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

Vijayrajsinh Gohil [*1] **Diwei Sheng** [*1] **Chen Feng** [1]

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 on HuggingFace and the ViSAEBench evaluation library on GitHub.

[*]Equal contribution [1]New York University, Brooklyn, USA. Correspondence to: Vijayrajsinh Gohil <vg2097@nyu.edu>.

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

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., 2024a; DeepMind, 2025). 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, 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

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., 2024b). Recent work has extended SAEs to cross-modal settings (Gu et al., 2026; Zhao et al., 2026). 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.

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). The ViT-Prisma library (Joseph et al., 2025) established infrastructure for training SAEs on ViT activations. Parallel efforts have explored 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.

Vision representation benchmarks and spatial structure. 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. Spatial structure 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 Moran's I (Moran, 1950), the standard spatial-autocorrelation statistic in spatial statistics, which to our knowledge has not previously 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 & 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. We adopt BatchTopK because visual complexity varies substantially across natural images. The training objective is:

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

To verify that cross-backbone findings are not specific to BatchTopK, we replicate the full evaluation on JumpReLU SAEs (Rajamanoharan et al., 2024b) across all five backbones. Cross-family Spearman rank correlations range from 0.80 (M6) to 1.00 (M2, M5), with six of seven significant at $p < 0.05$: DINOv2 ranks first and MAE last under both SAE families on every metric. Absolute values differ between families (JumpReLU shows higher FVU and lower sparse probing AUC at matched L0), but the backbone ordering is preserved throughout, indicating that our central findings reflect properties of the pretrained backbone rather than the BatchTopK training objective. Full results are in Appendix Q.

ViT backbones evaluated. We evaluated five ViT-Base backbones spanning 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 (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 in Appendix A. The full sweep yields 60 SAEs ($5 \times 3 \times 4$), all released on HuggingFace.

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. We define the per-feature, per-image Moran's I (Moran, 1950) as

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

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 $\varepsilon = 10^{-8}$, 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 valid images. The SAE-level statistic is the mean across evaluable features:

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

where $\mathcal{E}$ is the set of evaluable features and $N_{\text{eval}} = |\mathcal{E}|$. We report $\bar{I}_{\text{SAE}}$ 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 $|z| > 2$; the exact z-statistic is given in Appendix F. We use mean Moran's I (not 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) (Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2024; Lieberum et al., 2024a; Karvonen et al., 2025). 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:

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

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:

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

where s_{baseline} is the mean pairwise cosine similarity of N randomly sampled validation images under the cross-model embedding. M5=0 means the feature's top images are no more similar than random; 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 (cha) 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.

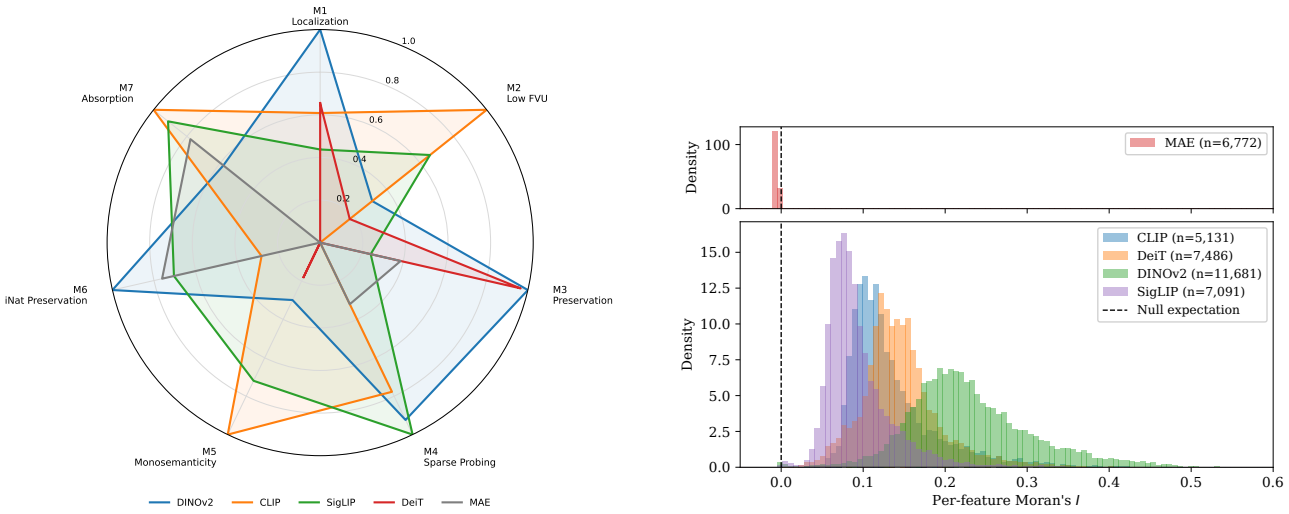
Definition. For each ImageNet hierarchy sibling group with at least three classes, we construct class-versus-sibling probing tasks. SAE codes are pooled to image-level features, and features are ranked by a univariate class-discrimination score on the probe-training split. For each task, we train sparse 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, each linking one SAE-internal metric to a downstream task through an explicit mechanism, before computing any cross-task correlations. Each hypothesis is tested cross-backbone with $n = 5$ by aggregating the SAE-internal metric across the 12 configurations per backbone and correlating it against a per-backbone downstream number, using Spearman's ρ as the test statistic. Confirmation thresholds



(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.

Table 1. The four pre-registered hypotheses. Each links one SAE-internal metric to a downstream task through a stated mechanism, with a confirmation threshold on the cross-backbone Spearman correlation fixed in advance.

	SAE metric → task	Mechanism	Threshold
H1	M1 → ADE20K segmentation mIoU	Spatially coherent features carry the structure a segmentation head exploits	$\rho \geq 0.7$
H2	Norm. M5 → CUB-200 fine-grained accuracy	Within-feature consistency under an independent embedding should track semantically coherent content	$\rho \geq 0.6$
H3	M3 → image-image retrieval R@1	Retrieval is a near-direct test of representation similarity	$\rho \geq 0.6$
H4	M4 → iNaturalist sparse probing AUC	Concept-discriminative features on ImageNet should remain so on related natural-image tasks	$\rho \geq 0.7$

were fixed in advance and are reported alongside each hypothesis in Table 1.

For H4, CUB-200 and DTD are reported as additional cross-domain checks. 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.

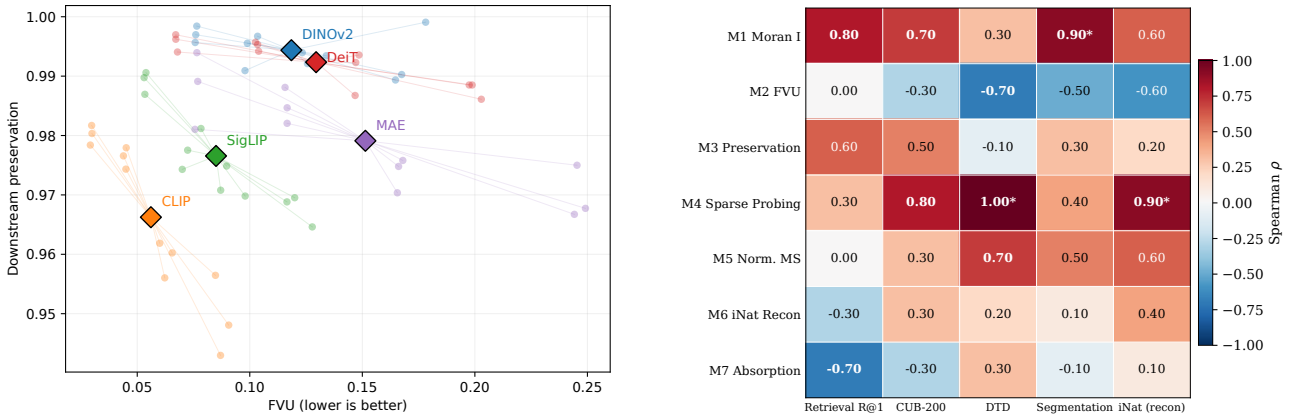
5. Experiments and Findings

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: $\mathrm{M1} = -0.005$, matching the



(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.

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 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 FVU $= 0.129$. DINOv2 reaches 0.994 preservation at 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 $\rho = 0.60$ ($p = 0.285$), meet-

ing 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.

The failure is not incidental to our particular operationalization; it follows from three structural features of within-feature consistency as a quality signal. First, M5 contains no relational term. M5 scores each feature in isolation, as the consistency of its own top-activating set. Fine-grained discrimination is intrinsically relational: it depends on how feature responses *differ* between neighbouring classes, not on how internally coherent any single feature is. A metric defined entirely on a feature's marginal activation set therefore has no quantity that rewards between-class separation, so high consistency and high discriminability can co-occur or dissociate freely. Second, consistency is measured at the wrong granularity. M5 is scored in an independent evaluator's embedding space, whose geometry is organized around the coarse semantic structure that backbone was trained to capture. Two images from different fine-grained classes that share coarse semantics—two bird species on similar perches—sit close together in that space. "Consistency" under the evaluator thus rewards coarse perceptual similarity, which is largely orthogonal to the sub-category distinctions a fine-grained probe must recover.

The third mechanism is made explicit by the convex decomposition of M5. We decompose each feature's score into within-class and between-class components by computing mean pairwise cosine similarity within the same ImageNet validation class versus across ImageNet classes, yielding $\mathrm{M5}_j = \rho_j^{\text{within}}\mathrm{MS}_j^{\text{within}} + \rho_j^{\text{between}}\mathrm{MS}_j^{\text{between}}$ (Appendix K). A feature can saturate M5 through the between-class term alone: if its top images are perceptually similar but span many classes, $\mathrm{MS}^{\text{between}}$ is high and the feature reads as monosemantic while contributing nothing to class separation. High-M5 features often retain substantial between-class similarity, confirming that M5 captures visual consistency but not necessarily fine-grained class separability. 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.

These mechanisms explain why the dissociation surfaces on fine-grained rather than coarse tasks. Coarse classification tolerates broadly-firing features, since the relevant distinctions are large and redundantly encoded. Fine-grained discrimination depends on precisely the small inter-class differences that a broadly-coherent feature averages over, so the gap between consistency and utility is widest exactly where M5 is most likely to be cited as evidence of quality.

This finding has implications for SAE evaluation practice. The broader lesson is that within-feature consistency is necessary but not sufficient for a feature set to serve as a good basis for downstream tasks: a dictionary of perceptually coherent features can still be a poor coordinate system if those features are coarse, redundant, or non-discriminative. 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, where the analogous between-token or between-context separability has not been systematically tested against within-feature consistency. 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)

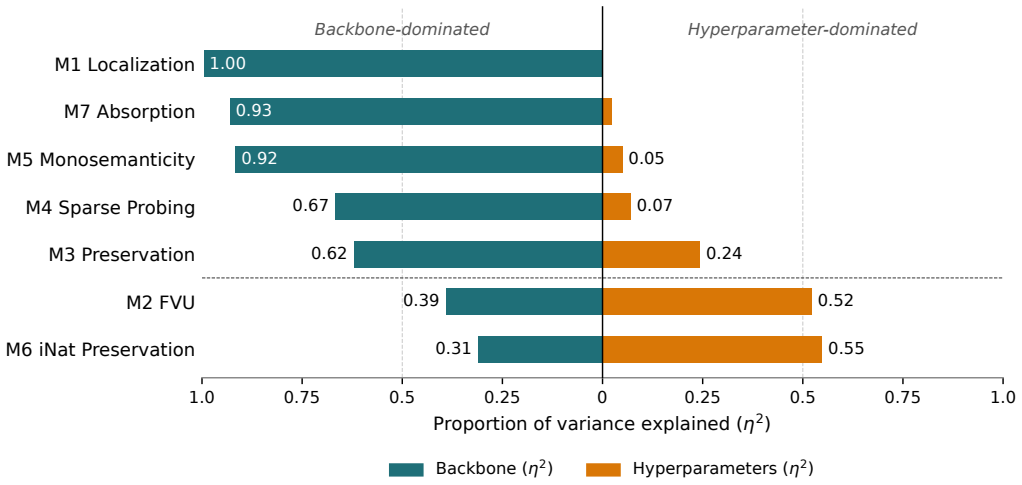


Figure 3. Variance decomposition of each metric across the 60 SAE configurations.

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 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 (Rajamanoharan et al., 2024c), Gated SAE (Rajamanoharan et al., 2024a), Matryoshka SAE (Bussmann et al., 2025)) 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. The cross-backbone rank ordering is further replicated under JumpReLU SAEs, with cross-family Spearman correlations of at least 0.80 on all seven metrics, supporting the generality of these findings beyond the BatchTopK training objective.

Acknowledgements

This work is supported by NSF Grant 2238968. We thank the NYU HPC team for their assistance and support.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning interpretability. ViSAEBench is a diagnostic tool for evaluating sparse representations of vision models. The primary positive impact is improved transparency and comparability of SAE evaluation methodologies, which can support safer and more accountable AI systems. We do not foresee direct negative societal consequences from this benchmark, as it evaluates existing models rather than introducing new generative capabilities.

References

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. https://transformer-circuits.pub/2023/monosemantic-features/index.html, 2023.

Bussmann, B., Leask, P., and Nanda, N. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

Bussmann, B., Nabeshima, N., Karvonen, A., and Nanda, N. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Chatzoudis, G., Polyzos, K. D., Li, Z., Gu, D., Moran, G. E., Wang, H., and Metaxas, D. N. Can cross-layer transcoders replace vision transformer activations? an interpretable perspective on vision, 2026. URL https://arxiv.org/abs/2604.13304.

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

DeepMind, G. Gemma scope 2. https://huggingface.co/google/gemma-scope-2, 2025.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Gu, D., Gao, Y., Chatzoudis, G., Dong, Z., Zhang, G., Guo, B., Zhou, Y., Zhou, M., and Metaxas, D. LUCID-SAE: Learning unified vision-language sparse codes for interpretable concept discovery, 2026. URL https://arxiv.org/abs/2602.07311.

Han, S., Kim, Y., and Kwak, N. Causal interpretation of sparse autoencoder features in vision, 2025. URL https://arxiv.org/abs/2509.00749.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. 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.

Joseph, S., Suresh, P., Hufe, L., Stevinson, E., Graham, R., Vadi, Y., Bzdok, D., Lapuschkin, S., Sharkey, L., and Richards, B. A. Prisma: An open source toolkit for mechanistic interpretability in vision and video, 2025. URL https://arxiv.org/abs/2504.19475.

Karvonen, A., Rager, C., Lin, J., Tigges, C., Bloom, J., Chanin, D., Lau, Y.-T., Farrell, E., McDougall, C., Ayonrinde, K., Till, D., Wearden, M., Conmy, A., Marks, S., and Nanda, N. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability, 2025. URL https://arxiv.org/abs/2503.09532.

Korznikov, A., Galichin, A., Dontsov, A., Rogov, O., Tutubalina, E., and Oseledets, I. OrtSAE: Orthogonal sparse autoencoders uncover atomic features, 2025. URL https://arxiv.org/abs/2509.22033.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., and Nanda, N. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 1st Workshop on Behavioral ML*. Association for Computational Linguistics, 2024a. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., and Nanda, N. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024b. URL https://arxiv.org/abs/2408.05147.

Makhzani, A. and Frey, B. A winner-take-all method for training sparse convolutional autoencoders. In *NIPS Deep Learning Workshop*, 2014.

Moran, P. A. P. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950. ISSN 0006-3444. URL http://www.jstor.org/stable/2332142.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Pach, M., Karthik, S., Bouniot, Q., Belongie, S., and Akata, Z. Sparse autoencoders learn monosemantic features in vision-language models, 2025. URL https://arxiv.org/abs/2504.02821.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., and Nanda, N. Improving dictionary learning with gated sparse autoencoders, 2024a.

Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., and Nanda, N. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders, 2024b. URL https://arxiv.org/abs/2407.14435.

Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., and Nanda, N. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders, 2024c.

Rubinstein, A., Prabhu, A., Bethge, M., and Oh, S. J. Are we done with object-centric learning?, 2025. URL https://arxiv.org/abs/2504.07092.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E. J., Casper, S., Tegmark, M., Saunders, W., Bau, D., Todd, E., Geiger, A., Geva, M., Hoogland, J., Murfet, D., and McGrath, T. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html, 2024. Accessed: 2026-05-03.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention, 2021. URL https://arxiv.org/abs/2012.12877.

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., and Houlsby, N. A large-scale study of representation learning with the visual task adaptation benchmark, 2020. URL https://arxiv.org/abs/1910.04867.

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

Zhao, T. Z., Kiblawi, S., Yang, J., Usuyama, N., Tan, R., Codella, N. C., Naumann, T., Poon, H., and Wei, M. Learning sparse visual representations via spatial-semantic factorization, 2026. URL https://arxiv.org/abs/2602.01905.

A. Experimental Setup and Hyperparameters

We report the full hyperparameters used for SAE training and evaluation in Tables 2 and 3. 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 2. 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 normalisation	Per-dimension mean subtraction and scalar standard-deviation division
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 3. Evaluation hyperparameters used for ViSAEBench metrics (M1–M7).

Metric	Hyperparameters / Protocol
M1 (Spatial coherence)	Grid inferred from patch count (14×14 for 196 patches, 16×16 for 256 patches); 8-neighbour connectivity; minimum active patches per image $= 5$; activation variance threshold $\varepsilon = 10^{-8}$; minimum valid images per feature $= 50$; SAE encoding batch size $= 512$.
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)	Cached ImageNet-1K validation activations, internal stratified 80/20 split, seed $= 42$; image-level features via max-pooling; multinomial logistic regression with `lbfgs`, $C = 1.0$, `max_iter` $= 1000$.
M4 (Sparse probing)	Cached ImageNet-1K validation activations, 80/20 split; SAE codes max-pooled; features ranked by ANOVA F-statistic; $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$; main score at $k = 128$.
M5 (Monosemanticity)	Top-N activating validation images per feature with $N = 16$; up to 2048 live features scored; independent cross-backbone embedding model; baseline uses 1000 random pseudo-features.
M6 (Cross-domain)	Evaluated on iNaturalist, CUB-200, and DTD; probes trained on dataset's own training split; preservation = reconstructed-probe accuracy ÷ raw-feature accuracy.
M7 (Probe-based absorption)	Max-pooled SAE codes; ImageNet hierarchy sibling groups with minimum size $= 3$; default thresholds $\alpha = 0.8$, $\beta = 0.5$; sensitivity sweep in Appendix N.

B. Backbone and Hyperparameter Variance Decomposition

Table 4. 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 Pres.	0.311	0.549	0.141

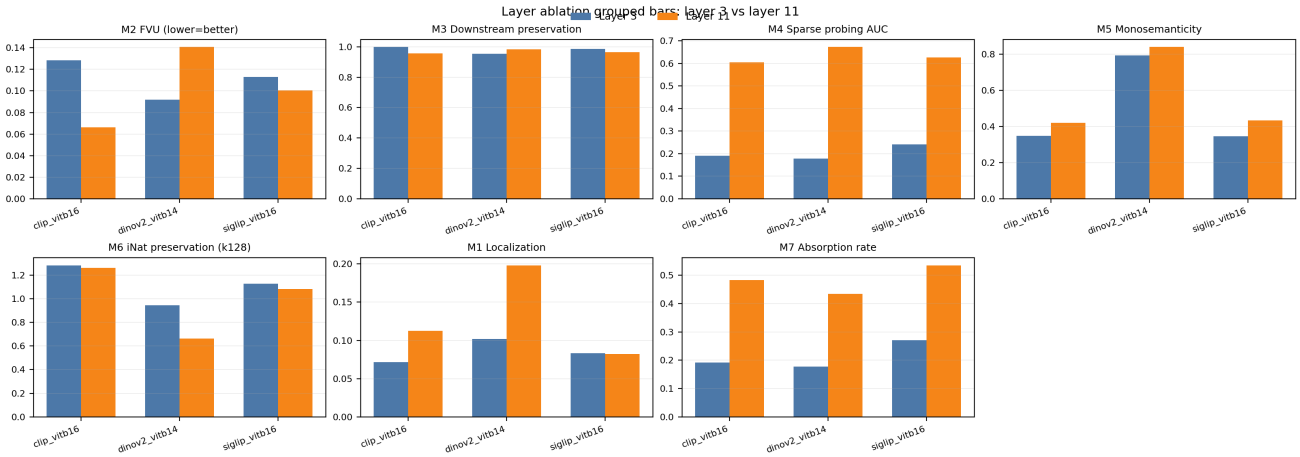


Figure 4. **Layer ablation comparing early-layer and late-layer SAEs.** We compare SAEs trained on layer 3 and layer 11 across seven ViSAEBench metrics for CLIP, DINOv2, and SigLIP. The comparison shows that some metrics are relatively stable across layers, while others change substantially with layer 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.

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.

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 produce large z-values; 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 5. 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

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 $16\times$, $K = 192$ 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

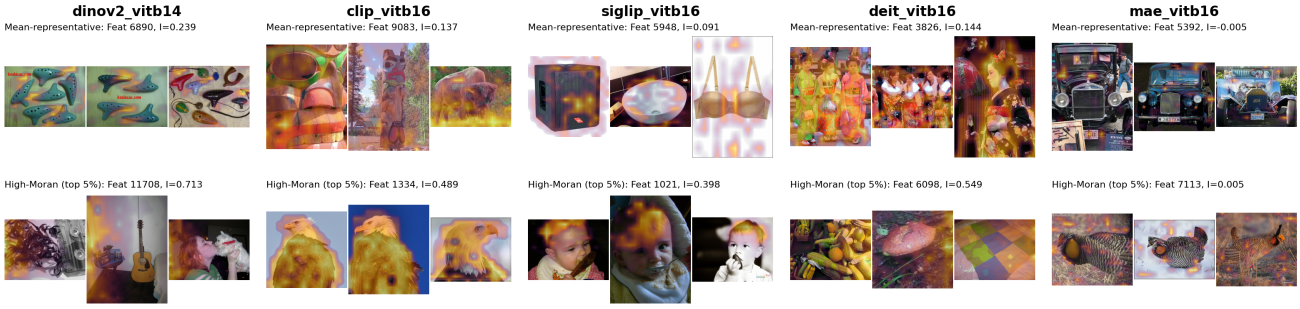


Figure 5. **Qualitative audit of M1 spatial coherence across backbones.** Each column corresponds to one pretrained backbone and shows two SAE features from the BatchTopK 16×, $K = 192$ configuration. The first row shows a mean-representative feature, selected as the evaluable feature whose per-feature 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.

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.

Figure 6 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.

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

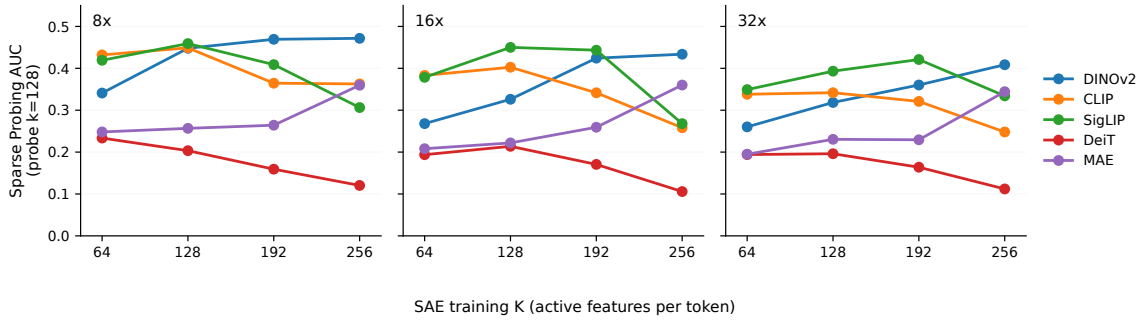


Figure 6. **Sparse probing curves.** ImageNet sparse-probing AUC across SAE training sparsity K and expansion factors. DINOv2 and SigLIP consistently support stronger sparse concept readout than MAE and DeiT across the sweep.

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.

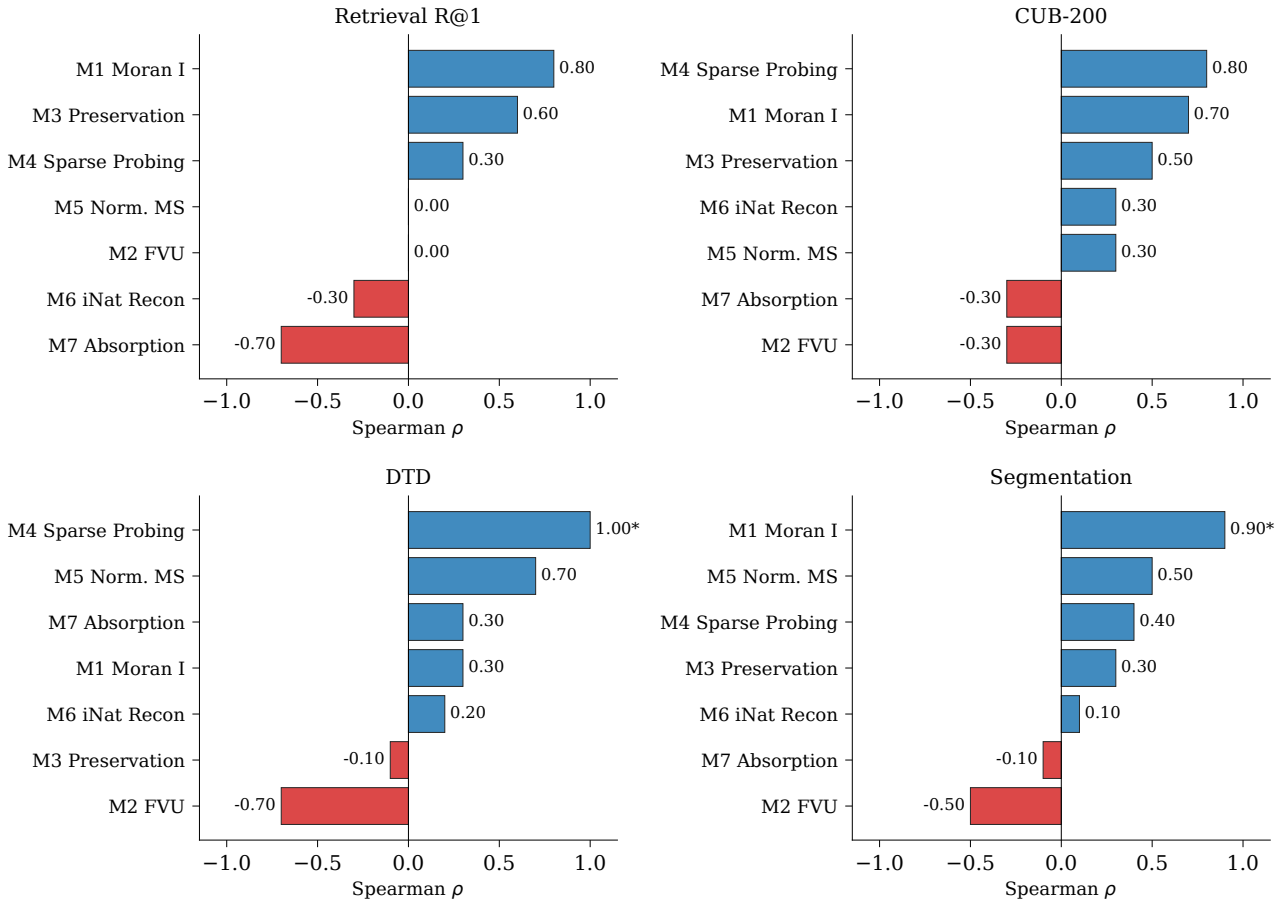


Figure 7. **Per-task metric correlations.** Cross-backbone Spearman correlations between ViSAEBench metrics and downstream tasks. M4 sparse probing best predicts CUB-200 and DTD performance, while M1 spatial coherence best predicts segmentation.

K. M5 monosemanticity protocol

Table 6. 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 6, 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 $\mathrm{Acc}_{\text{raw}}^{\mathcal{D}}$ and (ii) SAE-reconstructed features $\mathrm{Acc}_{\text{recon}}^{\mathcal{D}}$, and report the preservation ratio $P^{\mathcal{D}} = \mathrm{Acc}_{\text{recon}}^{\mathcal{D}}/\mathrm{Acc}_{\text{raw}}^{\mathcal{D}}$.

Table 7 aggregates over the 12 SAE configurations per backbone (reporting mean $\pm$ std across the sweep). $\mathrm{Acc}_{\text{raw}}^{\mathcal{D}}$ is a property of the backbone alone and is therefore reported once per row.

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 ($P^{\text{CUB}} = 0.984 \pm 0.008$).

Table 7. 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

Table 8. 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**

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 8 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.

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 9).

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 9. 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 ($n = 5$ 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.

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., 2024b). 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., 2024b) 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 $16\times$ at L_0=128, SigLIP $32\times$ at L_0=128, DINOv2 $32\times$ 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 10 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

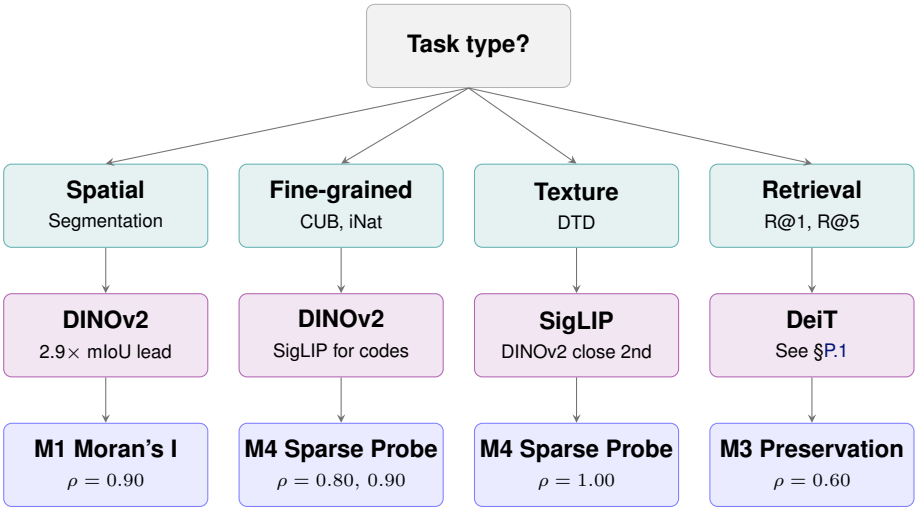


Figure 8. Backbone selection flowchart. Each column maps a task family (teal) to the recommended backbone (purple) and the supporting capability metric (blue), with ρ values from cross-backbone Spearman correlation on $n = 5$ backbones. Caveats and limitations are discussed in the surrounding text.

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 $16\times$ 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 $32\times$ 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 10. 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