CAPABILITY ≠ INTERPRETABILITY: HUMAN INTERPRETABILITY OF VISION FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

How interpretable are the features of leading vision models? The question is increasingly pressing as these models move from research benchmarks into high-stakes deployments, yet existing methods cannot answer it reliably. We address this gap with a framework for measuring and comparing the human interpretability of vision models, built around two complementary psychophysics protocols: (1) *localizability*—can an observer predict *where* a feature fires on a novel image?—and (2) *semanticity*—can an observer accurately describe *what* the feature represents? Features are recovered via sparse autoencoders, and a chance-anchored scoring function places every model on a common scale. Applying the framework to six vision transformers—two supervised ViTs and four foundation models (DINOv2, DINOv3, CLIP, SigLIP)—we collected more than $15{,}000$ behavioral responses, of which we analyzed $13{,}400$ responses from the 377 participants who passed our pre-specified quality checks. We find that foundation models are consistently *less* interpretable than their size-matched supervised counterparts, and the gap is not a capability tradeoff: interpretability does not correlate with downstream task performance on any benchmark we examine. What does correlate is the *locality* of a feature's activations and *coarse-grained* semantic alignment with humans—models with focal activations and representations that reflect the world's broad categorical structure produce more interpretable features, whereas fine-grained perceptual alignment does not. The two protocols yield strongly correlated rankings and share the same predictors, establishing interpretability as an independent, measurable dimension of representation quality—and, surprisingly, one on which the foundation models we tested generally fall below their size-matched supervised baseline. Capability alone cannot close that gap, but locality and coarse-grained alignment might.

1 INTRODUCTION

The dominant paradigm in modern computer vision is to pretrain large foundation models and fine-tune them for specific tasks. Built on the vision transformer (Dosovitskiy et al., 2021), models like CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), DINOv2 (Oquab et al., 2024), and DINOv3 (Siméoni et al., 2026) already serve as general-purpose visual backbones across an ever-expanding range of downstream tasks, including classification (Chiu et al., 2024), segmentation (Liu et al., 2024), object tracking (Faber et al., 2024; Tumanyan et al., 2024), and robotic perception (Kim et al., 2025). The same models are increasingly deployed in high-stakes settings, such as clinical decision support (Baharoon et al., 2023; Moutakanni et al., 2026; Jiao et al., 2026), and autonomous driving (Gao et al., 2025a; Jiang et al., 2025), where a user often needs to understand what a model is actually responding to. How interpretable, then, are the features that today's leading vision models have learned?

Surprisingly little work compares vision models along this dimension, because existing evaluation protocols cannot place them on a common, reliable scale of feature interpretability (Section 2).

To fill this gap, we introduce a model evaluation framework comprising two psychophysics protocols that quantify two complementary dimensions of feature interpretability: (1) feature *localizability*, i.e., can an observer predict *where* a feature fires on a novel image?; and (2) feature *semanticity*, i.e., can an observer accurately describe *what* a feature represents? Together they form a comprehensive evalu-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

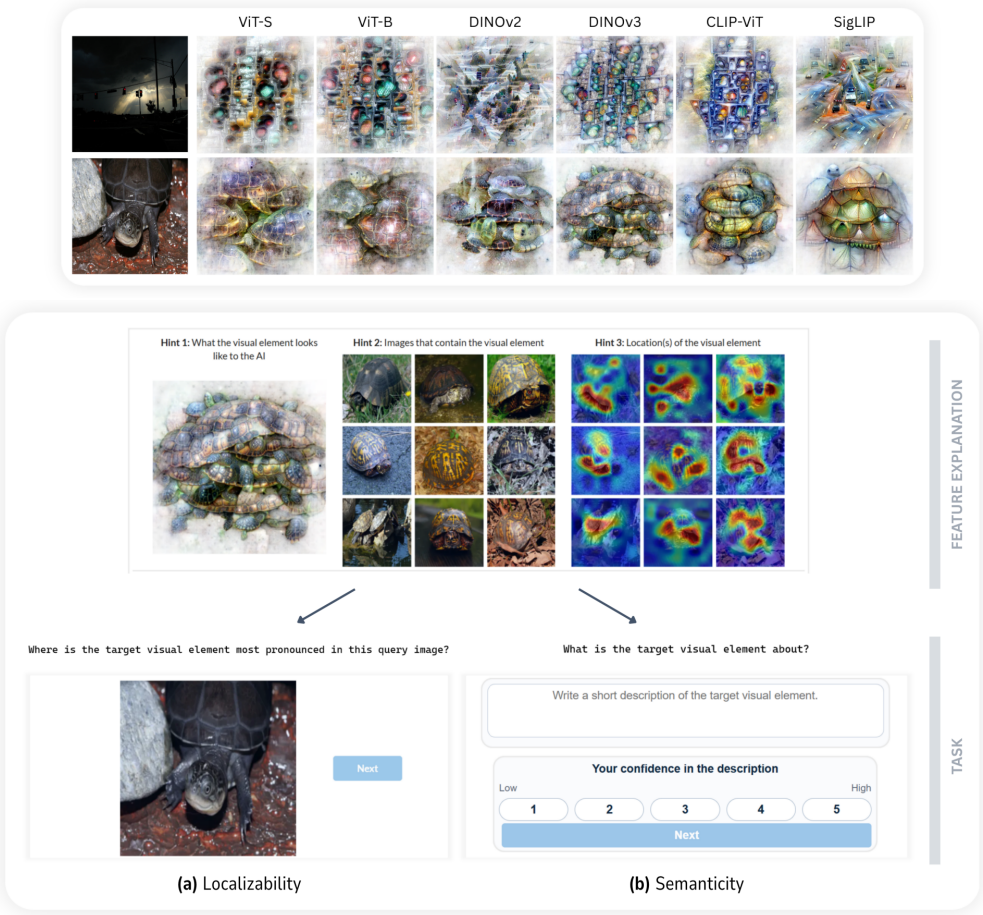


Figure 1: Models trained under different objectives learn different internal representations and, consequently, different features. Which of these features are more interpretable to humans remains an open question. **Top:** the most important feature recovered by each of the six models we study, for the prediction of *traffic light* (first row) and *mud turtle* (second row). **Bottom:** Our framework, illustrated for DINOv3. Given a feature and its visual explanation, composed of a synthetic visualization, highly activating images, and corresponding heatmaps showing where the feature fires (top panels), participants either click where they think the feature fires on a novel image (left, *localizability*) or describe what the feature represents in free text (right, *semanticity*).

ation of the interpretability of a feature (see Fig. 1, bottom): a feature can fire in a predictable location yet defy verbal description, or be easy to name yet hard to point to. The proposed framework has three properties that support a fair cross-model comparison: (i) it replaces the polysemantic neuron basis used in prior work with more monosemantic directions recovered via sparse autoencoders (Huben et al., 2024; Gao et al., 2025b); (ii) it evaluates *functionally analogous* features across models by anchoring feature selection to a shared set of input images (Fig. 1, top); and (iii) it uses calibrated scores to a common chance level across features and models, so scores are directly comparable. Furthermore, the evaluation is scaled to $\sim$6,000 features across six models at a fraction of the cost of previous protocols (Zimmermann et al., 2023).

We apply the framework to six vision transformers—two supervised ViTs and four foundation models (DINOv2, DINOv3, CLIP, SigLIP)—and collect more than 15,000 behavioral responses. We find that the foundation models tested generally produce features that are *less* interpretable than those of their size-matched supervised baseline. The gap is not predicted by downstream task performance, but it is associated with two representational properties: feature *locality* and *coarse-grained* semantic alignment with human similarity judgments. DINOv3, whose training objective explicitly promotes

local features, provides an encouraging counterexample, suggesting that this interpretability gap is not an inevitable consequence of current training paradigms.

Our contributions are: **(1)** a framework for measuring and comparing the human interpretability of vision models, with two complementary psychophysics protocols—*localizability* and *semanticity*; **(2)** a human study across six vision transformers showing that the foundation models we tested generally produce features that are *less* interpretable than those of their size-matched supervised baseline, yet this gap is not a consequence of their greater capability: downstream performance does not explain it; and **(3)** the identification of two representational properties that predict interpretability—*locality* of feature activations and *coarse-grained* semantic alignment with humans.

2 Related Work

Sparse autoencoders and monosemanticity. Under the superposition hypothesis (Elhage et al., 2022), networks encode more features than they have neurons by superimposing concepts onto shared activation channels, producing polysemantic neurons. Sparse autoencoders (SAEs) recover a dictionary of more monosemantic directions and have emerged as the dominant remedy (Huben et al., 2024; Gao et al., 2025b; Templeton et al., 2024; Fel et al., 2025). We extract features via SAEs to avoid the polysemy that complicates neuron-level analysis.

Evaluation of model interpretability. Bau et al. (2017) introduced Network Dissection, to our knowledge the first framework to compare interpretability across models (architectures and training regimes). They score every convolutional unit from four CNNs against pixel-level masks from a set of human-labeled concepts (Broden), and models are ranked by the number of units that align with a concept above a given threshold. The approach is bounded by Broden's vocabulary, restricted to the neuron basis, and conflates concept alignment with interpretability, which, as Section 4.2 shows, do not necessarily coincide.

Borowski et al. (2021) pioneered psychophysics-based evaluation of DNN interpretability: participants viewed maximally and minimally activating images for a given neuron, then identified which of two novel query images also strongly activated it. Zimmermann et al. (2023) scaled this protocol across nine vision and vision-language architectures (80 features per model, 720 in total, a small fraction of the thousands a typical model encodes) and concluded that more capable models are not necessarily more interpretable. Zimmermann et al. (2024) later automated the evaluation by replacing human judgments with pairwise perceptual similarity. Their metric correlates with human results, but cross-model differences arise primarily from a tail of highly uninterpretable units linked to superposition. Applying a similar protocol to CNNs, Colin et al. (2026) showed that reliance on a neuron basis can mask cross-model differences because the impact of superposition is itself model-specific. This line of work presents two methodological limitations that prevent fair cross-model comparison: reliance on the polysemantic neuron basis and a forced-choice task whose process-of-elimination shortcuts produce model-specific chance levels. Our framework resolves both and extends model coverage to several foundation models that postdate prior studies.

3 Method

3.1 Current protocols are ill-suited for model interpretability.

Prior protocols (Borowski et al., 2021; Zimmermann et al., 2023) frame the task as a contrastive forced-choice: participants view maximally and minimally activating images for a feature, then select which of two novel query images also strongly activates it (see Fig. A1 in Appendix C for an example trial). While intuitive, this design has a fundamental confound: participants can identify the correct answer by understanding what the feature is not about—*e.g.*, not a corn detector (Fig. A1)—rather than genuinely interpreting the feature itself.

To test whether this confounding effect is consequential, we performed a control experiment on two models from Zimmermann et al. (2023) (ResNet-50 and ViT-B32). We ran two versions of a given trial: the original version (Fig. A1) and a variation where we replaced the maximally activating images with images ranking 25,000 out of 50,000 in each feature's activation distribution, thus carrying little to no information about the feature's visual content (Fig. A2). In this case, participants were

expected to perform at chance (50%). Participants nevertheless performed significantly above the 50% chance level for both models. For ResNet-50, accuracy was 60.0% across $2{,}480$ trials from 62 participants (binomial test, 95% CI $[58.1\%, 62.0\%]$, $p < .001$). For ViT-B/32, accuracy was 53.6% across $1{,}840$ trials from 46 participants (95% CI $[51.3\%, 55.9\%]$, $p = .001$). The protocol therefore produces significantly inflated and *model-dependent baselines* even when the maximally activating images carry little information about the feature.

This model-specific baseline inflation makes raw scores incomparable across architectures. Zimmermann et al. (2023) reported ResNet-50 as more interpretable than ViT-B32 (83% vs. 80%). Yet, when each score is normalized relative to its model-specific baseline, the ranking *reverses*: ViT-B32 is 27% above its baseline while ResNet-50 is only 23% above its own. Ignoring this baseline effect can therefore lead to misleading model comparisons.

3.2 Framing interpretability as identification.

Overview. We address the limitations described above with two complementary protocols that operationalize interpretability: *localizability*, *i.e.*, can a participant predict *where* a feature fires on a novel image?; and *semanticity*, *i.e.*, can a participant accurately describe *what* a feature represents? Both protocols share the same feature extraction pipeline (Fig. A3 in Appendix E) and visual explanations (Fig. 1), isolating differences in measured interpretability to differences in the models themselves. They differ only in the task and scoring function. Because neither task provides a distractor to exploit, performance requires genuine understanding of the feature's visual content. Localizability scores are mapped to a common, chance-anchored scale (Section 3.2), enabling direct comparison across features, models and protocols. Semanticity requires no chance-level correction: although images and descriptions vary across models, both are scored in a shared embedding space, enabling fair model comparisons.

Unlike Zimmermann et al. (2023), who studied multiple layers, we evaluate only the final latent representation: the penultimate layer for supervised models, the last layer for DINO models, and the vision-encoder output for VLMs.

Feature extraction. For each model, we extract features from its vision encoder. Because individual neurons are often polysemantic (Elhage et al., 2022), we use TopK sparse autoencoders (Gao et al., 2025b) (SAEs; $K = 38$) to recover feature directions from patch-token activations that are more likely to be monosemantic. Training details, reconstruction quality, dead-feature rates, and stability are reported in Appendix F.

Feature relevance and linear probing. For each model, we train a linear classification head on ImageNet (Deng et al., 2009) on top of the frozen backbone, from which we derive (i) per-feature importance scores that quantify each feature's contribution to the classifier's predictions, computed as Gradient$\times$Input (Shrikumar et al., 2017), shown to be particularly faithful in latent space (Fel et al., 2023b); and (ii) a downstream accuracy baseline used in Section 4.1.

Feature selection. A pilot following Zimmermann et al. (2023) showed that, under a fixed annotation budget, evaluating more features yields more stable model-level estimates than repeating judgments on fewer features (Appendix D). We therefore prioritize breadth in the main experiment.

To obtain a representative subset of features, we select features that are both diverse and relevant to the model's decisions. To this end, we randomly sample one image for each of the 559 superordinate classes underlying the 1,000 ImageNet labels, yielding a fixed set of 559 images. For each image, we select the three visual features that rank as the most important for the model's prediction on that image. During selection, we exclude features that activate on more than 10% of patch tokens, following Sun et al. (2025) (in practice, this removes about 7–90 candidate features per model). This procedure yields a total of $559 \times 3 = 1{,}677$ trials per model, but because the same feature can be selected for multiple images, it results in 500–1,300 unique features evaluated per model (see Appendix G).

Feature explanation. Both protocols present participants with the same visual explanation of a feature (Fig. 1, top): (a) a synthetic feature visualization synthesized via MACO (Fel et al., 2023a) that maximally activates the feature (left panel in Fig. 1); (b) nine highly activating natural images

(middle panel in Fig. 1); and (c) their corresponding RISE heatmaps (Petsiuk et al., 2018) indicating where the feature fires on each image (right panel in Fig. 1). Participants use these three sources of information to form an understanding of the feature before solving the task.

Localizability (*where*). Each *localizability* trial also presents (d) a novel query image (bottom, left in Fig. 1) and participants are asked to click on the region of the image where they believe the feature appears. Each click is scored against the query image's RISE heatmap, smoothed with a Gaussian filter. Because attribution maps are only proxy ground truth, we test whether our results depend on this choice by rescoring the original human clicks using HSIC (Novello et al., 2022), an independent black-box attribution method. The broad model ordering is largely preserved, although the middle of the ranking shifts (see Appendix I.1 for full results).

Let v denote the heatmap value at the clicked location and $p = \widehat{F}(v)$ its empirical cumulative distribution function (ECDF) over heatmap pixels. Because heatmap distributions vary across features (*e.g.*, sparse vs. diffuse activations), raw percentiles are not directly comparable across features or models. We therefore introduce a *chance-anchored normalization* that sets chance to $s = 0.5$. Letting $p_\mu = \widehat{F}(\mu)$ denote the percentile rank of the mean activation μ, we map p_μ to 0.5 and the extremes to 0 and 1:

$$s(v) = \begin{cases} 0.5 - 0.5 \cdot \dfrac{p_\mu - p}{p_\mu}, & \text{if } p < p_\mu, \\ 0.5 + 0.5 \cdot \dfrac{p - p_\mu}{1 - p_\mu}, & \text{otherwise.} \end{cases}$$

By anchoring chance ($s = 0.5$) to the mean activation, the score has the same interpretation across features and models: clicks above the mean score above chance, and those below the mean score below chance, regardless of whether a feature activates focally or diffusely. We refer to this score as the *localizability* score.

Semanticity (*what*). Each *semanticity* trial asks participants to (d) write a short free-text description of the feature (bottom right in Fig. 1) and rate their confidence on a 5-point Likert scale. We evaluate two complementary properties: (1) Description *plausibility*, which measures whether the proposed description is supported by the feature's visual evidence; and (2) *agreement*, which measures whether independent participants produce similar descriptions of the feature. *Confidence* helps distinguish ambiguity, reflected by low agreement and low confidence, from confident misinterpretation, reflected by low plausibility despite high confidence.

To measure description *plausibility*, we extract a 96×96 crop around the peak RISE activation from each of the nine most-activating images and encode the crops and the participant's description using CLIP ViT-B/16 (Radford et al., 2021). The plausibility score is the mean cosine similarity between the text embedding and the nine image embeddings. We validate that the results are stable across six alternative vision-language encoders ($\rho = .70$–$.91$, all $p < .001$) and when using full images rather than crops (Appendix K). For reference, the mean similarity between image and text embeddings from different features is 0.19.

To measure the descriptions' *agreement*, we embed all descriptions of features with at least two responses using a popular sentence-encoder, MPNet (Song et al., 2020), and compute their mean pairwise cosine similarity. High agreement indicates a clear, shared interpretation, whereas low agreement suggests greater ambiguity.

Because agreement requires multiple descriptions per feature, we evaluate it on a representative subset sampled across the localizability distribution; see Appendix J for sampling details.

4 Results

We evaluated six vision transformers: two supervised baselines (`ViT-S/16` and `ViT-B/16`) and four foundation models—`DINOv2 ViT-B/14`, `DINOv3 ViT-B/16`, `CLIP ViT-B/16`, and `SigLIP ViT-B/16`. Across both experiments, we recruited about 440 participants via Prolific.

We analyze 13,400 behavioral responses from the 377 participants who passed our pre-specified quality checks. Recruitment, compensation, and exclusion criteria are detailed in Appendix H.

Table 1: Localizability, Semanticity (plausibility and agreement) and confidence scores across six vision transformers. The size-matched supervised model (ViT-B) consistently outperforms foundation models on both dimensions. Best result per row is in bold, second best is underlined.

	Supervised		**Foundation**			
	ViT-S	ViT-B	DINOv2	DINOv3	CLIP	SigLIP
Localizability (↑)	<u>80.3</u>	**86**	71	79.6	79.7	71.4
Semanticity (plausibility) (↑)	**0.274**	<u>0.273</u>	0.259	0.260	0.266	0.253
Semanticity (agreement) (↑)	<u>0.493</u>	**0.519**	0.433	0.458	0.440	0.443
Semanticity confidence (↑)	3.43	<u>3.61</u>	**3.68**	3.41	3.51	3.38

Foundation models are less interpretable than supervised ones. Table 1 reports the median *localizability* (L) and *semanticity* (*plausibility* (P) and *agreement* (A)) scores for each model[1]. Scores differ significantly across models (L: Kruskal–Wallis, $H(5) = 118.40$, $p < .001$; P: $H(5) = 295.36$, $p < .001$; A: $H(5) = 30.3$, $p < .001$). Foundation models score consistently lower than their size-matched supervised counterpart, ViT-B (Dunn's test, all $p < .001$), regardless of whether they used language supervision (CLIP, SigLIP) or self-supervised objectives (DINOv2, DINOv3). Although *localizability* and *semanticity* are designed to capture two different dimensions of interpretability, they yield highly similar model rankings (L-P: $r = 0.84$, $p = 0.036$; L-A: $r = 0.82$, $p = 0.046$; P-A: $r = 0.80$, $p = 0.059$), suggesting that they capture a shared property of interpretability. Qualitative examples spanning the four quadrants of *localizability* and *plausibility* are shown in Appendix Fig. A8.

Models can feel more interpretable than they are. Average confidence is relatively homogeneous across models (3.38–3.68, Table 1, bottom row), but its association with interpretability is not (P, $\rho = 0.34$; A, $\rho = 0.30$). DINOv2 elicits the highest confidence despite ranking near the bottom on interpretability: last on *localizability* and *agreement*, and second-to-last on *plausibility*. SigLIP, by contrast, performs poorly across all measures. Thus, confidence can diverge substantially from measured interpretability: a model can feel more interpretable than it is. Appendix Fig. A7 shows examples of confidently but inaccurately described features (and vice versa), a caution worth keeping in mind given that DINOv2 features have been used as a reference for interpretable representations (Dreyer et al., 2024).

4.1 Alignment with performance

A natural question is whether interpretability is related to task performance: models with richer, more semantically structured representations might be both more capable on downstream tasks and more legible to humans. We therefore compute Spearman correlations between each model's interpretability scores and its performance on three vision tasks: ImageNet-1k classification, semantic segmentation, and perceptual grouping. We report the results for the first two tasks in Fig. 2.

ImageNet-1k classification. Foundation models achieve higher ImageNet top-1 accuracy but lower interpretability. Across the six models, neither *localizability* ($\rho = -0.48$, $p = 0.333$) nor *semanticity* (P: $\rho = -0.60$, $p = 0.209$; A: $\rho = -0.48$, $p = 0.334$) are significantly correlated with classification accuracy.

Semantic segmentation. Segmentation performance (mIoU) on ADE20K (Zhou et al., 2017), measured using a linear probe, is not significantly correlated with either *localizability* ($\rho = -0.54$, $p = 0.273$) or with *semanticity* (P: $\rho = -0.77$, $p = 0.072$; A: $\rho = -0.69$, $p = 0.133$). Perceptual grouping yields the same pattern, with small correlations ($|\rho| \leq 0.12$, all $p > 0.8$). See Appendix M for details.

Thus, across all three benchmarks, we find no evidence that task performance tracks interpretability.

[1]We report the median per model because the scores are not normally distributed; see Appendix L.

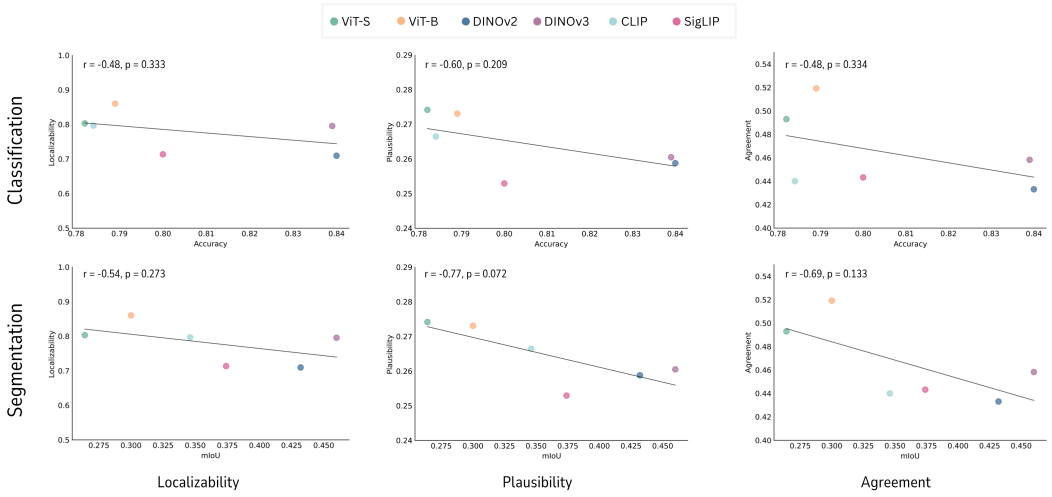


Figure 2: **Interpretability is not significantly associated with downstream task performance.** Each panel plots an interpretability score against a capability benchmark across the six models. Top row: Classification (ImageNet-1k top-1 accuracy (Deng et al., 2009)); bottom row: Segmentation (ADE20K semantic segmentation (Zhou et al., 2017)). Columns: *Localizability*, *Plausibility*, and *Agreement*. Spearman ρ and p-values are shown in each panel. None of the correlations are statistically significant, suggesting that interpretability and downstream task performance capture largely orthogonal properties of a visual representation.

Locality of the representation. Beyond task performance, we ask whether *locality*, *i.e.*, the degree to which a feature's activations concentrate on confined image regions, predicts interpretability: a focal activation supplies a concrete spatial anchor (an object part, a texture), while a diffuse one leaves the participant to guess what the activating regions share. We quantify locality with the Hoyer metric applied to each feature's RISE heatmaps (Hoyer, 2004):

$$H(\mathbf{x}) = \frac{(\sqrt{n} - \|\mathbf{x}\|_1/\|\mathbf{x}\|_2)}{(\sqrt{n} - 1)}$$

with n the number of pixels, and where $H = 0$ corresponds to uniform activation across the image and $H = 1$ to all mass on a single pixel. We average H across each feature's 9 most activating images and then across the model's features. In contrast to task performance, locality emerges as a strong and consistent predictor of interpretability, correlating with *localizability* ($\rho = 0.91$, $p = 0.011$; Fig. 3, top left) and, importantly, with both *semanticity* metrics (P: $\rho = 0.99$, $p < .001$; Fig. 3, top middle; A: $\rho = 0.82$, $p = 0.048$; Fig. 3, top right). This consistency across distinct protocols and scoring functions suggests that locality captures a broader property associated with human interpretability. See Appendix N for additional results.

4.2 Alignment with human perception

Next, we consider human alignment, *i.e.*, the degree to which a model aligns with some properties of human perception, as alignment has been shown to improve representation quality over a wide range of downstream tasks (Sucholutsky & Griffiths, 2023; Muttenthaler et al., 2023; Sundaram et al., 2024; Muttenthaler et al., 2025). Because not all measures of alignment capture the same properties of human perception (Ahlert et al., 2024), we study 3 kinds of alignment: alignment with human (a) visual strategies; (b) coarse-grained similarity judgments (semantic similarity); and (c) fine-grained similarity judgments (lower-level visual similarity).

Alignment with human visual strategies. Strategy alignment—the overlap between model-driven and human-attended image regions (Linsley et al., 2019; Fel et al., 2022)—does not correlate significantly with *localizability* ($\rho = 0.14$, $p = 0.798$) or *semanticity* (P: $\rho = 0.21$, $p = 0.696$; A: $\rho = 0.30$, $p = 0.564$).

Alignment with human similarity judgments. Human similarity judgments operate at multiple levels of abstraction (Muttenthaler et al., 2025). *Fine-grained* similarity captures within-category distinctions (*e.g.*, how similar two butterfly species are to each other), while *coarse-grained* similarity captures broad semantic structure across categories (*e.g.*, how similar a buffalo is to a spider relative to grass). We measure alignment at the different levels of similarity using three datasets: THINGS similarity (Hebart et al., 2020; 2023), Levels (Muttenthaler et al., 2025), and NIGHTS (Fu et al., 2023).

Alignment with *fine-grained* similarity does not correlate with either *localizability* (NIGHTS: $\rho = -0.41$, $p = 0.418$; Levels: $\rho = 0.20$, $p = 0.705$), *plausibility* (NIGHTS: $\rho = -0.47$, $p = 0.347$; Levels: $\rho = 0.26$, $p = 0.617$) or *agreement* (NIGHTS: $\rho = -0.33$, $p = 0.523$; Levels: $\rho = -0.06$, $p = 0.915$).

In contrast, alignment with *coarse-grained* similarity is generally lower and shows a strong correlation with interpretability (Fig. 3, bottom), which is significant for the THINGS similarity dataset for both *localizability* and *plausibility*: *localizability* (THINGS: $\rho = 0.84$, $p = 0.036$; Levels: $\rho = 0.71$, $p = 0.110$), *plausibility* (THINGS: $\rho = 0.85$, $p = 0.030$; Levels: $\rho = 0.6$, $p = 0.205$), and *agreement* (THINGS: $\rho = 0.85$, $p = 0.03$; Levels: $\rho = 0.6$, $p = 0.21$). See Appendix O for additional results.

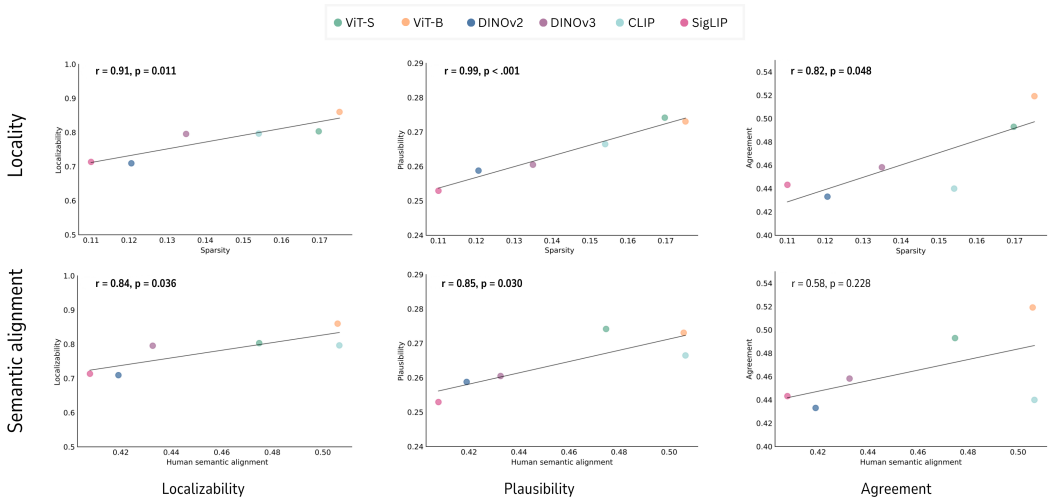


Figure 3: **Locality and coarse-grained semantic alignment track interpretability.** Each panel plots an interpretability score against a representational property across the six models. Top row: Locality (mean Hoyer sparsity over feature heatmaps); bottom row: Human semantic alignment (coarse-grained alignment with human similarity judgments on THINGS (Hebart et al., 2020)). Columns: *Localizability*, *Plausibility*, and *Agreement*. Spearman ρ and p-values per panel. Locality is the strongest predictor of interpretability, consistent with focal activations providing a clearer visual anchor for understanding a feature. Coarse-grained alignment also tracks interpretability, particularly for *localizability* and *plausibility*, suggesting that interpretable representations are not only local but also organized in ways that better reflect human categorical similarity judgments.

This dissociation suggests that interpretability depends less on fine-grained perceptual fidelity than on whether a model organizes features around the high-level semantic categories that structure human perception.

5 Discussion

Two protocols, three measures, one interpretability. The two protocols differ in nearly every respect: one asks participants to point, the other to write; one is scored against a spatial heatmap, the other against a language or vision–language embedding. That they yield concordant rankings and surface the same predictors suggests that we are measuring a stable property of the representations rather than an artifact of any one task, and either protocol may suffice on its own in future work.

Foundation models are less interpretable than their size-matched supervised counterpart. The gap appears in both vision-only SSL models (DINOv2, DINOv3) and vision-language models (CLIP, SigLIP), so neither pretraining signal is the cause. The common factor is foundation-style pretraining itself. One reading is that these models reach—and on many benchmarks exceed—human performance in part by developing *superhuman* features: representations effective for the task but not the kind a human observer can readily make sense of. The very generality that makes foundation models capable may also push their representations away from a human-interpretable basis. This gap is easy to miss without direct evaluation—DINOv2, often treated as a reference for interpretable representations (Dreyer et al., 2024), elicits the highest rater confidence yet ranks among the least interpretable models on *localizability* and *semanticity*.

Interpretability is orthogonal to task performance. Across every benchmark we examined, capability and interpretability are uncorrelated. The relationship is neither a tradeoff nor a free lunch: how well a model classifies, segments, or groups says little about whether a human can interpret its features. The leaderboards by which the field currently judges progress are silent on interpretability, a property independently desirable in high-stakes settings (*e.g.*, clinical decision support or autonomous driving) where these same backbones are increasingly considered. Interpretability has to be designed for explicitly; it will not arrive as a byproduct of capability, and prior evidence suggests that scale alone does not guarantee it either (Zimmermann et al., 2023).

Locality is the proximate signature of an interpretable representation. The locality of feature activations is the strongest predictor of interpretability we identify. Foundation models tend to develop diffuse features that fire across larger, less focused regions, blending local content with broader scene context. This may help on downstream tasks but leaves a human observer without a clear visual anchor. DINOv3 is the encouraging exception: one of the most capable models in our set and also among the most interpretable foundation models, a pattern consistent with its training objective explicitly promoting local features. This finding suggests that locality-aware objectives can partially close the interpretability gap, though whether they do so without capability cost requires additional direct experimental verification.

Coarse-grained semantic alignment, not perceptual fidelity, predicts interpretability. Of the alignment measures we considered, only coarse-grained alignment with human similarity judgments tracks interpretability. A model can attend to the same image regions humans do, yet still encode what it sees there in features humans cannot make sense of. Fine-grained perceptual alignment has been found to be uniformly high and largely shared across models (Gröger et al., 2026). Coarse-grained alignment, on the other hand, is consistently lower, leaving real room for improvement. Interpretability appears driven less by perceptual fidelity than by how a representation organizes the visual world at the categorical level. Training signals built from human odd-one-out judgments (Hebart et al., 2020) seem like a promising direction to close part of the gap.

6 Conclusion

We introduced a framework for measuring and comparing the human interpretability of vision models, built around two complementary psychophysics protocols: *localizability* (where a feature fires) and *semanticity* (what it represents). Across six vision transformers and more than $15{,}000$ behavioral responses, foundation models, whether trained with language supervision or self-supervised objectives, were consistently *less* interpretable than their size-matched supervised counterpart—a surprising result for models that have surpassed those same baselines on most other axes. The gap is not a capability tradeoff: interpretability is uncorrelated with downstream task performance on every benchmark we examined, and both protocols agree on this and on its predictors. Two properties track interpretability: the *locality* of a feature's activations and *coarse-grained* semantic alignment with human similarity judgments. Together, they point to a concrete recipe for building more interpretable vision models—training objectives that promote local feature activations and align representations with the coarse semantic structure of human perception—and DINOv3, whose objective explicitly promotes locality, suggests that capability and interpretability need not trade off.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript for readability, including correcting typos and grammar. All authors reviewed the final text, and we take responsibility for the full content of this work, including any text produced with the aid of generative AI.

REFERENCES

Jannis Ahlert, Thomas Klein, Felix A. Wichmann, and Robert Geirhos. How aligned are different alignment metrics? In *ICLR 2024 Workshop on Representational Alignment (Re-Align)*, 2024. URL https://openreview.net/forum?id=cHlKB28bjV.

Mohammed Baharoon, Waseem Qureshi, Jiahong Ouyang, Yanwu Xu, Abdulrhman Aljouie, and Wei Peng. Evaluating general purpose vision foundation models for medical image analysis: An experimental study of DINOv2 on radiology benchmarks. *arXiv preprint arXiv:2312.02366*, 2023.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6541–6549, 2017.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *OpenAI Blog*, 2023. URL https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Judy Borowski, Roland S. Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain CNN activations better than state-of-the-art feature visualization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Christopher Chiu, Maximilian Heil, Teresa Kim, and Anthony Miyaguchi. Fine-grained classification for poisonous fungi identification with transfer learning. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*, pp. 2016–2026, 2024.

Julien Colin, Lore Goetschalckx, Thomas Fel, Victor Boutin, Thomas Serre, and Nuria Oliver. Choosing the right basis for interpretability: Psychophysical comparison between neuron-based and dictionary-based representations. *Transactions on Machine Learning Research*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Maximilian Dreyer, Erblina Purelku, Johanna Vielhaben, Wojciech Samek, and Sebastian Lapuschkin. PURE: Turning polysemantic neurons into pure features by identifying relevant circuits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8212–8217, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Niels G Faber, Seyed Sahand Mohammadi Ziabari, and Fatemeh Karimi Nejadasl. Leveraging foundation models via knowledge distillation in multi-object tracking: Distilling DINOv2 features to FairMOT. *arXiv preprint arXiv:2407.18288*, 2024.

Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 9432–9446, 2022.

Thomas Fel, Thibaut Boissin, Victor Boutin, Agustin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom Rousseau, Rémi Cadène, Lore Goetschalckx, Laurent Gardes, and Thomas Serre. Unlocking feature visualization for deep network with MAgnitude constrained optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 37813–37826, 2023a.

Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 54805–54818, 2023b.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E. Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16543–16572. PMLR, 2025.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 50742–50768, 2023.

Haoxiang Gao, Zhongruo Wang, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A survey for foundation models in autonomous driving. In *2025 6th International Conference on Computer Vision and Data Mining (ICCVDM)*, pp. 63–71. IEEE, 2025a.

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025b.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, 2020.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Adam H. Rockter, Charles Y. Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I. Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023. doi: 10.7554/eLife.82580.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research (JMLR)*, 5:1457–1469, 2004.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, et al. A survey on vision-language-action models for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 4583–4595, 2025. doi: 10.1109/ICCVW69036.2025.00476.

Licheng Jiao, Jiayao Hao, Ruiyang Li, Lingling Li, Xu Liu, Fang Liu, Wenping Ma, Puhua Chen, Zhongjian Huang, Jingyi Yang, Jiaxuan Zhao, and Qigong Sun. Foundation models meet medical image interpretation. *Research*, 9:1024, 2026. doi: 10.34133/research.1024. URL https://spj.science.org/doi/abs/10.34133/research.1024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, pp. 2679–2713, 2025.

Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul MB Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.

Yihao Li, Saeed Salehi, Lyle Ungar, and Konrad Kording. Does object binding naturally emerge in large pretrained vision transformers? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 3394–3423, 2025.

Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Théo Moutakanni, Piotr Bojanowski, Guillaume Chassagnon, Céline Hudelot, Armand Joulin, Yann LeCun, Matthew Muckley, Maxime Oquab, Marie-Pierre Revel, and Maria Vakalopoulou. Advancing human-centric AI for robust X-ray analysis through holistic self-supervised learning. *Nature Communications*, 17:9931, 2026. doi: 10.1038/s41467-026-76076-4.

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 50978–51007, 2023.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, 2025. doi: 10.1038/s41586-025-09631-6.

Paul Novello, Thomas Fel, and David Vigouroux. Making sense of dependence: Efficient black-box explanations using dependence measure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tuomas Oikarinen and Tsui-Wei Weng. CLIP-Dissect: Automatic description of neuron representations in deep vision networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *Transactions on Machine Learning Research*, 2026.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 16857–16867, 2020.

Ilia Sucholutsky and Tom Griffiths. Alignment with human representations supports robust few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 73464–73479, 2023.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025.

Xiaoqing Sun, Alessandro Stolfo, Joshua Engels, Ben Peng Wu, Senthooran Rajamanoharan, Mrinmaya Sachan, and Max Tegmark. Dense SAE latents are features, not bugs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 140252–140282, 2025. doi: 10.52202/085713-4686.

Shobhita Sundaram, Stephanie Fu, Lukas Muttenthaler, Netanel Tamir, Lucy Chai, Simon Kornblith, Trevor Darrell, and Phillip Isola. When does perceptual alignment benefit vision representations? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 55314–55341, 2024.

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, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, 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. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. DINO-Tracker: Taming DINO for self-supervised point tracking in a single video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 367–385, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.

Roland S. Zimmermann, Thomas Klein, and Wieland Brendel. Scale alone does not improve mechanistic interpretability in vision models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Roland S. Zimmermann, David Klindt, and Wieland Brendel. Measuring per-unit interpretability at scale without humans. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 48448–48483. Curran Associates, Inc., 2024. doi: 10.52202/079017-1535. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/56ed2bd15b66f709cd81cb1aaa0496b9-Paper-Conference.pdf.