detection but achieves the highest reconstruction preservation; MAE ranks last on five of seven metrics. No backbone wins across all dimensions. The profiles are consistent with the view that pretraining objective shapes which kinds of structure are SAE-decomposable, rather than producing a uniform "better" or "worse" backbone for interpretability.

5.1 Finding 1: Categorical separation in spatial coherence

Four of the five backbones produce SAE features with strongly non-random spatial structure. M1 (mean Moran's I) values are 0.232 (DINOv2), 0.149 (DeiT), 0.139 (CLIP), and 0.099 (SigLIP), all rejecting the spatial-randomness null $\mathbb{E}[I] = -1/(P-1)$ at $|z| > 150$. MAE produces SAE features whose spatial activation patterns are at the spatial-randomness null in effect size: M1 $= -0.005$, matching the analytic null expectation $\mathbb{E}[I] = -1/(P-1) = -0.005128$ (for P=196) to four decimals, across all 12 MAE configurations.

The interquartile range of MAE's per-feature distribution is 0.0008, two orders of magnitude tighter than DINOv2's 0.091, ruling out the possibility that flatness reflects cancellation between positively



(a) FVU preservation scatter.

(b) Metric-task correlation heatmap.

Figure 2: **Metric dissociations.** (a) FVU does not reliably predict downstream preservation across backbones. (b) Cross-backbone metric–task correlations show that different ViSAEBench metrics capture distinct downstream-relevant properties.

and negatively localized features. Figure 1b shows the per-feature distributions against the null. A diagnostic on raw pre-SAE backbone activations (Appendix G) confirms that MAE's spatial flatness originates in the backbone, not the SAE. Raw MAE activations sit on the null at $\bar{I} = -0.005$. Raw activations from the other four backbones reject the null at $z > 150$, with mean Moran's I between 0.16 and 0.38. SAE decomposition preserves a substantial fraction of the raw spatial coherence (mean ratio 0.64 across the four non-null backbones), so the categorical gap exists in the source representation and the SAE inherits it. Figure 5 shows representative activation maps that visualize this gap directly.

This pattern is consistent with H1: the cross-backbone Spearman correlation between M1 and ADE20K segmentation mIoU is $\rho = 0.90$ ($p = 0.037$, $n = 5$), above the pre-registered threshold $\rho \geq 0.7$

5.2 Finding 2: FVU and downstream preservation dissociate

Figure 2a plots per-configuration FVU against downstream preservation. The two metrics are not monotonically related across backbones. CLIP achieves the lowest mean FVU of any backbone (0.056) but a mean preservation ratio of 0.966, the second-lowest. DeiT achieves the highest mean preservation (0.992) at $\text{FVU} = 0.129$. DINOv2 reaches 0.994 preservation at $\text{FVU} = 0.118$. Within-backbone Spearman correlation between FVU and preservation is weak across all five backbones, and the cross-backbone rank order produced by FVU inverts the order produced by preservation for the highest-fidelity backbones.

The dissociation has a clear practical consequence: comparing vision SAE quality across backbones using FVU alone produces misleading conclusions. CLIP's contrastive pretraining concentrates label-aligned variance on directions used by the contrastive head, and the BatchTopK objective is unaware of which directions those are. Optimizing FVU under this objective can reduce reconstruction error in directions the linear probe never reads while damaging directions it does. Within a single backbone, FVU remains a useful relative metric for comparing SAE configurations. Across backbones, only task-conditional metrics (M3 or M6) support quality claims.

5.3 Finding 3: Sparse probing transfers across natural-image domains (H4)

Cross-backbone Spearman correlation between M4 (ImageNet sparse probing AUC at $k = 128$) and iNaturalist sparse probing AUC at $k = 128$ is $\rho = 0.90$ ($p = 0.037$), clearing the H4 threshold. The same trend holds on DTD ($\rho = 1.00$) and CUB-200 ($\rho = 0.80$), with a weaker association to segmentation transfer ($\rho = 0.40$). M4 is the strongest single predictor of cross-task generalization in our suite (Figure 2b): backbones whose SAE features are concept-discriminative on ImageNet remain so on related natural-image tasks.

5.4 Finding 4: Preservation predicts retrieval but below threshold (H3)

Cross-backbone Spearman correlation between M3 and image-image retrieval R@1 is $\rho = 0.60$ ($p = 0.285$), meeting but not exceeding the H3 threshold given the small sample size. The directional prediction holds, and Figure 2b shows M1 ($\rho = 0.80$) and M3 ($\rho = 0.60$) jointly tracking retrieval. We treat H3 as supported but underpowered.

5.5 Finding 5: Monosemanticity does not predict fine-grained classification

Monosemanticity, as operationalized by top-activating-image consistency under an independent embedding, does not predict fine-grained classification ability across backbones. Cross-backbone Spearman correlation between normalized M5 and CUB-200 accuracy is $\rho = 0.30$ ($p = 0.624$, $n = 5$), below the pre-registered threshold of 0.6, with the directional prediction also unsupported.

The result has a clean mechanism. Within-feature consistency is a property of a feature's activation set in isolation: how perceptually similar are the images on which feature i fires most strongly. Between-class separability is a property of the joint geometry of features across classes: do the features collectively distinguish bird species A from species B. A feature can fire consistently on visually similar images that span multiple fine-grained categories, registering as highly monosemantic while contributing nothing to fine-grained discrimination. The two properties are formally orthogonal, and our data show they are also empirically dissociated.

We decompose M5 into within-class and between-class components by computing mean pairwise cosine similarity within the same ImageNet validation class versus across ImageNet classes. High-M5 features often retain substantial between-class similarity, confirming that M5 captures visual consistency but not necessarily fine-grained class separability (Appendix K). Figure 5 shows representative high-M5 features whose top activating images are perceptually consistent (the bird-on-branch case) but straddle multiple fine-grained classes.

This finding has implications for SAE evaluation practice. Monosemanticity is a central quality criterion in language SAE work [Bricken et al., 2023, Templeton et al., 2024]; the dissociation we observe in vision raises the question of whether the metric tracks downstream task utility in language either. We recommend reporting M5 jointly with concept-detection metrics that test between-feature distinctness, and we suggest that future monosemanticity metrics incorporate between-class separability, for example through a discriminability-weighted consistency score, rather than treating within-feature consistency as a sufficient quality signal.

5.6 Finding 6: DeiT's absorption paradox

DeiT achieves the lowest probe-based absorption rate under the default thresholds ($\alpha = 0.8$, $\beta = 0.5$), with M7 = 0.298 compared to 0.415 (MAE), 0.423 (DINOv2), 0.494 (CLIP), and 0.547 (SigLIP), yet produces the second-worst sparse probing AUC, so low absorption does not imply high concept-detection performance. DeiT remains the lowest-absorption backbone in 7 of 9 settings under a sweep of $\alpha \in \{0.7, 0.8, 0.9\}$ and $\beta \in \{0.3, 0.5, 0.7\}$, breaking only in the strictest corner ($\alpha = 0.9, \beta \geq 0.5$) where MAE and DeiT converge; the paradox is a stable directional pattern rather than an ordering invariant (Appendix N). Recent work on decoder orthogonality constraints [Korznikov et al., 2025] suggests absorption-prone failure modes may be addressable architecturally, motivating future investigation into whether orthogonal decoders reconcile DeiT's behavior.

5.7 Variance decomposition

A natural concern with cross-architecture SAE benchmarking is whether the metric differences reflect SAE properties or are predictable from the pretrained backbone. We address this with a two-way ANOVA on each metric, with backbone (5 levels) and hyperparameter configuration (expansion factor $\times K$, 12 levels) as factors.

Table 3 (Appendix B) reports η^2, the proportion of total sum-of-squares attributable to each factor.

As shown in Fig. 3, backbone explains over 90% of total variance on M1, M5, and M7, and roughly three-quarters on M4 and M6. Differences in expansion factor and K account for almost none of the cross-configuration spread on spatial coherence, monosemanticity, or absorption. M2 inverts the pattern: backbone explains 39.2% of FVU variance while hyperparameters explain 52.3%. M3 sits

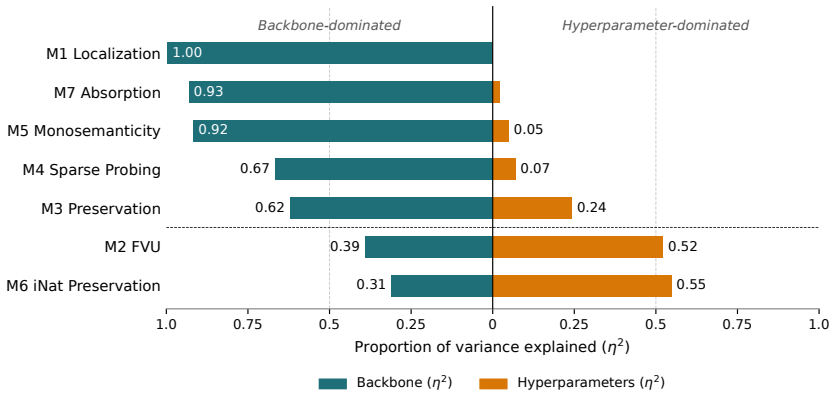


Figure 3: Variance decomposition of each metric across the 60 SAE configurations.

between, with backbone at 62.2% and hyperparameters at 24.3%. Restricting the ANOVA to the four non-MAE backbones yields $\eta^2_{\text{backbone}} = 0.987$ for M1, confirming that the headline result is not driven by MAE's null-indistinguishable cluster.

5.8 Hypothesis summary

Of the four pre-registered hypotheses, three are confirmed (H1, H3, H4). H1 and H4 clear the pre-registered threshold at $\rho \geq 0.7$. H3 meets but does not exceed its threshold. The fourth, H2, is not supported: monosemanticity, as commonly defined, does not track fine-grained classification ability across backbones, with implications for SAE evaluation practice in both vision and language. Figure 2b reports the full 7×5 correlation table across all metric–task pairs, including the 31 non-pre-registered combinations.

6 Limitations and Open Questions

ViSAEBench evaluates five ViT-B backbones at a single layer with a single SAE family (BatchTopK). The $n = 5$ backbone count limits the statistical power of cross-backbone correlations, and we treat single-task Spearman p-values as supporting rather than confirmatory evidence. SAEs are trained on ImageNet-1K activations, so features specific to non-natural-image domains (medical imaging, satellite imagery, synthetic data) may be underfit, and other SAE architectures (JumpReLU, Gated SAE, Matryoshka) may exhibit different cross-backbone profiles. M1 measures whether features activate in spatially contiguous regions, not whether those regions are semantically meaningful (Figure 5). Methodologically, the variance decomposition (§5.7) implies that single-backbone evaluations of spatial coherence, monosemanticity, and absorption measure backbone properties more than SAE properties; we recommend backbone standardization as a prerequisite for SAE-method comparison on these dimensions. FVU is the exception, with hyperparameter choice as its primary driver, making cross-backbone reporting comparatively safer there.

7 Conclusion

We introduced ViSAEBench, an evaluation suite spanning seven metrics across four interpretability dimensions, and applied it to 60 BatchTopK SAEs across five ViT-B backbones. The cross-backbone study surfaces three dissociations that constrain how vision SAEs should be evaluated: MAE-derived SAEs are statistically indistinguishable from a spatially-random null while the other four backbones are not; FVU and downstream preservation diverge across backbones; and monosemanticity does not predict fine-grained classification, despite its central role in language SAE work. Three of four pre-registered hypotheses are confirmed, and the fourth yields a substantive negative result with implications for SAE evaluation in both vision and language.

References

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. https://transformer-circuits.pub/2023/monosemantic-features/index.html, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Gerasimos Chatzoudis, Konstantinos D. Polyzos, Zhuowei Li, Difei Gu, Gemma E. Moran, Hao Wang, and Dimitris N. Metaxas. Can cross-layer transcoders replace vision transformer activations? an interpretable perspective on vision, 2026. URL https://arxiv.org/abs/2604.13304.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Difei Gu, Yunhe Gao, Gerasimos Chatzoudis, Zihan Dong, Guoning Zhang, Bangwei Guo, Yang Zhou, Mu Zhou, and Dimitris Metaxas. Lucid-sae: Learning unified vision-language sparse codes for interpretable concept discovery, 2026. URL https://arxiv.org/abs/2602.07311.

Sangyu Han, Yearim Kim, and Nojun Kwak. Causal interpretation of sparse autoencoder features in vision, 2025. URL https://arxiv.org/abs/2509.00749.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2111.06377.

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, 2025. URL https://arxiv.org/abs/2504.19475.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability, 2025. URL https://arxiv.org/abs/2503.09532.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Elena Tutubalina, and Ivan Oseledets. Ortsae: Orthogonal sparse autoencoders uncover atomic features, 2025. URL https://arxiv.org/abs/2509.22033.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Alireza Makhzani and Brendan Frey. A winner-take-all method for training sparse convolutional autoencoders. In *NIPS deep learning workshop*, 2014.

P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2332142.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models, 2025. URL https://arxiv.org/abs/2504.02821.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders, 2024. URL https://arxiv.org/abs/2407.14435.

Alexander Rubinstein, Ameya Prabhu, Matthias Bethge, and Seong Joon Oh. Are we done with object-centric learning?, 2025. URL https://arxiv.org/abs/2504.07092.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html, 2024. Accessed: 2026-05-03.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021. URL https://arxiv.org/abs/2012.12877.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark, 2020. URL https://arxiv.org/abs/1910.04867.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language–image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2303.15343.

Theodore Zhengde Zhao, Sid Kiblawi, Jianwei Yang, Naoto Usuyama, Reuben Tan, Noel C Codella, Tristan Naumann, Hoifung Poon, and Mu Wei. Learning sparse visual representations via spatial-semantic factorization, 2026. URL https://arxiv.org/abs/2602.01905.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three concrete claims: (1) ViSAEBench provides seven metrics across four interpretability dimensions including a novel spatial coherence metric; (2) backbone explains over 90% of variance on three of seven metrics and over 60% on four of seven, with FVU being the sole hyperparameter-dominated exception; (3) MAE-derived SAEs are categorically separated from the other four backbones in spatial coherence at $z>150$. All three claims are supported by experiments in Section 5 and the variance decomposition in Section 5.7 and Table 3. The framing as a benchmarking suite rather than a leaderboard is reflected in the explicit "no composite score" position in Section 4.5.

   Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses limitations explicitly: (1) scope is restricted to five ViT-B backbones at a single layer with a single SAE family (BatchTopK), with n=5 limiting cross-backbone correlation power; (2) SAEs are trained on ImageNet-1K and may underfit non-natural-image domains such as medical imaging or satellite imagery; (3) other SAE architectures including JumpReLU, Gated, and Matryoshka are not benchmarked; (4) the MAE mechanism is reported empirically without direct manipulation of the pretraining objective; (5) M1 measures spatial contiguity but not semantic meaningfulness and must be read jointly with M4. The paper also acknowledges single-task Spearman pp p-values are treated as supporting rather than confirmatory given n=5n=5 n=5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not contain theoretical results requiring proofs. M1 uses Moran's I, which is a standard spatial statistics quantity with a well-known null distribution; the formula is stated and cited but no novel theorems are claimed.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3.3 specifies the SAE training protocol: 12 configurations per backbone sweeping expansion factors $\{8, 16, 32\}$ and $K \in \{64, 128, 192, 256\}$ on 1.28M ImageNet-1K activations from block 11 of each backbone, with identical optimizer, learning rate, and step counts across backbones. Backbones are listed in Table 1 with sources. Section 4 defines all seven metrics formally, with thresholds, parameter values ($N = 16$ for M5, $\alpha = 0.8$ and $\beta = 0.5$ for M7, $\varepsilon = 10^{-8}$ for M1 evaluability), and evaluation datasets. Full hyperparameters and layer-sensitivity analysis are deferred to Appendix A and C. All 60 SAE checkpoints are released on HuggingFace under the `visaebench` organization.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All 60 SAE checkpoints will be released on HuggingFace under the visaebench organization in safetensors format under Apache 2.0 license, with one repository per backbone containing the 12 configuration subfolders. The evaluation library is released will be released as a pip-installable python package once the paper is accepted. ImageNet-1K, iNaturalist 2021, CUB-200-2011, DTD, and ADE20K are all public datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3.3 specifies the SAE sweep grid (expansion factors, K values, 1.28M training activations, block 11 tap point, identical optimizer settings across backbones). Section 4 specifies metric-level details: M1 evaluability thresholds (5 patches, variance above 10^{-8}, 50 evaluable images), M4 selection rule (top-k by mutual information at $k \in \{32, 128, 512\}$), M5 cross-model embedding pairing with $N = 16$ maximally activating images, M6 datasets (iNat, CUB, DTD, ADE20K) with EuroSAT exclusion documented, and M7 thresholds $\alpha = 0.8$, $\beta = 0.5$. Train/test splits follow the standard ImageNet-1K validation split for training-distribution evaluation. Full optimizer hyperparameters and layer-sensitivity analyses are in Appendix A and C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports statistical significance in three ways. (1) M1 is itself a z-statistic against a permutation null with the null expectation $\mathbb{E}[\bar{I}_{\text{null}}] = -1/(P-1)$ stated explicitly, and individual MAE configurations are flagged as null-indistinguishable. (2) The four pre-registered hypotheses report Spearman ρ with p-values ($n = 5$): H1 $\rho = 0.90$, $p = 0.037$; H2 $\rho = 0.30$, $p = 0.624$; H3 $\rho = 0.60$, $p = 0.285$; H4 $\rho = 0.90$, $p = 0.037$. (3) The two-way ANOVA in Section 5.7 reports η^2 for backbone and hyperparameter factors across all seven metrics in Table 3. The paper is also explicit that with $n = 5$ backbones, single-task p-values are treated as supporting rather than confirmatory, with consistency across multiple downstream tasks Figure 2b as primary evidence.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Compute used: NYU Torch H200 cluster (1 node, 8 GPU) and ai4ce-shannon cluster, with prototyping on a local RTX 3090 (24GB). Slurm with Singularity containers was used for full sweeps. The full sweep is 60 SAEs (5 backbones $\times$ 12 configurations) trained on 1.28M ImageNet-1K activations. Per-SAE training time, total GPU-hours, and storage requirements for the cached activation tensors should be reported in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. The work uses publicly available pretrained backbones and standard public datasets (ImageNet-1K, iNaturalist 2021, CUB-200-2011, DTD, ADE20K), involves no human subjects or crowdsourcing, and the released artifacts are SAE checkpoints and an evaluation library, neither of which presents elevated misuse risk relative to the underlying backbones.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The work is interpretability infrastructure for vision models. Positive impacts: standardized cross-backbone evaluation supports more rigorous mechanistic interpretability research, helps identify representational pathologies (such as MAE's null-indistinguishable spatial structure), and reduces wasted compute from non-comparable single-paper SAE evaluations. Negative impacts are indirect: more reliable interpretability tooling could be applied to vision systems in surveillance or content moderation pipelines, but the contribution itself is methodological infrastructure rather than a deployable system, and SAE evaluation does not enable capabilities beyond those already present in the underlying pretrained backbones.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released artifacts are SAE checkpoints trained on activations from publicly released pretrained backbones, plus an evaluation library.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All five backbones are publicly released under their original licenses and properly cited (Table 1): DINOv2 (Oquab et al., 2024, CC-BY-NC 4.0), MAE (He et al., 2022, CC-BY-NC 4.0), CLIP (Radford et al., 2021, OpenAI License), SigLIP (Zhai et al., 2023, Apache 2.0), DeiT (Touvron et al., 2021, Apache 2.0). Datasets used are ImageNet-1K (custom research license), iNaturalist 2021 (CC-BY-NC 4.0), CUB-200-2011 (CC0 1.0), DTD (CC-BY 4.0), and ADE20K (custom research license), all publicly available with their respective licenses. The overcomplete library BatchTopK SAE (Apache 2.0) and Prisma library (Joseph et al., 2025, MIT) are cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Two new assets are released: (1) 60 SAE checkpoints under the HuggingFace, organized as five repositories (one per backbone) each containing the 12 configuration, released under Apache 2.0 license; (2) a pip-installable evaluation library implementing the seven metrics. Documentation of training hyperparameters, configuration naming, and metric usage will be in the repository README and Appendix A and C.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or human subjects. All evaluations use publicly available pretrained backbones and standard public benchmark datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects research and therefore does not require IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification:

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Appendix

A Experimental Setup and Hyperparameters

We report the full hyperparameters used for SAE training and evaluation in Tables 1 and 2. Unless otherwise stated, all values are held fixed across the five backbones and across all 60 SAE configurations. The only swept SAE hyperparameters are the expansion factor and the BatchTopK sparsity level K.

Table 1: SAE training hyperparameters used in the main sweep.

Hyperparameter	**Value**
SAE family	BatchTopK SAE
Backbones	CLIP ViT-B/16, DINOv2 ViT-B/14, SigLIP ViT-B/16, MAE ViT-B/16, DeiT ViT-B/16
Layer	11
Input dimension	768
Expansion factors	$\{8, 16, 32\}$
Dictionary sizes	$\{6144, 12288, 24576\}$
K values	$\{64, 128, 192, 256\}$
Training activations	Cached ImageNet-1K train activations from backbone's layer-11 patch tokens
Activation normalization	Per-dimension mean subtraction and scalar standard-deviation division using cached activation statistics
Optimizer	Adam
Learning rate	1×10^{-3}
Batch size	8192 patch tokens per step
Epochs	1
Seed	42
Precision	FP16
Inference threshold	Fixed BatchTopK threshold estimated from validation activations and saved with each checkpoint

Table 2: Evaluation hyperparameters used for ViSAEBench metrics (M1–M7).

Metric	Hyperparameters / Protocol
M1 (Spatial coherence)	Grid inferred from patch count (14×14 for 196 patches and 16×16 for 256 patches); 8-neighbor connectivity; minimum active patches per image $= 5$; activation variance threshold $\epsilon = 10^{-8}$; minimum valid images per feature $= 50$; SAE encoding batch size $= 512$; image-group size $= 64$.
M2 (FVU)	Computed on cached ImageNet-1K validation activations from layer 11; reports token-level FVU, mean L_0, and dead-feature count.
M3 (Downstream preservation)	Uses cached ImageNet-1K validation activations with an internal stratified 80/20 split, seed $= 42$; image-level features are obtained by max-pooling patch features; multinomial logistic regression with `lbfgs`, $C = 1.0$, `max_iter`=1000, and `class_weight=None`; the same split is used for raw and reconstructed activations; preservation is reconstructed accuracy divided by raw-activation accuracy.
M4 (Sparse probing)	Uses cached ImageNet-1K validation activations with an internal stratified 80/20 split, seed $= 42$; SAE codes are max-pooled to image-level features; features are ranked once by descending ANOVA F-statistic using `f_classif`; feature budgets $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$; probes use logistic regression with `lbfgs`, $C = 1.0$, `max_iter`=500, and `class_weight=None`; $k = 128$ is the main reported sparse-probing score.
M5 (Monosemanticity)	Top-N activating validation images per feature with $N = 16$; up to 2048 live features are scored; live features are identified by positive maximum activation; embeddings are computed with an independent cross-backbone embedding model; score uses activation-weighted mean pairwise cosine similarity of L2-normalized embeddings; baseline uses 1000 random pseudo-features, each sampling 16 validation images; normalized score is $(s_{\text{obs}} - s_{\text{baseline}})/(1 - s_{\text{baseline}})$.
M6 (Cross-domain)	Evaluated on iNaturalist, CUB-200, and DTD. For each dataset, activations are extracted from the same frozen backbone at layer 11 and normalized with ImageNet statistics; raw activations, SAE reconstructions, and SAE codes are max-pooled to image-level features. Matched linear probes are trained separately for each representation. Preservation is reconstructed-probe accuracy divided by raw-feature accuracy; sparse SAE-code probing uses $k = 128$.
M7 (Probe-based absorption)	Uses max-pooled SAE codes and ImageNet hierarchy sibling groups with minimum group size $= 3$; features are ranked by `f_classif` for class-vs-sibling probes; absorption is detected when $F1(k = 1) < \alpha$ and $F1(k = 4) > \beta$; default thresholds are $\alpha = 0.8$ and $\beta = 0.5$; probes use logistic regression with `lbfgs`, $C = 1.0$, `max_iter`=500, and `class_weight=None`; sensitivity analysis sweeps $\alpha \in \{0.7, 0.8, 0.9\}$ and $\beta \in \{0.3, 0.5, 0.7\}$.

B Backbone and Hyperparameter Variance Decomposition

Table 3: Variance decomposition η^2 values for each ViSAEBench metric.

Metric	η^2_{backbone}	η^2_{hp}	η^2_{resid}
M1 Localization	0.996	0.002	0.002
M7 Absorption	0.931	0.023	0.046
M5 Monosemanticity	0.916	0.050	0.034
M4 Sparse Probing	0.668	0.070	0.262
M3 Preservation	0.622	0.243	0.135
M2 FVU	0.392	0.523	0.086
M6 iNat Preservation	0.311	0.549	0.141

C Probe training and normalization details

All probe-training choices are matched across backbones to avoid introducing evaluation bias. For SAE training, all backbones use the same BatchTopK architecture and the same sweep protocol: expansion factor in $\{8, 16, 32\}$, sparsity $K \in \{64, 128, 192, 256\}$, learning rate 10^{-3}, batch size 8192, one training epoch, and seed 42. The primary regularizer is the explicit BatchTopK sparsity constraint, so the sparsity regularization strength is harmonized by construction across backbones.

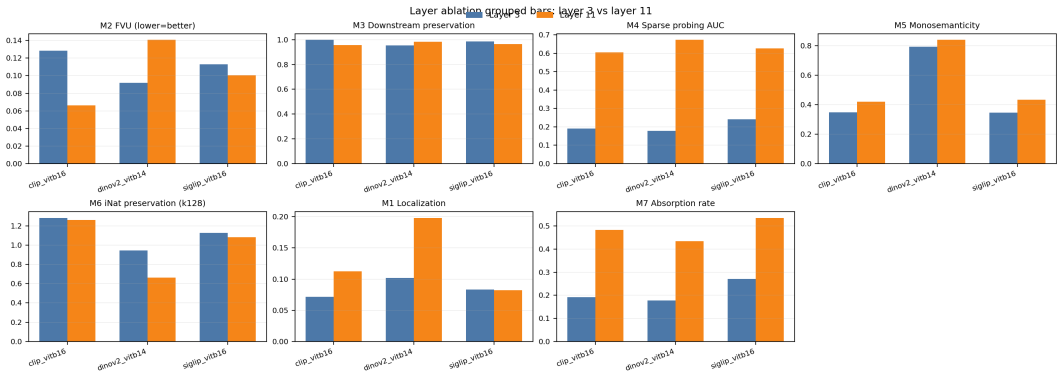


Figure 4: **Layer ablation comparing early-layer and late-layer SAEs.** We compare SAEs trained on layer 3 and layer 11 across seven ViSAEBench metrics for CLIP, DINOv2, and SigLIP. The comparison shows that some metrics are relatively stable across layers, while others change substantially with layer depth. Reconstruction error (M2) and downstream preservation (M3) do not exhibit a single uniform direction across backbones, suggesting that the effect of layer choice is backbone dependent. In contrast, sparse probing performance (M4) and feature absorption (M7) are consistently higher for layer 11, indicating that later-layer SAE features better support sparse concept readout and show stronger absorption structure. Monosemanticity (M5) changes only mildly, preserving the same broad backbone ordering, while localization (M1) increases most clearly for DINOv2 at layer 11.

For linear probing in M3 and related probe-based metrics, we use logistic regression with the L-BFGS solver and $C = 1.0$. In scikit-learn this corresponds to L2-regularized logistic regression under the default penalty. The same solver, regularization strength, and optimization settings are used for all backbones and tasks. We do not use class weighting in any probe; equivalently, `class_weight=None`, the scikit-learn default. Thus, differences in preservation or probing accuracy are not attributable to backbone-specific probe regularization or class reweighting.

Feature normalization is also applied consistently. Before SAE encoding and evaluation, activations are standardized using the cached dataset statistics: we subtract the activation mean and divide by the stored standard deviation. The same normalization procedure is used for every backbone, SAE configuration, and evaluation metric.

We did not evaluate whitening-based preprocessing, such as PCA or ZCA whitening, before SAE training or before probing. Whitening may change the geometry of both the SAE training objective and the linear probe problem, and we leave this as a follow-up ablation rather than including untested whitening claims in the present study.

D Layer ablation

Our main experiments train SAEs on the canonical late-layer representation of each backbone. To test whether the reported SAE properties are specific to this layer choice, we additionally train matched SAEs on an earlier transformer layer and compare them against the layer used in the main benchmark. Specifically, we compare SAEs trained on layer 3 with SAEs trained on layer 11 for CLIP, DINOv2, and SigLIP, using the same evaluation protocol as in the main paper. This ablation asks whether properties such as reconstruction fidelity, sparsity, downstream preservation, spatial localization, monosemanticity, cross-domain preservation, and absorption change systematically as a function of layer depth.

Figure 4 shows that layer choice affects SAE behavior, but not all metrics respond in the same way. For reconstruction fidelity, measured by M2 FVU, the early and late layers do not follow a universal trend: CLIP and SigLIP obtain lower FVU at layer 11, whereas DINOv2 obtains lower FVU at layer 3. This suggests that reconstruction difficulty is not determined by depth alone, but also by the geometry of the underlying backbone representation at that layer.

Downstream preservation, measured by M3, is comparatively stable across the two layers. All three backbones remain close to the raw-feature reference, and the layer-3 versus layer-11 differences are

small relative to the larger cross-backbone effects observed in the main benchmark. This supports the interpretation that SAE reconstruction can preserve much of the downstream signal at both early and late layers, even when other interpretability metrics differ.

The clearest systematic layer effect appears in sparse probing and absorption. M4 sparse probing AUC is substantially higher for layer 11 than layer 3 across all three backbones, indicating that late-layer SAE features are more directly useful for sparse concept readout. M7 absorption rate also increases consistently at layer 11, suggesting that later representations exhibit stronger feature-sharing or absorption structure under our criterion. These two trends are consistent with the intuition that later transformer layers contain more semantically organized features than earlier layers.

For M5 monosemanticity, the absolute values change only modestly between layers, and the backbone ordering is largely preserved: DINOv2 remains substantially higher than CLIP and SigLIP. Thus, although layer depth affects sparse probing and absorption, the relative monosemanticity differences across backbones are not explained away by the choice of layer. Finally, M1 localization shows a notable increase for DINOv2 at layer 11, while CLIP increases more moderately and SigLIP is nearly unchanged. This again suggests that the spatial organization of SAE features depends jointly on the backbone and the layer at which the SAE is trained.

Overall, this ablation supports using a fixed canonical late layer for the main controlled study, while also showing that layer choice is a meaningful experimental axis. The main conclusions are not simply artifacts of evaluating a single layer: backbone-dependent differences persist across layers, but late-layer SAEs tend to produce stronger sparse probing performance and higher absorption rates.

E BatchTopK inference threshold

BatchTopK training applies sparsity by retaining the top $B \cdot K$ activations across a batch of B tokens. At inference time, however, evaluation should not depend on the other samples that happen to appear in the same batch. We therefore replace batch-coupled top-k selection with a fixed activation threshold estimated after training.

After each SAE is trained, we compute a robust inference threshold on held-out validation activations and serialize it with the run artifacts. Specifically, the threshold is saved both in the model checkpoint, under `_running_threshold`, and in the corresponding `config.yaml`, under `threshold`. This explicit serialization is necessary because the underlying Overcomplete implementation stores `running_threshold` as a Python attribute rather than as a registered PyTorch buffer, so it would not otherwise be captured by `state_dict`.

At evaluation time, all metric scripts restore the saved threshold before encoding activations. Thus, SAE inference is deterministic and reproducible from the released checkpoint and configuration alone. We verified that all 60 checkpoints in the main 1M-activation sweep include `_running_threshold` in `sae.pt` and that all corresponding run configurations include the explicit `threshold` field.

F M1 Null-Check Statistic

For completeness, the SAE-level null-check statistic used for M1 is

$$z = \frac{\bar{I}_{\mathrm{SAE}} - \mathbb{E}[I]}{\widehat{\mathrm{SE}}}, \qquad \widehat{\mathrm{SE}} = \frac{\mathrm{sd}(\{\bar{I}_i\}_{i \in \mathcal{E}})}{\sqrt{N_{\mathrm{eval}}}},$$

where $\mathcal{E}$ is the set of evaluable features and $N_{\mathrm{eval}} = |\mathcal{E}|$. Because N_{eval} is on the order of 10^4, even small deviations from the analytic null can produce large z-values; throughout the paper we therefore use z only to test null consistency and use mean Moran's I for cross-backbone comparison.

G Raw-Backbone Spatial Coherence Diagnostic

Finding 1 shows that MAE-derived SAE features are near the spatial-random null under M1, whereas the other four backbones produce strongly spatially structured SAE features. To test whether this gap is introduced by SAE training or inherited from the pretrained representation, we repeat the Moran's-I analysis directly on raw pre-SAE backbone activations.

For each backbone, we use the cached ImageNet-1K validation activations from the same layer used for SAE training. Each raw activation channel is treated as a spatial feature map over the ViT patch grid. We compute Moran's I using the same patch-grid statistic as M1, with 8-neighbor connectivity and the same analytic spatial-random null expectation,

$$\mathbb{E}[I_{\text{null}}] = -\frac{1}{P-1},$$

where P is the number of patches. Thus, the null expectation is -0.0051 for patch-16 backbones with $P = 196$, and -0.0039 for DINOv2-B/14 with $P = 256$.

Table 4: Raw-backbone spatial coherence diagnostic. Moran's I is computed directly on pre-SAE backbone activation channels using the same patch-grid statistic as M1.

Backbone	**Mean raw I**	**Null $\mathbb{E}[I]$**	**Raw z**	**Conclusion**
DINOv2	0.3829	-0.0039	202.03	strongly non-random
CLIP	0.2065	-0.0051	196.51	strongly non-random
SigLIP	0.1616	-0.0051	219.64	strongly non-random
DeiT	0.2225	-0.0051	168.08	strongly non-random
MAE	-0.0052	-0.0051	-6.17	near-null effect size

The raw-backbone diagnostic mirrors the SAE-level pattern. DINOv2, CLIP, SigLIP, and DeiT all have large positive raw Moran's-I values and reject the spatial-random null by a large margin. MAE, in contrast, has mean raw Moran's $I = -0.0052$, numerically almost identical to the analytic null expectation of -0.0051. Although its raw z-score is below the $|z| < 2$ null-indistinguishability threshold, the effect size is essentially at the null. This supports the interpretation that MAE's SAE-level spatial flatness is inherited from the pretrained backbone representation rather than introduced by the BatchTopK SAE.

Sensitivity of M1 to design choices. M1 has three design choices that warrant brief comment. First, the evaluability threshold of 5 active patches per image was chosen to ensure sufficient spatial support for Moran's I to be meaningful while admitting features that activate on small object regions; the cross-backbone ranking is stable to thresholds in $\{1, 5, 20\}$ on the panel we tested. Second, we use 8-connected (queen contiguity) neighbours, the standard choice in spatial statistics; the (N/S) normalization in Moran's I absorbs the row-sum difference between rook and queen contiguity, and on blob-shaped ViT features the two yield rank-correlated SAE scores. Third, DINOv2 ViT-B/14 produces a 16×16 patch grid (P=256) while the other four backbones produce 14×14 (P=196). The null expectation $\mathbb{E}[I] = -1/(P-1)$ differs by 0.0012 between these two grids, two orders of magnitude smaller than the cross-backbone effect sizes we observe (DINOv2 leads the next-best P=196 backbone, DeiT, by 0.083 in mean Moran's I). The raw-backbone diagnostic in this appendix shows the same DINOv2-leading ranking on raw pre-SAE activations, indicating that the spatial-coherence gap is not an artifact of the M1 statistic interacting with SAE encoding at different grid resolutions. A resampling-based control that interpolates DINOv2 feature maps to 14×14 before computing I would further isolate the effect of pretraining from grid resolution; we leave this to future work.

H Qualitative audit of M1 spatial coherence

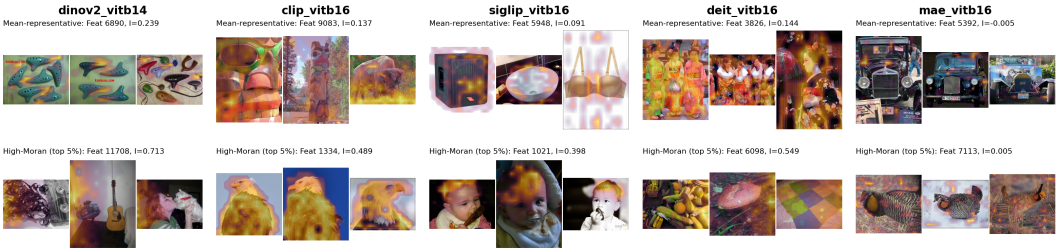


Figure 5: **Qualitative audit of M1 spatial coherence across backbones.** Each column corresponds to one pretrained backbone and shows two SAE features from the BatchTopK 16×, $K = 192$ configuration. The first row shows a mean-representative feature, selected as the evaluable feature whose per-feature Moran's I is closest to the backbone mean. The second row shows a high-Moran feature from the top 5% of the same distribution. Each cell displays top-activating validation images with patch-level activation overlays. DINOv2, CLIP, SigLIP, and DeiT show spatially contiguous activations that often align with localized visual structure, while MAE remains close to the spatial-random null even in the high-Moran row. This figure illustrates that M1 captures spatial organization, not semantic purity.

M1 is designed to measure whether an SAE feature activates in a spatially contiguous region of an image. It is not, by itself, a semantic-purity metric: a feature can have high spatial autocorrelation because it activates on a compact object part, a localized texture, a background region, or another spatially coherent but semantically broad pattern. We therefore include a qualitative audit to show what different regions of the per-feature Moran's-I distribution look like across backbones.

Figure 5 shows representative SAE features from the BatchTopK 16×, $K = 192$ configuration for each backbone. For each backbone, we show two features. The first row selects a *mean-representative* feature, defined as the evaluable feature whose per-feature Moran's I is closest to the backbone's mean Moran's I for that configuration. This row visualizes what a typical spatial-coherence feature looks like for the backbone. The second row selects a *high-Moran* feature from the top 5% of the same per-feature Moran's-I distribution, visualizing the upper tail of spatially coherent features. For each selected feature, we display its top-activating validation images with the corresponding patch-level activation overlay.

The qualitative pattern matches the quantitative M1 separation in Section 5.1. DINOv2, CLIP, SigLIP, and DeiT produce features whose activations often concentrate on contiguous image regions. In several cases these regions align with visually interpretable content, such as object parts, foreground objects, or localized textures. This supports the use of M1 as a filter for finding spatially organized, concept-like SAE features. However, the examples also show why M1 should not be interpreted as a standalone monosemanticity score: spatially coherent features may still correspond to broad visual structure rather than a single class-pure concept.

MAE provides the clearest contrast. Both the mean-representative MAE feature and the top-5% MAE feature remain near the spatial-random null, with per-feature Moran's I close to -0.005. Their activation overlays are diffuse and fragmented compared with the other backbones. This qualitative behavior is consistent with the aggregate result that MAE-derived SAEs are statistically indistinguishable from the spatial-random null under M1, whereas the other four backbones produce strongly spatially structured SAE features.

Overall, the grid illustrates the intended interpretation of M1: high Moran's I identifies features with spatially organized activation patterns and can help surface candidate localized concepts, but semantic meaningfulness must be assessed jointly with concept-detection metrics such as M4/M5 and with qualitative inspection.

I M4 Sparse Probing Curves

In the main text, M4 is summarized by sparse-probing AUC at a fixed feature budget $k = 128$. Here we report the full sparse-probing curves across SAE training sparsity levels and expansion factors. For each SAE, image-level SAE codes are obtained by max-pooling patch-level feature activations. For each ImageNet class, features are ranked by a univariate class-discrimination score on the probe-training split, and logistic-regression probes are trained using the top-ranked features. We report mean AUC across classes.

Appendix I shows the resulting curves for all five backbones. The cross-backbone ordering is broadly stable across expansion factors and SAE sparsity levels: DINOv2 and SigLIP consistently provide the strongest sparse concept readout, CLIP is intermediate, and MAE and DeiT are weaker. This supports the use of $k = 128$ as a representative summary point in the main text, while also showing that the M4 trends are not an artifact of a single SAE configuration.

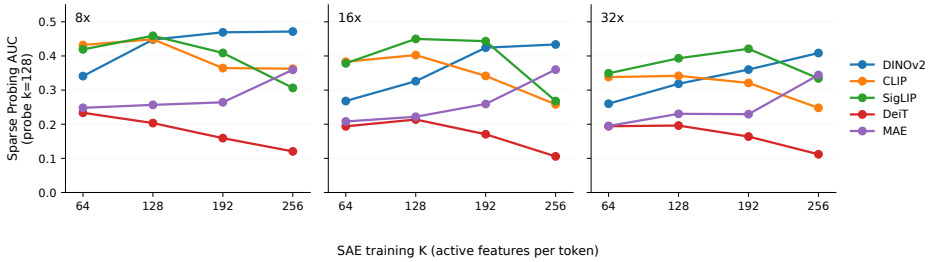


Figure 6: **Sparse probing curves.** ImageNet sparse-probing AUC across SAE training sparsity K and expansion factors. DINOv2 and SigLIP consistently support stronger sparse concept readout than MAE and DeiT across the sweep.

J Per-Task Metric Correlation Panels

The main text summarizes metric–task relationships with a heatmap. Here we provide the same cross-backbone Spearman correlations as per-task panels, which make it easier to compare which ViSAEBench metrics are most predictive for each downstream task.

Figure 7 reports correlations between the seven ViSAEBench metrics and four downstream tasks: retrieval, CUB-200 fine-grained classification, DTD texture classification, and ADE20K segmentation. M4 sparse probing is the strongest predictor for the two natural-image concept-recognition tasks: it reaches $\rho = 0.80$ on CUB-200 and $\rho = 1.00$ on DTD. This supports the interpretation that ImageNet sparse probing captures concept-discriminative structure that transfers to related natural-image domains. In contrast, M1 Moran's I is most predictive for segmentation, reaching $\rho = 0.90$, consistent with the role of spatial coherence in dense prediction. M2 FVU is negatively correlated with most downstream tasks, reinforcing that reconstruction error alone is not a reliable cross-backbone quality criterion.

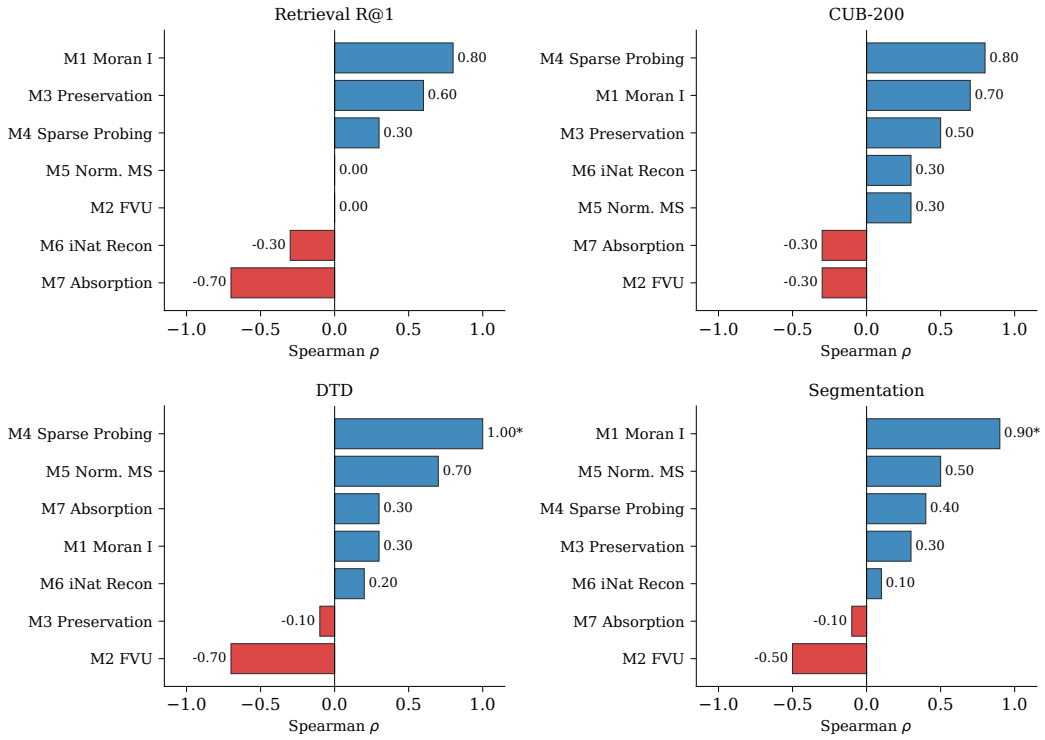


Figure 7: **Per-task metric correlations.** Cross-backbone Spearman correlations between ViSAEBench metrics and downstream tasks. M4 sparse probing best predicts CUB-200 and DTD performance, while M1 spatial coherence best predicts segmentation.

K M5 monosemanticity protocol

Table 5: Exact cross-backbone pairing used for M5 monosemanticity evaluation. Each SAE is evaluated using image embeddings from a backbone different from the one on which the SAE was trained, avoiding circularity in the top-activating-image similarity score.

SAE training backbone	M5 evaluation embedding backbone
DINOv2-B/14	CLIP-ViT-B/16
CLIP-ViT-B/16	DINOv2-B/14
SigLIP-ViT-B/16	DINOv2-B/14
DeiT-ViT-B/16	CLIP-ViT-B/16
MAE-ViT-B/16	CLIP-ViT-B/16

M5 implementation and normalization. For each SAE, we encode all ImageNet validation activations and obtain one image-level SAE code vector per image by max-pooling over patches, i.e., for each feature we keep its maximum patch activation within the image. A feature is considered *live* if its maximum image-level activation over the validation set is strictly positive; features with maximum activation equal to zero are treated as dead and excluded from scoring. From the live features, we evaluate up to 2,048 features, sampled uniformly without replacement when more than 2,048 are live and otherwise using all live features.

For each selected feature j, we rank validation images by that feature's image-level activation and take the top-k images, with $k = 16$ unless the dataset slice contains fewer than 16 images. Let a_{ij} denote the activation of feature j on top image i. We convert activations to nonnegative weights by

min–max normalizing feature j over all validation images:

$$w_{ij} = \frac{a_{ij} - \min_n a_{nj}}{\max_n a_{nj} - \min_n a_{nj}}.$$

The selected top images are then embedded with the designated cross-backbone evaluator from Table 5, and all embeddings are L2-normalized. The feature-level monosemanticity score is the activation-weighted mean pairwise cosine similarity

$$\mathrm{MS}_j = \frac{\sum_{u<v} w_{uj} w_{vj} \cos(e_u, e_v)}{\sum_{u<v} w_{uj} w_{vj}},$$

computed only when the denominator is nonzero. The raw M5 score, $\mathrm{M5}_{\mathrm{raw}}$, is the mean of MS_j over all scored features.

Raw M5 reflects two factors: (i) SAE-dependent image selection, i.e., which images a feature activates on, and (ii) evaluator-dependent embedding geometry, i.e., how tightly those selected images cluster in the evaluator's representation space. Even if image sets were sampled at random, different evaluation backbones can induce different expected pairwise cosine similarities because of anisotropy, class structure, and feature scaling in their embedding spaces. Raw M5 is therefore not a pure SAE property.

To correct for this evaluator-specific similarity floor, we estimate a backbone-specific geometric baseline using random pseudo-features. Concretely, we sample 1,000 random image groups, matching the same top-k size used for real features, from the exact embedded image pool used in scoring. Each pseudo-feature uses uniform pair weights, equivalently the unweighted mean pairwise cosine similarity. The baseline $\mathrm{M5}_{\mathrm{baseline}}$ is the average over these 1,000 pseudo-features and measures the expected similarity induced by the evaluator's embedding geometry alone.

We report normalized M5 as

$$\mathrm{M5}_{\mathrm{norm}} = \frac{\mathrm{M5}_{\mathrm{raw}} - \mathrm{M5}_{\mathrm{baseline}}}{1 - \mathrm{M5}_{\mathrm{baseline}}}.$$

This rescales the baseline-corrected score by the available cosine-similarity headroom above baseline, so normalized values can be interpreted as the fraction of possible improvement over the evaluator-specific random baseline. As a result, cross-backbone rankings can differ between raw and normalized M5: raw rankings include both SAE signal and evaluator-geometry bias, whereas normalized rankings more directly reflect how far SAE-selected image sets rise above evaluator-specific chance structure.

Within-class vs. between-class decomposition of M5. To test whether high M5 reflects strict class-pure selectivity or broader visual clustering, we decompose each feature's pairwise similarity into within-class and between-class terms using ImageNet validation labels. For a scored feature j with top-k images, let embeddings be $\{e_i\}_{i=1}^k$, labels $\{y_i\}_{i=1}^k$, and activation-derived weights $\{w_{ij}\}_{i=1}^k$ as defined above. Define pair weight $\omega_{uv} = w_{uj} w_{vj}$ and cosine similarity $c_{uv} = \cos(e_u, e_v)$ for $u < v$.

We partition pairs into

$$\mathcal{W}_j = \{(u,v) : y_u = y_v,\ u < v\}, \qquad \mathcal{B}_j = \{(u,v) : y_u \neq y_v,\ u < v\},$$

and compute

$$\mathrm{MS}_j^{\mathrm{within}} = \frac{\sum_{(u,v)\in\mathcal{W}_j} \omega_{uv} c_{uv}}{\sum_{(u,v)\in\mathcal{W}_j} \omega_{uv}}, \qquad \mathrm{MS}_j^{\mathrm{between}} = \frac{\sum_{(u,v)\in\mathcal{B}_j} \omega_{uv} c_{uv}}{\sum_{(u,v)\in\mathcal{B}_j} \omega_{uv}},$$

whenever the corresponding denominator is nonzero. We also compute the pair-mass fractions

$$\rho_j^{\mathrm{within}} = \frac{\sum_{(u,v)\in\mathcal{W}_j} \omega_{uv}}{\sum_{u<v} \omega_{uv}}, \qquad \rho_j^{\mathrm{between}} = 1 - \rho_j^{\mathrm{within}},$$

which yield the exact convex decomposition

$$\mathrm{MS}_j = \rho_j^{\mathrm{within}} \mathrm{MS}_j^{\mathrm{within}} + \rho_j^{\mathrm{between}} \mathrm{MS}_j^{\mathrm{between}}.$$

At the aggregate level, averaging over scored features, high-M5 features typically have elevated $\mathrm{MS}^{\mathrm{within}}$ and nontrivial $\mathrm{MS}^{\mathrm{between}}$. Thus, their top-activating images cluster strongly within class but can also remain highly similar across different classes. This supports the claim in §5.6 that M5 captures within-feature embedding-space consistency rather than class-pure selectivity or fine-grained discriminability.

L M6 cross-domain protocol

For each out-of-domain dataset, we train probes using only that dataset's own training split and evaluate on its held-out test split. No ImageNet-trained classifier is reused for M6. Instead, each representation is probed independently: raw backbone activations, SAE-reconstructed activations, and sparse SAE feature vectors each receive their own classifier trained from scratch under the same hyperparameters. For the CPU implementation we use scikit-learn logistic regression with L-BFGS, $C = 1.0$, and `max_iter` $= 500$; for the GPU implementation we use Adam with learning rate 10^{-2}, weight decay 10^{-4}, 50 epochs, and batch size 2048. The M6 preservation ratio is then computed as reconstructed-probe accuracy divided by raw-activation-probe accuracy on the same OOD dataset. This protocol makes the ratio an estimate of SAE-induced information loss under matched OOD supervision, rather than a measure of how well an ImageNet classifier transfers to the OOD label space. In practice the ratio is usually at or below 1, with values near 1 indicating that SAE reconstruction preserves nearly all task-relevant OOD information available in the raw representation. This protocol isolates SAE-induced loss rather than backbone capacity or ImageNet-to-OOD classifier transfer.

M Full Cross-Domain Probe Accuracies (M6)

This appendix reports the raw, reconstructed, and preserved linear-probe accuracies underlying the M6 (cross-domain preservation) score for all five backbones across the BatchTopK SAE sweep (n=60: $\{8\times, 16\times, 32\times\} \times \{k$=$64, 128, 192, 256\}$). For each dataset $\mathcal{D} \in \{\text{iNaturalist}, \text{CUB-200}, \text{DTD}\}$ we train matched linear probes on (i) raw backbone features $\mathrm{Acc}^{\mathcal{D}}_{\text{raw}}$ and (ii) SAE-reconstructed features $\mathrm{Acc}^{\mathcal{D}}_{\text{recon}}$, and report the preservation ratio $P^{\mathcal{D}} = \mathrm{Acc}^{\mathcal{D}}_{\text{recon}} / \mathrm{Acc}^{\mathcal{D}}_{\text{raw}}$.

Table 6 aggregates over the 12 SAE configurations per backbone (reporting mean $\pm$ std across the sweep). $\mathrm{Acc}^{\mathcal{D}}_{\text{raw}}$ is a property of the backbone alone and is therefore reported once per row.

Table 6: Cross-domain probe accuracies and preservation ratios. Recon and P columns report mean $\pm$ std across n=12 BatchTopK SAE configurations per backbone; raw is the backbone-only probe.

	iNaturalist (10K classes)			**CUB-200** (200 classes)			**DTD** (47 classes)		
Backbone	Raw	Recon	P	Raw	Recon	P	Raw	Recon	P
CLIP-B/16	0.093	0.074 ± 0.007	0.793 ± 0.078	0.575	0.517 ± 0.018	0.900 ± 0.032	0.793	0.773 ± 0.014	0.974 ± 0.018
SigLIP-B/16	0.119	0.101 ± 0.006	0.852 ± 0.050	0.693	0.658 ± 0.011	0.949 ± 0.016	0.836	0.825 ± 0.006	0.987 ± 0.007
DINOv2-B/14	0.235	0.210 ± 0.009	0.893 ± 0.037	0.831	0.817 ± 0.007	0.984 ± 0.008	0.831	0.816 ± 0.005	0.983 ± 0.007
DeiT-B/16	0.074	0.057 ± 0.008	0.770 ± 0.107	0.623	0.528 ± 0.052	0.847 ± 0.084	0.699	0.672 ± 0.009	0.961 ± 0.012
MAE-B/16	0.037	0.032 ± 0.002	0.864 ± 0.062	0.420	0.381 ± 0.020	0.908 ± 0.048	0.691	0.661 ± 0.014	0.957 ± 0.021

Observations. (i) Raw probe accuracy varies by an order of magnitude across datasets and backbones (e.g. MAE-B/16 reaches 3.7% on iNaturalist-10K vs. 69.1% on DTD), which is exactly why $P^{\mathcal{D}}$ rather than raw $\mathrm{Acc}_{\text{recon}}$ is the correct preservation signal. (ii) Preservation is consistently highest on DTD ($P \geq 0.957$ for all backbones) and lowest on iNaturalist, where the long-tailed 10K-way label space amplifies any reconstruction error. (iii) DeiT-B/16 shows the largest variance across SAE configurations (e.g. $P^{\text{CUB}} = 0.847 \pm 0.084$), indicating its representations are more sensitive to the choice of SAE width and sparsity than the other backbones; DINOv2-B/14 is the most stable ($P^{\text{CUB}} = 0.984 \pm 0.008$).

N M7 threshold sensitivity

We evaluate the sensitivity of M7 absorption to the overlap threshold α and size-ratio threshold β by sweeping $\alpha \in \{0.7, 0.8, 0.9\}$ and $\beta \in \{0.3, 0.5, 0.7\}$. Table 7 reports the per-backbone absorption rates. DeiT has the lowest absorption in 7 of 9 settings. The two exceptions occur at the strictest overlap threshold, $\alpha = 0.9$, with $\beta \in \{0.5, 0.7\}$, where MAE becomes lower than DeiT. This confirms that the DeiT absorption paradox is not an artifact of the default threshold choice, while also showing that the exact rank ordering is threshold-sensitive in the high-stringency regime.

Table 7: Sensitivity of M7 absorption rate to the absorption thresholds α and β. Lower is conventionally better. Bold indicates the lowest-absorption backbone for each threshold setting.

α	β	DINOv2	CLIP	SigLIP	DeiT	MAE
0.7	0.3	0.385	0.562	0.502	**0.155**	0.521
0.7	0.5	0.294	0.404	0.408	**0.121**	0.291
0.7	0.7	0.098	0.200	0.208	**0.038**	0.106
0.8	0.3	0.513	0.653	0.642	**0.332**	0.645
0.8	0.5	0.423	0.494	0.547	**0.298**	0.415
0.8	0.7	0.223	0.291	0.347	**0.211**	0.226
0.9	0.3	0.725	0.796	0.796	**0.604**	0.785
0.9	0.5	0.634	0.638	0.702	0.570	**0.555**
0.9	0.7	0.434	0.434	0.502	0.483	**0.366**

O The Codes-lift Inversion

Across the four downstream tasks where we evaluate both raw features and SAE codes (retrieval R@1, CUB-200, DTD, segmentation), we observe a consistent and counterintuitive pattern: backbones with weaker raw-feature performance gain proportionally more from SAE decomposition than backbones with stronger raw-feature performance. We refer to this as the *codes-lift inversion*.

Setup For each (backbone, task) pair, we compute the lift ratio $L = \mathrm{perf}(\text{SAE codes})/\mathrm{perf}(\text{raw features})$, averaged across the 12 SAE configurations per backbone. We then compute the Spearman rank correlation between mean raw performance and mean lift across the five backbones ($n = 5$).

Results On three of four tasks, the correlation is strongly negative (Table 8).

Task	$\rho(\text{raw}, L)$	p	Range of lifts
Retrieval R@1	-0.900	0.037	DeiT 0.99 to MAE 1.18
DTD	-0.900	0.037	SigLIP 1.02 to DeiT 1.06
CUB-200	-0.700	0.188	DINOv2 0.99 to MAE 1.19
Segmentation mIoU	$+0.500$	0.391	DeiT 0.43 to DINOv2 0.75

Table 8: Cross-backbone Spearman correlation between raw-feature performance and SAE codes lift. Negative correlations indicate the inversion: weaker raw backbones gain more from SAE decomposition. Segmentation is the lone exception.

The magnitude of the inversion is substantial. MAE, the worst raw performer on retrieval (R@1 $= 0.335$), exhibits a 17.7% codes lift, while DeiT, the best raw retrieval performer (R@1 $= 0.692$), shows a slight codes regression (lift $= 0.988$). On CUB-200, MAE gains 19.4% from SAE decomposition while DINOv2 loses 1.4%. The pattern is stable across all 12 configurations within each backbone: per-config lift standard deviations are small (between 0.006 for DTD on DeiT and 0.094 for segmentation on MAE), indicating the effect is a property of the backbone rather than configuration noise.

Segmentation is the lone exception. The correlation flips sign to $+0.50$ and no backbone reaches a lift above 1.0 (every backbone loses mIoU under SAE decomposition, with DINOv2 retaining the largest fraction at 0.75 and DeiT the smallest at 0.43). This dissociation matches the broader spatial-coherence finding from Section 5.1: the spatial structure is a property of the backbone that SAE decomposition cannot manufacture, only preserve to varying degrees.

Interpretation We interpret the inversion as a ceiling effect on SAE-extractable concept structure rather than as evidence that weaker backbones are secretly better. A backbone with raw representations that already align well with task labels has less headroom for the SAE's sparse decomposition to add discriminative structure on top. A backbone with weaker raw alignment offers more room for the sparse code to contribute task-relevant axes, and the BatchTopK reconstruction loss appears to

capture some of that available signal. The inversion does not hold for spatial tasks because the SAE's sparse codes are reconstructed from the backbone's feature space, and patch-level spatial coherence is a geometric property of that space which decomposition cannot create.

Practical implications The inversion has two consequences for practitioners. First, the relative gap between backbones narrows substantially after SAE decomposition on retrieval, fine-grained classification, and texture, which weakens the case for selecting a backbone purely on raw downstream performance when the downstream consumer is the SAE codes themselves. Second, the inversion does not generalize to spatial tasks; for segmentation and dense prediction, raw-feature ranking is preserved through SAE decomposition and DINOv2 retains its dominance regardless of lift considerations.

Limitations The result is computed across $n = 5$ backbones, so the ρ values should be read as suggestive of an effect rather than tight estimates of its magnitude. Per-backbone variance across the 12 SAE configurations is small (typically below 0.05 in lift units), but the cross-backbone trend would benefit from replication on additional pretraining paradigms before being treated as a general law. We do not have a mechanistic account of why the inversion is task-dependent in the specific pattern observed, and we leave a feature-level intervention study to future work.

P Practitioner guide for backbone selection

This appendix condenses the empirical task-to-backbone mapping into a single-page reference. Recommendations are derived from per-backbone means across 12 SAE configurations on the 1.28M ImageNet training split, and the supporting capability metric is the one with the highest cross-backbone Spearman rank correlation to the task in our pre-registered analysis (Section **??**, $n = 5$ backbones).

P.1 Decision summary

Figure 8 provides the top-level routing. In text form:

Dense spatial prediction. If the task involves semantic segmentation or dense feature transfer to spatial tasks, use DINOv2. The supporting metric is M1 (Moran's I), which correlates with raw segmentation mIoU at $\rho = 0.90$, $p = 0.037$ across backbones. DINOv2's mean raw mIoU on ADE20K is 0.171 versus 0.058 for the next backbone (CLIP), a 2.9$\times$ lead that is preserved through SAE decomposition (codes mIoU 0.127 versus 0.036).

Fine-grained concept discrimination. If the task involves CUB-200, iNaturalist transfer, or object classification with many similar classes, use DINOv2 for raw features and either DINOv2 or SigLIP for SAE codes. The supporting metric is M4 (Sparse Probing AUC), which correlates with iNat at $\rho = 0.90$ and CUB-200 at $\rho = 0.80$. DINOv2 leads CUB raw accuracy at 0.831 with SigLIP at 0.693; on SAE codes the gap narrows to 0.819 versus 0.748.

Texture and material classification. If the task involves DTD or analogous datasets, use SigLIP. The supporting metric is again M4, with $\rho = 1.00$ to DTD ($p = 0.017$, $n = 5$). SigLIP's raw DTD accuracy is 0.836 with DINOv2 essentially tied at 0.831; on SAE codes SigLIP edges DINOv2 by 0.002. The earlier prediction that masked-pretraining (MAE) would win on texture did not hold.

Image-image retrieval. If the task is retrieval (R@1, R@5, R@10) and interpretability is secondary, use DeiT. The supporting metric is M3 (Preservation), which correlates with retrieval at $\rho = 0.60$. DeiT achieves R@1 = 0.692 on raw features and 0.684 on SAE codes, well ahead of other backbones. We note a caveat: DeiT exhibits the worst sparse probing performance (0.418) and a high feature absorption rate, so applications requiring both retrieval accuracy and feature interpretability should consider DINOv2 (R@1 = 0.548) or SigLIP (R@1 = 0.526) as more balanced alternatives.

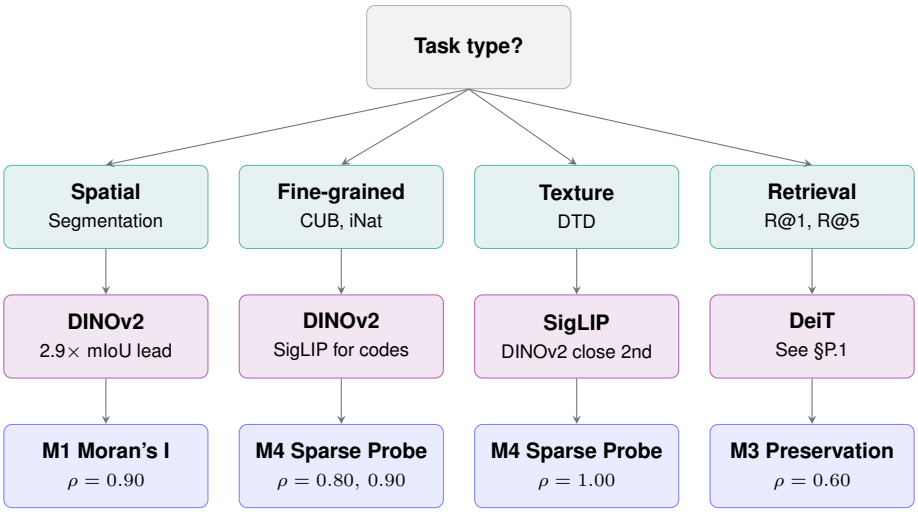


Figure 8: Backbone selection flowchart. Each column maps a task family (teal) to the recommended backbone (purple) and the supporting capability metric (blue), with ρ values from cross-backbone Spearman correlation on $n = 5$ backbones. Caveats and limitations are discussed in the surrounding text.

P.2 Backbones we do not recommend

MAE places last or near-last on every raw downstream task in our suite. We retain it in the benchmark as a reference point for the masked-pretraining paradigm and to enable backbone-specific failure-mode reporting, but it is not a recommended choice for any of our five task families.

P.3 What not to select on

M2 (Fractional Variance Unexplained) correlates negatively or near-zero with every downstream task we evaluate (retrieval 0.00, CUB -0.30, DTD -0.70, segmentation -0.50, iNat -0.60). Selecting a backbone on FVU alone is the most consistent way to make a poor choice in our results. M3 (Preservation) is a usable proxy when M1 or M4 cannot be computed, but it is weaker than either of those two metrics on every task except retrieval.

P.4 Codes-lift caveat

Practitioners selecting a backbone for SAE-codes consumption should expect the relative gap between backbones to compress on retrieval, CUB, and DTD, but not on segmentation. We document this effect in detail in Appendix [Rajamanoharan et al., 2024]. The strongest raw backbone is not always the strongest SAE-codes backbone, and the codes ranking should be checked directly against the target task before committing to a choice.

P.5 Caveat on sample size

All cross-backbone correlations are computed across $n = 5$ backbones. Three of four pre-registered hypotheses pass at $\rho \geq 0.7$ with $p \leq 0.05$, and one (M3 to retrieval) is directionally consistent but does not reach significance. The guidance above should be read as evidence-grounded but not as a substitute for evaluating SAE features on the practitioner's actual target task when stakes are high.

Q Architectural robustness: do these findings hold under JumpReLU?

Setup. We trained JumpReLU SAEs [Rajamanoharan et al., 2024] on the same five backbones at expansion factors $16\times$ and $32\times$ and target L_0 values of 64 and 128, using a uniform recipe (Silverman kernel, bandwidth 10^{-3}, adaptive sparsity coefficient with multiplicative controller). Of the 20 resulting configurations, 17 settled within ± 2 of their target L_0 and are reported here. Three

configurations (SigLIP 16× at L_0=128, SigLIP 32× at L_0=128, DINOv2 32× at L_0=128) failed to converge to the target L_0 within the one-epoch training budget under the uniform recipe and are excluded; we believe these reflect known JumpReLU sensitivity to controller hyperparameters rather than fundamental incompatibility, and leave per-backbone tuning to future work.

Cross-family rank consistency. Table 9 reports the Spearman rank correlation across the five backbones between BatchTopK and JumpReLU mean values for each of the seven metrics. Correlations range from 0.80 (M6) to 1.00 (M2, M5), with six of seven significant at $p < 0.05$ despite the small sample size of $n = 5$ backbones. DINOv2 ranks first overall under both SAE families and MAE ranks last under both. We interpret this as evidence that the cross-backbone findings reported in the main text reflect properties of the underlying vision backbones rather than artifacts of the BatchTopK SAE family.

Absolute differences. Although ordering is preserved, absolute values differ between SAE families. JumpReLU exhibits uniformly higher FVU than BatchTopK at matched L_0 (mean increase 0.037 across backbones), and uniformly lower sparse probing AUC (mean decrease 0.075). These differences are consistent across all five backbones, indicating they reflect a property of the SAE family rather than an interaction with any specific backbone.

A note on CLIP. The single most striking cross-family difference concerns dead-feature rates. At matched 16× configurations, CLIP under BatchTopK has 30.6% dead features on average, compared to 9.5% under JumpReLU; the other four backbones differ by less than 5 percentage points between families. Across the full BatchTopK sweep, CLIP dead rates range from 20.8% to 47.0%, an outlier relative to the other backbones (DINOv2: 0.0% to 6.9%; SigLIP: 2.9% to 13.6%). We do not have a definitive explanation but note that this interaction between CLIP's representation and the BatchTopK forced-allocation mechanism warrants further investigation.

Sensitivity check. Excluding CLIP 32× configurations (where dead-feature rates are highest) from the per-backbone aggregation leaves all seven cross-family rank correlations unchanged to three decimal places, indicating the robustness conclusion does not depend on the dead-feature outlier.

Table 9: Cross-family rank consistency across the five backbones. Rankings are best to worst per metric (lower-is-better metrics inverted accordingly). Spearman ρ computed on the per-backbone means across all configurations satisfying $|L_0 - \text{target}| \leq 2$. $^*p < 0.05$.

Metric	BatchTopK ranking	JumpReLU ranking	ρ	p
M1 Localization	DINOv2, DeiT, CLIP, SigLIP, MAE	DINOv2, CLIP, DeiT, SigLIP, MAE	0.900^*	0.037
M2 FVU	CLIP, SigLIP, DINOv2, DeiT, MAE	CLIP, SigLIP, DINOv2, DeiT, MAE	1.000^*	0.000
M3 Preservation	DINOv2, DeiT, MAE, SigLIP, CLIP	DINOv2, DeiT, SigLIP, MAE, CLIP	0.900^*	0.037
M4 Sparse Probing	SigLIP, DINOv2, CLIP, DeiT, MAE	DINOv2, SigLIP, CLIP, DeiT, MAE	0.900^*	0.037
M5 Norm. MS	CLIP, SigLIP, DINOv2, DeiT, MAE	CLIP, SigLIP, DINOv2, DeiT, MAE	1.000^*	0.000
M6 iNat	SigLIP, DINOv2, CLIP, MAE, DeiT	DINOv2, SigLIP, CLIP, DeiT, MAE	0.800	0.104
M7 Absorption	DeiT, DINOv2, MAE, SigLIP, CLIP	DeiT, DINOv2, MAE, CLIP, SigLIP	0.900^*	0.037