Capability $\neq$ Interpretability: Human Interpretability of Vision Foundation Models

Julien Colin
ELLIS Alicante, Spain
Brown University, USA
julien_colin@brown.edu

Lore Goetschalckx
IMEC, Leuven, Belgium

Nuria Oliver*
ELLIS Alicante, Spain

Thomas Serre*
Brown University, USA

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 close 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) *nameability*—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 collect more than $15{,}000$ behavioral responses, analyzing the $13{,}400$ responses from the 377 participants who passed our pre-specified quality checks. Foundation models are consistently *less* interpretable than their 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 every foundation model we tested falls below the supervised baselines that came before. Capability alone cannot close that gap; locality and coarse-grained alignment can.

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 [1], models like CLIP [2], SigLIP [3], DINOv2 [4], and DINOv3 [5] already serve as general-purpose visual backbones across an ever-expanding range of downstream tasks, including classification [6], segmentation [7], object tracking [8, 9], and robotic perception [10]. The same models are increasingly deployed in high-stakes settings—clinical decision support [11–14], autonomous driving [15, 16]—where a user often needs

*co-principal investigators

Preprint.

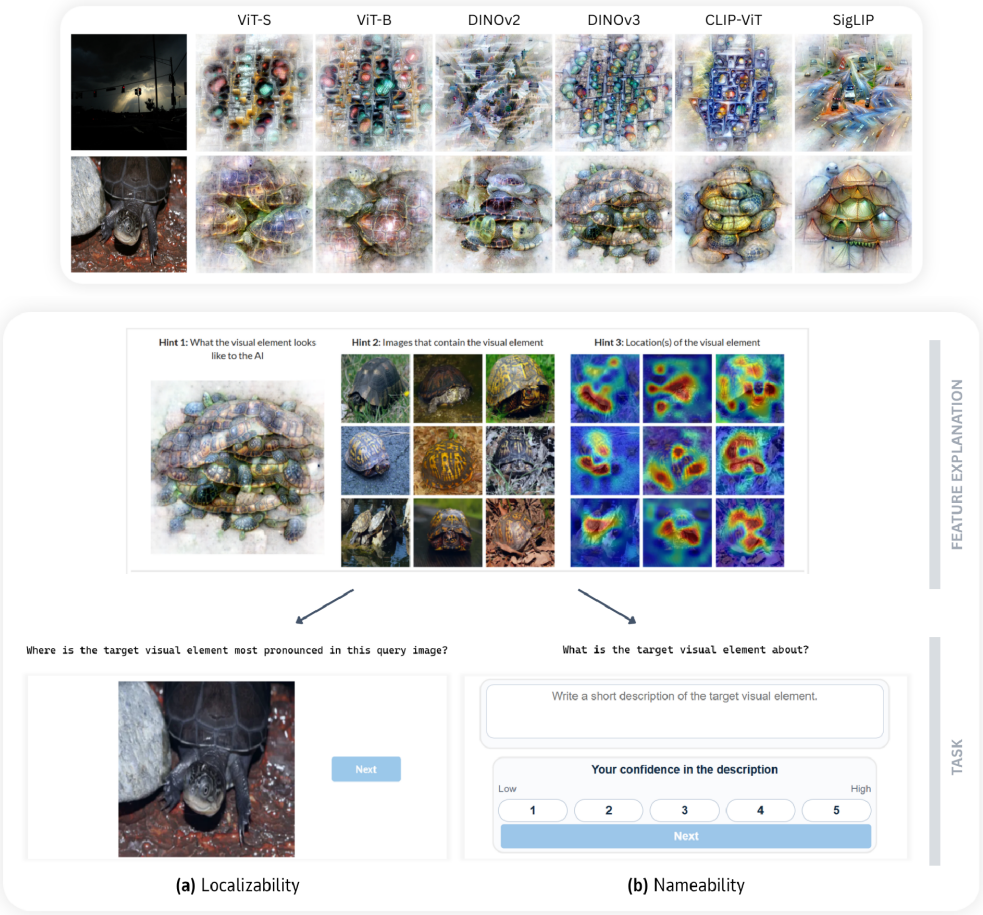


Figure 1: Models trained under different objectives learn different internal representations and, consequently, different features. Which of these features are more interpretable to humans remains an open question. **Top:** the most important feature recovered by each of the six models we study, for 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—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, *nameability*).

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 on this dimension, because existing evaluation protocols cannot place different models 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 understanding: feature *localizability* — can an observer predict *where* a feature fires on a novel image? — and feature *nameability* — can an observer accurately describe *what* a feature represents? Together they form a comprehensive evaluation (see Fig. 1): 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 enable a fair cross-model comparison: (i) It replaces the polysemantic neuron basis used in prior work with monosemantic directions recovered via sparse autoencoders [17, 18], so each tested unit corresponds to a single concept; (ii) it evaluates *functionally analogous* features across models by anchoring feature selection to a shared set of input images (Fig. 1), rather than sampling features randomly from

each model; and (iii) it uses calibrated scores to a common chance level across features and models, so raw 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 [19].

We apply the framework to six vision transformers—two supervised ViTs and four foundation models (DINOv2 [4], DINOv3 [5], CLIP [2], SigLIP [3])—and collect more than $15{,}000$ behavioral responses, of which the $13{,}400$ from the 377 participants who passed our pre-specified quality checks enter the analyzes. The headline finding is sobering: every foundation model we tested produces features that are *less* interpretable than those of the supervised baselines that came before, and downstream task performance does not predict the gap. Two representational properties do: feature *locality* and *coarse-grained* semantic alignment with human similarity judgments. DINOv3—whose training objective explicitly promotes local features—is the encouraging counterexample, suggesting that the gap is not architectural destiny.

Our contributions are: **(1)** a framework for measuring and comparing the human interpretability of vision models, with two complementary psychophysics protocols—*localizability* and *nameability*—whose convergence on the same rankings and predictors validates the approach; **(2)** a human study across six vision transformers showing that every foundation model we tested produces features that are, surprisingly, *less* interpretable than those of the supervised baselines that came before, with no correlation to downstream task performance on any benchmark; and **(3)** the *locality* of feature activations and *coarse-grained* semantic alignment with humans as the two predictors of interpretability we identify, both of which capability alone does not deliver.

2 Related Work

Sparse autoencoders and monosemanticity. The superposition hypothesis [20] accounts for polysemantic neurons: networks encode more features than they have neurons by superimposing concepts onto each activation channel. Sparse autoencoders (SAEs) recover a sparse dictionary of monosemantic directions and have emerged as the dominant remedy [17, 18, 21, 22]. We extract features via SAEs to avoid the polysemy that complicates neuron-level analysis.

Evaluation of model interpretability. Bau et al. [23] 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. [24] 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. [19] 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. [25] 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. [26] showed that reliance on a neuron basis can mask cross-model differences because the impact of superposition is itself model-specific. Two methodological limitations, therefore, prevent fair cross-model comparison in this lineage: exclusive 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.

We build on this lineage with broader model coverage—several of the foundation models in our evaluation postdate prior studies—and a model-agnostic framework that resolves two limitations: exclusive reliance on the neuron basis, and a confound in the forced-choice task that produces model-specific chance levels and prevents fair cross-model comparison.

3 Method

3.1 Current protocols are ill-suited for models' interpretability.

Prior protocols [19, 24] 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. A.1 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. A.1)—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. [19] (ResNet-50 and ViT-B32). We ran two versions of a given trial: the original version (Fig. A.1) 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. A.2). In this case, participants were expected to perform at chance (50%). Instead, participants reliably exceeded chance with *model-specific baselines*: 53% for ViT-B32 and 60% for ResNet-50.

This model-specific baseline inflation makes raw accuracy scores incomparable across architectures. Zimmermann et al. [19] reported ResNet-50 as more interpretable than ViT-B32 (83% vs. 80%). However, once each score is expressed 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. Any protocol that does not account for this artifact risks drawing erroneous conclusions.

3.2 Framing interpretability as identification.

Overview. We address the limitations described above with a framework comprising of two complementary protocols that operationalize interpretability: *localizability*, *i.e.*, can a participant predict *where* a feature fires on a novel image?; and *nameability*, *i.e.*, can a participant accurately describe *what* a feature represents? Both protocols share the same feature extraction pipeline and the same visual explanations presented to participants (Fig. 1). They differ only in the task and scoring function. Because there is no distractor to exploit, participants must rely on 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. Nameability requires no chance-level correction: images and descriptions vary across models, but they are scored in a shared embedding space, making cross-model comparisons fair.

By evaluating the interpretability of features, we aim to derive a more general score representative of model interpretability. Whereas Zimmermann et al. [19] studied features across multiple layers, our focus on foundation models leads us, moving forward, to examine only the last latent representation of each model, *i.e.*, the output of the penultimate layer for supervised models, the output of the last layer for DINO models, and the output of the vision encoder for VLMs.

Feature extraction. For each model, we extract features from its vision encoder. Because individual neurons in vision models are often polysemantic [20], we use sparse autoencoders (SAEs) to recover a dictionary of monosemantic directions. SAEs are trained on activations from the ImageNet training split ($\approx$1.28M images), using all patch tokens per image (256 for patch size 14 in DINOv2, 196 for patch size 16 in all other models). Each SAE uses an expansion factor of $\times 10$, yielding $3{,}840$ features for ViT-S and $7{,}680$ features from all the other models.

Feature relevance and linear probing. For each model, we train a linear classification head on ImageNet 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 input [27], shown to be particularly faithful in latent space [28]; and (ii) a downstream accuracy baseline used in Section 4.1.

Feature selection. Given that current computer vision models are expected to encoded thousands of features, we aim to scale the number of features evaluated. Building on Zimmermann et al. [19], we first ran a pilot with 80 features and 10 trials per feature to obtain reliable per-feature interpretability

estimates and a model interpretability score. We then evaluated the depth-vs-breadth tradeoff: we found that the number of features evaluated is more important than the number of repetitions per feature to obtain stable model-level estimates (see Appendix D). We therefore adopted this design in the main experiment.

Following common practice[29, 30], we exclude dense features—*i.e.*, SAE features that activate on most inputs and do not encode specific concepts—leaving between 500 and $1{,}300$ unique features per model, which reflects architectural differences in feature reuse (see Appendix F).

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 [31] 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 [32] 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 additionally 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 RISE heatmap of the query image (smoothed with a Gaussian filter) as follows.

Let v be the heatmap value at the clicked location and $p = \widehat{F}(v)$ its empirical CDF percentile over the 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 pins chance to $s = 0.5$. Writing $p_\mu = \widehat{F}(\mu)$ for the percentile rank of the mean activation μ, we map p_μ to 0.5 and the extremes to 0 and 1:

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

A participant who clicks at random scores $s = 0.5$ in expectation, regardless of whether the feature activates focally or diffusely. We refer to this score as the *localizability* score.

Nameability (*what*). Each *nameability* trial asks participants to write a short free-text description of the feature (bottom, right in Fig. 1), based on the explanation panel shown at the top, and a confidence score about their description on a 5-point Likert scale. We aim to collect multiple descriptions per feature, which scales up the number of trials per feature. We compensate this increase by evaluating a representative subsample rather than the full feature set: for each model, we bin features into deciles by their *localizability* score and sample uniformly within each bin ($\sim$30 features per 10% bin, $\sim$300 features per model).

We score the description by comparing it against the feature's visual content in CLIP's joint vision–language space [2]. For each of the nine most-activating images, we take a 96×96 crop centered on the peak RISE activation ($\sim 1/5$ of the image) and encode it with the CLIP image encoder. We encode the participant's description with the CLIP text encoder. The *nameability* score is the mean cosine similarity between the text embedding and the nine crop embeddings.

Note that we compare against *crops* rather than whole images so the score rewards descriptions that go beyond naming the object, something that does not necessarily emerge when using the full image (Fig. A.3). Regardless, the results are robust to using the full image (see Appendix E).

Due to the CLIP modality gap, *nameability* scores are bounded well below 1. In practice, they range between 0.13 and 0.35 with a chance-level at 0.19—average similarity for an image-text pair sampled from two different features. See Appendix E for further details.

Participants. Across both experiments, we recruited about 440 participants via Prolific[2]. All were native English speakers without reported visual impairments and completed the study on a laptop

[2] `www.prolific.com`

Table 1: Localizability, Nameability and Confidence scores across six vision transformers. Supervised models consistently outperform 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	80	79.7	71.4
Nameability (↑)	**0.274**	<u>0.273</u>	0.259	0.260	0.266	0.253
Confidence (↑)	3.43	<u>3.61</u>	**3.68**	3.41	3.51	3.38

or desktop. Each provided informed consent and received \$2.95 ($\sim$\$16/hr; 10–13 minutes). The protocol was approved by the IRB of an author-affiliated institution. We analyze about $13,400$ behavioral responses from the 377 participants who (i) passed at least 4 of 6 practice trials, (ii) answered all 4 attentiveness catch trials correctly, and (iii) completed the experiment within 3 SDs of the mean duration.

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`.

Foundation models are less interpretable than supervised ones. We measure interpretability with both the *localizability* and the *nameability* scores[3], reflected in Table 1. Localizability (L) and *nameability* (N) scores differ significantly across models (L: Kruskal–Wallis, $H(5) = 118.40$, $p < .001$; N: Kruskal–Wallis, $H(5) = 295.36$, $p < .001$). As seen in the table, foundation models are consistently less interpretable than their supervised counterparts (Dunn's test and Tukey HSD's test, $p < .001$ vs. ViT-B), regardless of whether they were trained with language supervision (CLIP, SigLIP) or pure self-supervised objectives (DINOv2, DINOv3). Moreover, while we expect *localizability* and *nameability* to quantify two different dimensions of interpretability, in practice we find that they are highly correlated (r=0.84, p=0.036; Fig. A.8).

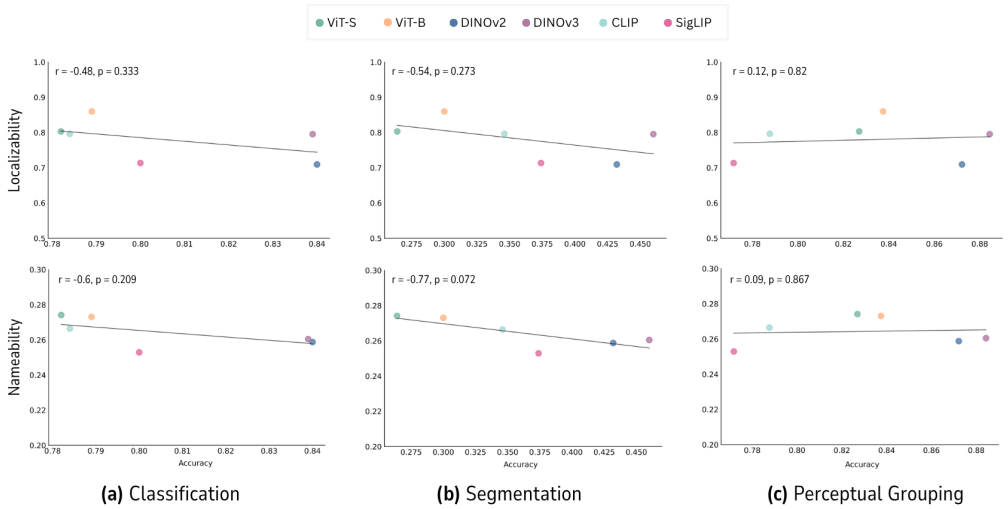
Models can feel more interpretable than they are. Average confidence is fairly homogeneous across models ($3.38 - 3.68$, Table 1, bottom row), but its alignment with interpretability is not. While DINOv2 elicits the highest confidence and ranks near the bottom on interpretability (last on *localizability*, second-to-last on *nameability*), SigLIP performs poorly on all fronts. Based on these results, a model can feel more interpretable than it is, a caution worth keeping in mind given that DINOv2 features have been treated as a reference for interpretable representations [33].

4.1 Alignment with performance

A natural question is whether interpretability and capability are linked: models with richer, more semantically structured representations might be both more capable on downstream tasks and more legible to humans. We measured the Spearman correlation between each model's interpretability scores and its performance on three vision tasks: ImageNet-1k classification, semantic segmentation, and perceptual grouping. We report those results in Fig. 2.

ImageNet-1k classification. Foundation models lead on ImageNet top-1 accuracy but trail on interpretability. Across the six models, neither *localizability* ($\rho = -0.48$, $p = 0.33$) nor *nameability* ($\rho = -0.6$, $p = 0.21$) correlate significantly with classification (Fig. 2). DINOv2 and SigLIP match CLIP on accuracy yet score roughly 8 points lower in *localizability*.

[3]For both measures, we report the median per model as scores do not follow a normal distribution, see Appendix G



(a) Classification **(b)** Segmentation **(c)** Perceptual Grouping

Figure 2: **Interpretability is uncorrelated with downstream task performance.** Each panel plots an interpretability score against a capability benchmark across the six models. Top row: *localizability*; bottom row: *nameability*. Columns: ImageNet-1k top-1 accuracy, ADE20K semantic segmentation, and perceptual grouping. Spearman ρ and p-values are shown in each panel. Correlations are non-significant for both interpretability measures across all three benchmarks, suggesting that interpretability and downstream task performance capture largely orthogonal properties of a visual representation.

Semantic segmentation. Segmentation performance on ADE20K[34]—using a linear probe[4]—does not correlate significantly with *localizability* ($\rho = -0.54$, $p = 0.27$) nor with *nameability* ($\rho = -0.77$, $p = 0.07$).

Perceptual grouping. On a perceptual grouping benchmark—a more constrained form of segmentation that probes the models' ability to bind individual object instances—we again find no significant correlation with either *localizability* ($\rho = 0.12$, $p = 0.82$) nor *nameability* ($\rho = 0.09$, $p = 0.87$; Fig. 2).

Across all three benchmarks, task performance does not predict interpretability. This dissociation suggests that the representational properties that make a model useful for downstream tasks are largely orthogonal to those that make its features interpretable to humans.

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 [35]:

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

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 correlates strongly with *localizability* ($\rho = 0.91$, $p = 0.01$; Fig. 3, top left) and *nameability* ($\rho = 0.99$, $p < .001$; Fig. 3, bottom left). See Appendix H for additional results.

[4]We follow the DINOv3 methodology but force all models to an input size of 224 × 224, since supervised ViTs from `timm` cannot take flexible input sizes. The same pipeline at 512 × 512 reproduces DINOv3 results.

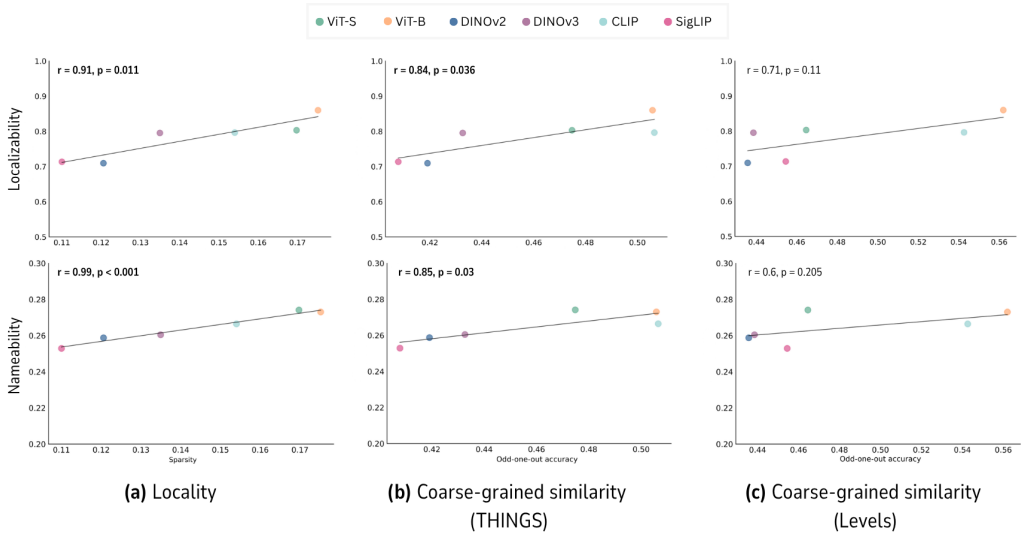
4.2 Alignment with human perception

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 [36–39]. Because not all measures of alignment capture the same properties of human perception [40], 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 [41, 42]—does not correlate significantly with *localizability* ($\rho = 0.14$, $p = 0.8$; Fig. A.9, left) or *nameability* ($\rho = 0.21$, $p = 0.7$; Fig. A.9, right).

Alignment with human similarity judgments. Human similarity judgments operate at multiple levels of abstraction [39]. *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 [43, 44], Levels [39], and NIGHTS [38, 45].

Alignment with *fine-grained* similarity does not correlate neither with *localizability* (NIGHTS: $\rho = -0.41$, $p = 0.42$; Levels: $\rho = 0.16$, $p = 0.76$; Fig. A.10, top row) nor *nameability* (NIGHTS: $\rho = -0.47$, $p = 0.35$; Levels:$\rho = 0.26$, $p = 0.62$; Fig. A.10, bottom row). In contrast, alignment with *coarse-grained* similarity is generally lower and shows a strong correlation with interpretability (Fig. 3, middle and right), which is significant for the THINGS similarity dataset: *localizability* (THINGS: $\rho = 0.84$, $p = 0.04$, Levels:$\rho = 0.71$, $p = 0.11$) and *nameability* (THINGS: $\rho = 0.85$, $p = 0.03$; Levels $\rho = 0.6$, $p = 0.21$). See Appendix I for additional results.



(a) Locality **(b)** Coarse-grained similarity (THINGS) **(c)** Coarse-grained similarity (Levels)

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: *localizability*; bottom row: *nameability*. Columns: locality of the representation (mean Hoyer sparsity over feature heatmaps), and coarse-grained alignment with human similarity judgments on THINGS [43] and Levels [39] (odd-one-out accuracy). 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 on THINGS, suggesting that interpretable representations are not only local but also organized in ways that better reflect human categorical similarity judgments.

This dissociation between granularity levels suggests that what matters for interpretability is not perceptual fidelity to fine visual detail, but rather the degree to which a model organizes its features around the same high-level semantic categories that structure human perception. A feature that

groups concepts according to what humans consider broadly similar is more likely to be a feature that humans can understand.

5 Discussion

Two protocols, 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 vision–language embedding. That they yield concordant rankings and surface the same predictors is unlikely to be coincidence—we appear to be 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 supervised counterparts. 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 [33], elicits the highest rater confidence yet ranks among the least interpretable models on *localizability* and *nameability*.

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 [19].

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, because its training objective explicitly promotes local features. This finding suggests that locality-aware objectives can partially closes 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 [46]. 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 [43] 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 *nameability* (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 supervised counterparts—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 the two structurally different protocols agree on this conclusion and on the predictors that drive it. 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 [39]—and DINOv3, whose objective explicitly promotes locality, suggests that capability and interpretability need not trade off when locality is an explicit training objective.

References

[1] 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`.

[2] 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)*, pages 8748–8763. PmLR, 2021.

[3] 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)*, pages 11975–11986, 2023.

[4] 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. *arXiv preprint arXiv:2304.07193*, 2023.

[5] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[6] Christopher Chiu, Maximilian Heil, Teresa Kim, and Anthony Miyaguchi. Fine-grained classification for poisonous fungi identification with transfer learning. *arXiv preprint arXiv:2407.07492*, 2024.

[7] Liu Shilong, Zeng Zhaoyang, Ren Tianhe, Li Feng, Zhang Hao, Yang Jie, Jiang Qing, Li Chunyuan, Yang Jianwei, Su Hang, Zhu Jun, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXIv:2303.05499*, 2024.

[8] 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.

[9] Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *European Conference on Computer Vision*, pages 367–385. Springer, 2024.

[10] 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 *Conference on Robot Learning*, pages 2679–2713. PMLR, 2025.

[11] 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.

[12] 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. *arXiv preprint arXiv:2405.01469*, 2024.

[13] 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`.

[14] Tugba Akinci D'Antonoli, Christian Bluethgen, Renato Cuocolo, Michail E Klontzas, Andrea Ponsiglione, and Burak Kocak. Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. *Diagnostic and Interventional Radiology*, 2025.

[15] 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)*, pages 63–71. IEEE, 2025.

[16] 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*, pages 4524–4536, 2025.

[17] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

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

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

[20] 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.

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

[22] 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*, pages 16543–16572. PMLR, 2025.

[23] 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*, pages 6541–6549, 2017.

[24] 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.

[25] 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, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 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`.

[26] 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. *arXiv preprint arXiv:2411.03993*, 2024.

[27] 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.

[28] 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. *Advances in Neural Information Processing Systems*, 36:54805–54818, 2023.

[29] Lewis Smith, Senthooran Rajamanoharan, Arthur Conmy, Callum McDougall, Tom Lieberum, János Kramár, Rohin Shah, and Neel Nanda. Negative results for saes on downstream tasks and deprioritising sae research (gdm mech interp team progress update 2. AI Alignment Forum, 2025.

[30] Xiaoqing Sun, Alessandro Stolfo, Joshua Engels, Ben Peng Wu, Senthooran Rajamanoharan, Mrinmaya Sachan, and Max Tegmark. Dense sae latents are features, not bugs. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

[31] 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 Deeper Networks with Magnitude Constrained Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 37813–37826, 2023.

[32] 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.

[33] 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*, pages 8212–8217, 2024.

[34] 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*, pages 633–641, 2017.

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

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

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

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

[39] 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.

[40] 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`.

[41] 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.

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

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

[44] 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.

[45] 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. *arXiv preprint arXiv:2306.09344*, 2023.

[46] Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

[47] 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. *arXiv preprint arXiv:2310.13018*, 2023.

[48] 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.

[49] 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`.

[50] 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.

[51] 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.

A Limitations and broader impact.

Limitations. We acknowledge several limitations. First, our model-level analyses span only six architectures, which limits statistical power and makes it difficult to disentangle correlated properties of representations (*e.g.*, sparsity and coarse-grained alignment may co-vary across architectures). Second, almost all models are evaluated at ViT-B scale. This controls for scale across architectures, but whether the interpretability gap between supervised and foundation models persists or reverses at larger scales remains an open question. Third, we focus on a specific set of representational properties: downstream task performance, locality, and alignment with human similarity judgments. Other factors, such as training data or robustness, may also shape interpretability. Finally, our *nameability* protocol relies on external vision-language models to score the accuracy of human descriptions. This is imperfect: using an external model introduces its own biases and cannot provide a direct ground-truth measure of whether a description captures the feature. However, directly testing free-form descriptions against the evaluated vision foundation models is not currently possible, since these models do not provide a reliable way to judge whether a natural-language description matches one of their internal visual features. We therefore use external image–text similarity as a practical proxy, keeping the scorer fixed across all evaluated models so that its biases are held constant in the comparisons.

Broader impact. This work contributes to human-centered evaluation of model interpretability. By testing whether people can localize and describe decision-relevant features, our protocol adds to a broader effort to assess whether vision models are transparent enough for settings where human oversight matters, including critical applications. However, interpretability evaluations should not be taken as evidence of safety or reliability on their own. A model whose features appear more interpretable may still fail under distribution shift or behave unpredictably in downstream applications. Our results should therefore be viewed as one component of a broader evaluation ecosystem, complementing robustness, fairness, calibration, and task-specific validation.

B Additional Related Work

Representational alignment with human perception. A growing body of work investigates how well DNN representations mirror human perceptual judgments [47], with parallel evidence that aligning model feature attributions with human importance maps improves object recognition [42]. Alignment with human similarity judgments has been shown to improve transfer [36, 37], though benefits are task-dependent: alignment helps retrieval but may hurt discriminative classification [38]. Crucially, Muttenthaler et al. [39] showed that this trade-off depends on granularity: coarse-grained and fine-grained alignment affect downstream tasks differently. While prior work treats alignment as a driver of task performance, we treat it as a potential correlate of *feature interpretability*—a connection that, to our knowledge, has not been investigated.

Automated feature scoring with vision-language models. Hernandez et al. [48] proposed labeling visual neurons with natural-language descriptions and measuring label quality via retrieval accuracy. Bills et al. [49] extended this idea to language model neurons, using GPT-4 to generate explanations scored by how well they predict held-out activations. CLIP-Dissect [50] matched each neuron to the most CLIP-similar concept from a large vocabulary, enabling scalable, open-vocabulary labeling without human annotation. Our *nameability* score builds on this spirit—participants describe a feature in free text and we measure CLIP cosine similarity between their description and feature-centered crops—while retaining the validity guarantee of human behavioral data.

C Details about the control experiment

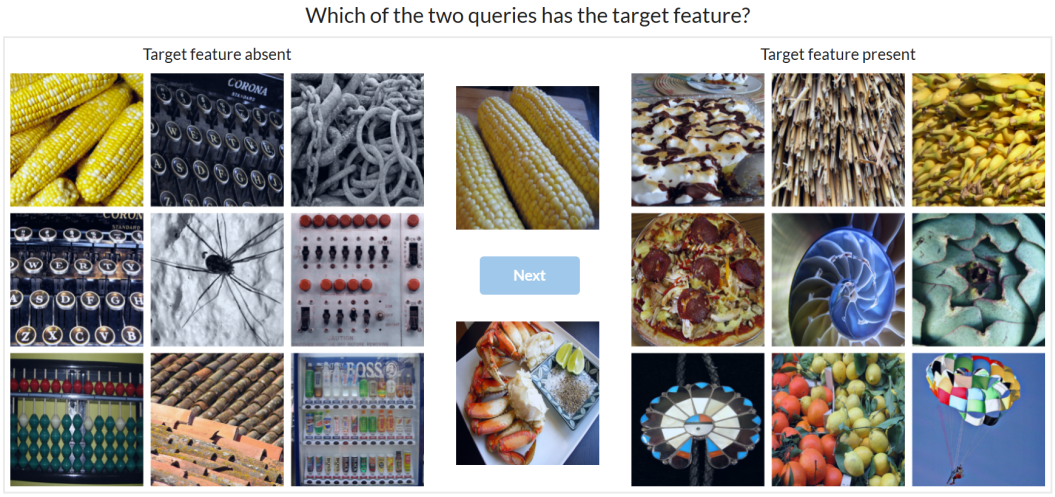


Figure A.1: Example of a trial from Zimmermann et al. [19]. In this protocol, the task is to pick the query image that matches the reference images illustrating the feature (left panel), by leveraging reference images that both highly activate the feature (right panel) and images that minimally activate it (left panel). In this example, the task can be trivially solved by relying on semantic grouping. By observation of the minimally activating stimuli (left panel), it is easy to conclude that the neuron of interest is not a corn detector, yet, it is hard to articulate what visual pattern is captured by the neuron (images in the right panel)

Figure A.2: Instantiation of the trial from Fig. A.1 in our control experiment (Sect. C). In this protocol, we replace the images that highly activate the feature with images that carry little to no information about the feature's visual content. If the protocol only quantify the degree to which participant understand the feature, we expect them to perform at chance (50%) in this setup

D Scaling model interpretability

Model-level interpretability is a property of a model's full feature population. Under a fixed annotation budget, we therefore face a choice: cover more features (*breadth*) or measure each feature more carefully (*depth*). This pilot asks which choice better approximates the population score.

What we do. We collect interpretability judgments on 80 DINOv2 neurons; this defines our reference score of 0.5555. We then use bootstrapping (200 resamples per design) to simulate two cheaper alternatives: subsampling fewer units (breadth) or reducing the number of images and trials per unit (depth). For each design we report the bootstrap mean and SD.

Strategy	**Setting**	**Bootstrap score**	**Δ vs. ref.**
Reference	All 80 units (full data)	0.5555	—
Breadth	10 units	0.5586 ± 0.0291	$+0.0031$
Breadth	30 units	0.5541 ± 0.0142	-0.0014
Breadth	60 units	0.5552 ± 0.0057	-0.0003
Depth	1 image, 1 trial	0.5475 ± 0.0376	-0.0080
Depth	1 image, 10 trials	0.5484 ± 0.0278	-0.0071
Depth	10 images, 1 trial	0.5509 ± 0.0083	-0.0046

Table A.1: **Breadth vs. depth under a fixed budget.** Bootstrap estimates of cheaper designs against the full-data reference. Breadth approaches the reference quickly; depth saturates only by exhaustively measuring the same 80 units.

What we find. Breadth converges quickly to the reference: 60 units already match it within ± 0.006. On the other hand, depth converges more slowly. Even 10 trials on a single image (± 0.0278) is less stable than 30 units alone (± 0.0142).

Why this points to scaling features. Depth makes our estimate of the units we picked more precise; breadth makes it more representative. Because the 80-unit reference is itself a stand-in for a much larger feature population, only breadth moves the estimate toward what we actually care about. We therefore allocate the main experiment's budget across as many features as possible rather than across repeated judgments on a small set.

E Nameability details

Crop vs full image For a given feature, we evaluate both the similarity of a general description—the name of the main object, and a more specific one which we believe encapsulates the feature better. Descriptions that go beyond simply naming the main object in the image generally receive a higher score when using crops (see Fig. A.3 for an illustration).

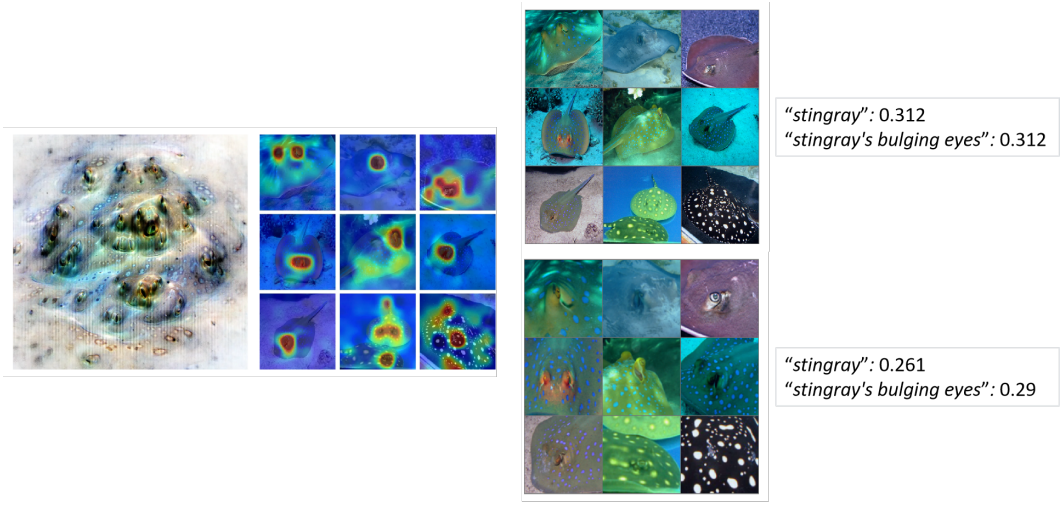


Figure A.3: For a given feature, we evaluate the similarity of two descriptions, a general one consisting of the name of the main object, and a more specific one which we believe encapsulates the feature better. Using crops allows us to reward descriptions that go beyond simply naming the main object in the image.

Regardless, Table A.2 highlights that results are fairly robust to the type of image used: *full images vs. crops*.

Table A.2: Nameability scores computed using the full image (bottom row) versus crops (top row) around the feature location. Best result per row is in bold, second best is underlined.

	Supervised		**Foundation**			
	ViT-S/16	ViT-B/16	DINOv2	DINOv3	CLIP	SigLIP
Crops (↑)	**0.274**	<u>0.273</u>	0.259	0.260	0.266	0.253
Full image (↑)	<u>0.276</u>	**0.277**	0.256	0.258	0.260	0.250

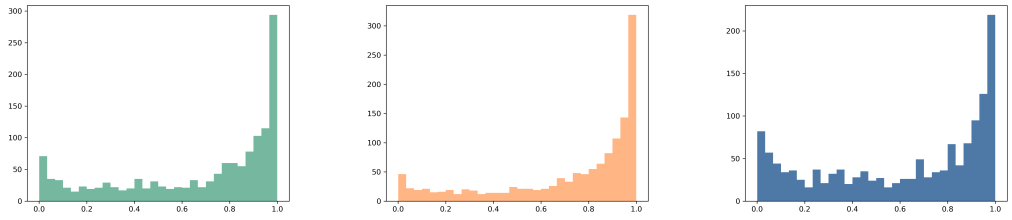
F Distribution of feature reuse across models

Table A.3: Number of unique SAE features retained after filtering dense features for each model. Differences reflect varying degrees of feature reuse across architectures.

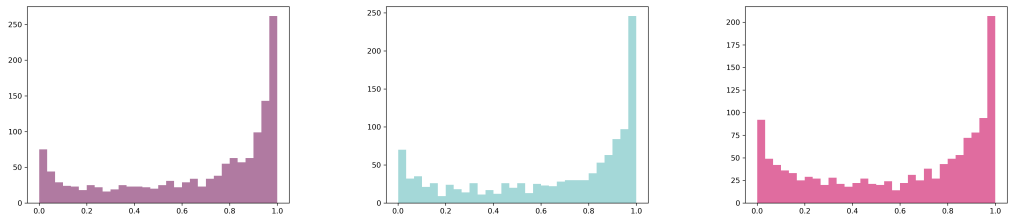
	Supervised		**Foundation**			
	ViT-S/16	ViT-B/16	DINOv2	DINOv3	CLIP	SigLIP
# Unique features	1012	1305	1119	1018	545	896

G Detailed results for the psychophysics experiments.

Across both protocols, interpretability scores were non-normally distributed (Fig. A.4 and Fig. A.5) for every model (Shapiro–Wilk test, $p < .001$). We therefore report the median score per model as a more representative measure of central tendency.

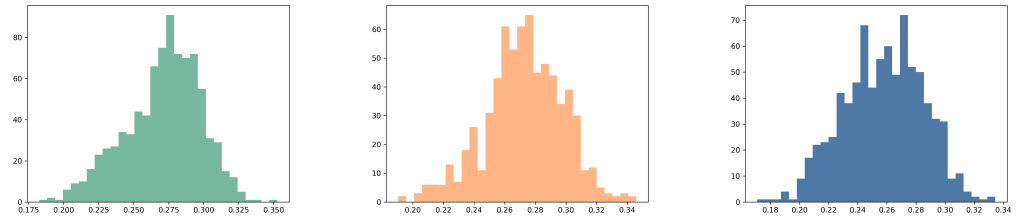


(a) ViT-S (b) ViT-B (c) DINOv2

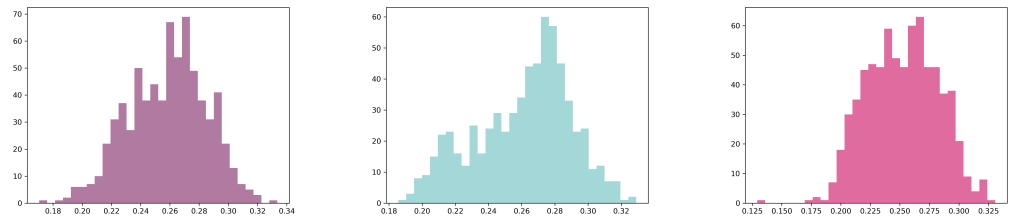


(d) DINOv3 (e) CLIP (f) SigLIP

Figure A.4: **Distribution of feature *localizability* scores across models.** Each panel shows the distribution of *localizability* scores for one model.



(a) ViT-S (b) ViT-B (c) DINOv2



(d) DINOv3 (e) CLIP (f) SigLIP

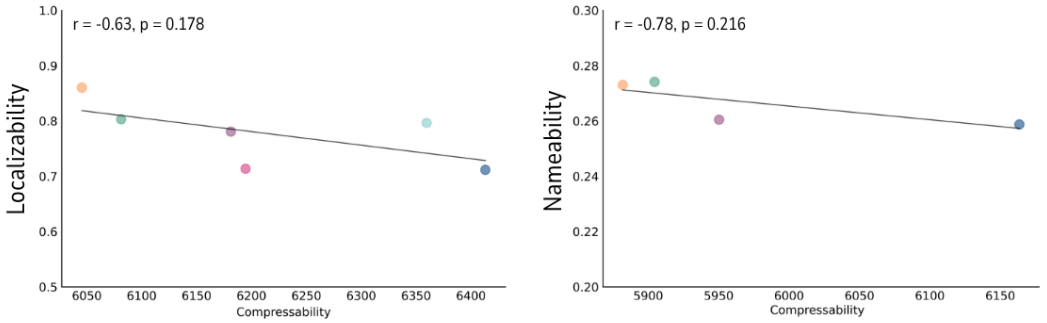
Figure A.5: **Distribution of feature *nameability* scores across models.** Each panel shows the distribution of *nameability* scores for one model.

H Additional results on locality of representations

A feature can be considered local if it fires on a few grouped pixels, i.e., (i) *few* pixels are active and (ii) those active pixels are *spatially grouped*. The Hoyer sparsity used in Section 4.1 primarily targets the first component. As a complementary proxy for the latter, we use the **Kolmogorov complexity**![51] of each feature's RISE heatmaps, approximated by their JPEG compressibility: the size of the losslessly compressed heatmap, normalized by its uncompressed size. Grouped activations form contiguous, redundant regions and compress well; scattered activations do not. Lower values therefore indicate more spatially grouped heatmaps. We average over the heatmaps of each feature's nine most activating images and then over all features for a model.

Results. Across the six models, compressibility correlates with both interpretability protocols in the expected direction (see Fig. A.6)—interpretable models tend to have more compressible heatmaps—although neither correlation reaches significance: *localizability* ($\rho = -0.63$, $p = 0.178$) and *nameability* ($\rho = -0.78$, $p = 0.216$).

Limitation. Compressibility alone does not distinguish a heatmap with a single active pixel and a heatmap where every pixel takes the same value are both highly compressible. The score therefore, captures the presence and, to some extent, the number of activation groups, but not whether those groups are spatially confined. We therefore read it as a useful complement to Hoyer sparsity: suggestive of grouping, but, on its own, an incomplete proxy for locality.



(a) Localizability **(b)** Nameability

Figure A.6: **Compressibility of feature heatmaps as a complementary measure of locality.** Each panel plots an interpretability score against the mean compressibility of a model's feature heatmaps (lower = more grouped activations). Left: *localizability*. Right: *nameability*. Spearman ρ and p-values shown per panel. Correlations point in the expected direction on both protocols but do not reach significance.

I Additional results on alignment with human similarity judgments

The Levels dataset [39] provides human similarity judgment at three different levels, by constructing triplets using 3 images from 1 class to 3 different classes. For completeness, in this section we report results from the third triplet category beyond the coarse- and fine-grained ones reported in Section 4.2. In this *class-level* category, two images share a basic-level class while the third belongs to a different basic-level class within the same superordinate category, probing a granularity that sits between fine and coarse similarity.

The Spearman correlations with class-level alignment fall between the coarse- and fine-grained results, with positive but non-significant trends on both protocols (see Fig.!A.7): *localizability* ($\rho = 0.58$, $p = 0.223$) and *nameability* ($\rho = 0.72$, $p = 0.11$). The pattern is consistent with interpretability tracking categorical alignment more closely than within-category alignment.

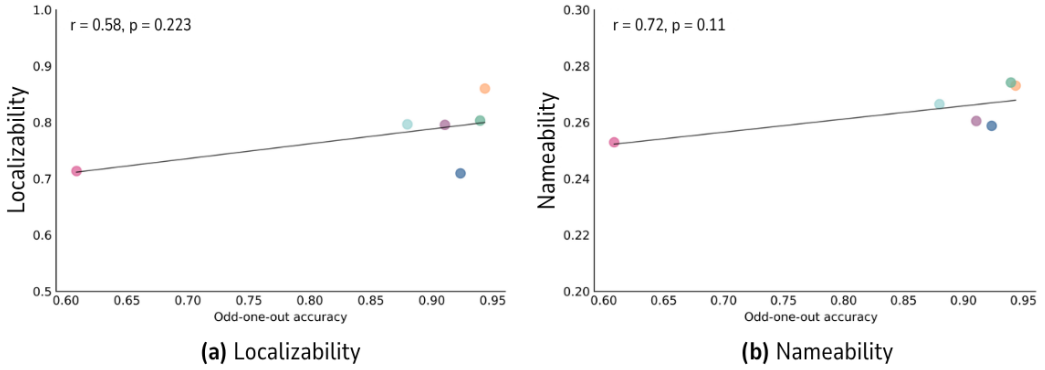


Figure A.7: **Class-level alignment on the Levels dataset.** Each panel plots an interpretability score against odd-one-out accuracy on the class-level triplets of Levels [**?**]. Left: *localizability*. Right: *nameability*. Spearman ρ and p-values per panel. Both correlations are positive but non-significant, sitting between the coarse- and fine-grained results.

J Additional figures

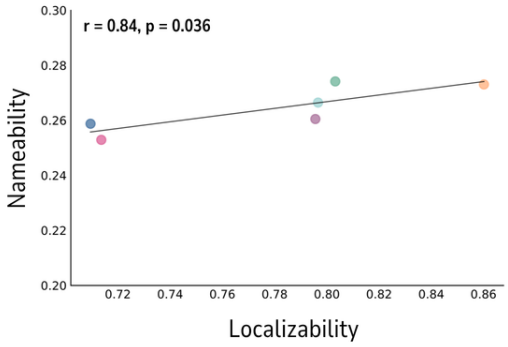


Figure A.8: **The two protocols converge on the same model ranking.** Median *nameability* plotted against median *localizability* across the six vision transformers. Despite differing in nearly every respect — pointing vs. writing, spatial heatmap vs. vision–language embedding — the two protocols produce a similar ranking of models, suggesting that we are measuring a stable property of the representation rather than an artifact of any one task.

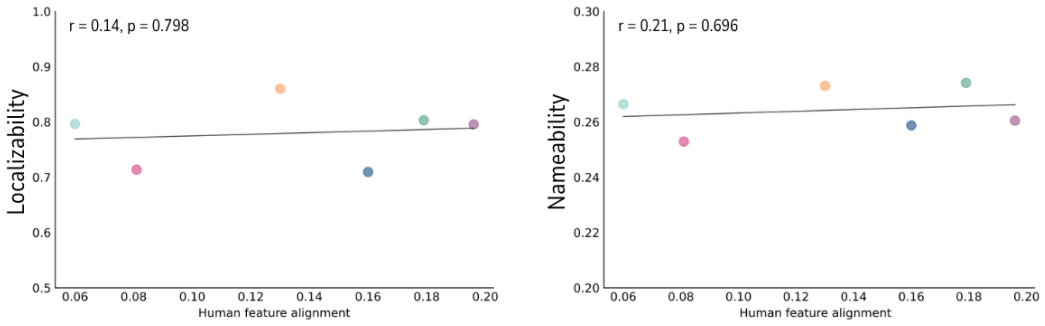
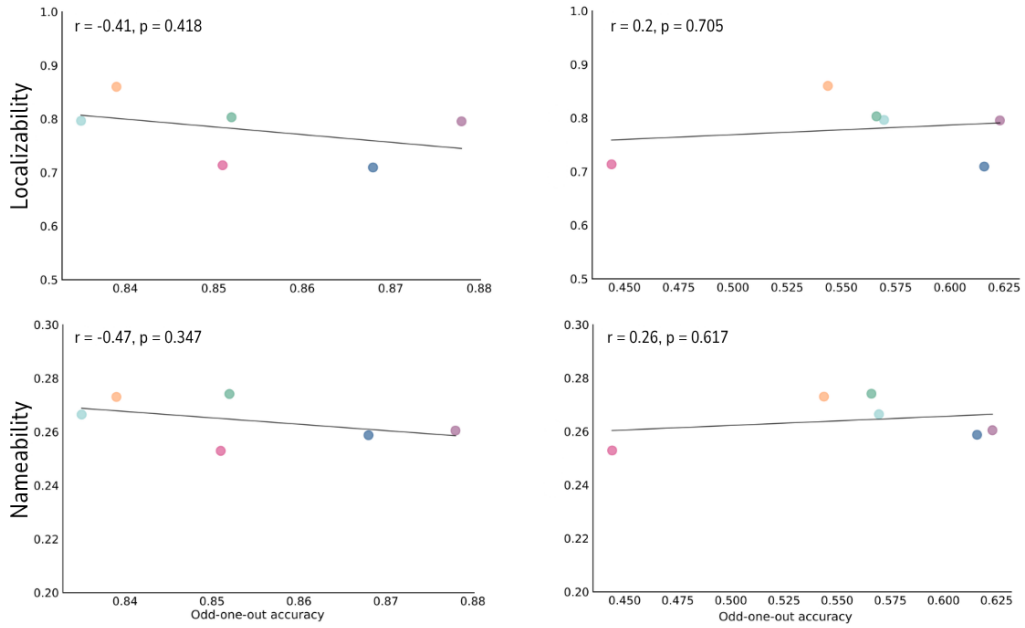


Figure A.9: **Alignment with human visual strategies does not predict interpretability.** Each panel plots an interpretability score against the overlap between model-driven and human-attended image regions [42**?**] across the six models. Left: *localizability*; right: *nameability*. Neither correlation is significant: a model can attend to the same image regions humans do and still encode what it sees there in features humans cannot readily make sense of.

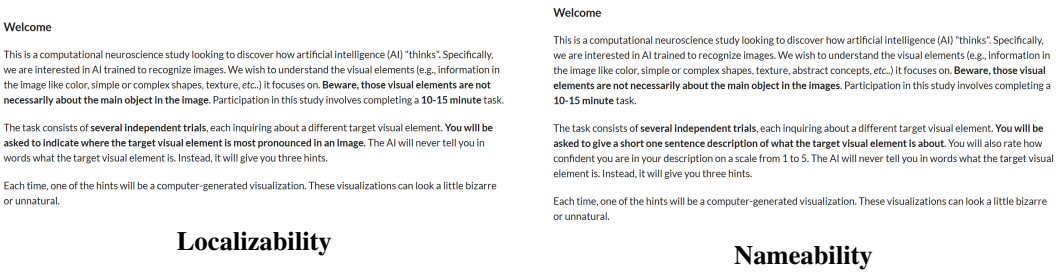


(a) Fine-grained similarity (THINGS)

(b) Fine-grained similarity (Levels)

Figure A.10: **Alignment with fine-grained similarity does not predict interpretability.** Each panel plots an interpretability score against an alignment measure across the six models. Top row: *localizability*; bottom row: *nameability*. Columns: fine-grained alignment with human similarity judgments on THINGS [43] and Levels [**?**] (odd-one-out accuracy). Spearman ρ and p-values per panel. None of the four correlations reaches significance, suggesting that what matters for interpretability is not perceptual fidelity to fine visual details.

K Psychophysics experiment



Welcome

This is a computational neuroscience study looking to discover how artificial intelligence (AI) "thinks". Specifically, we are interested in AI trained to recognize images. We wish to understand the visual elements (e.g., information in the image like color, simple or complex shapes, texture, *etc.*) it focuses on. **Beware, those visual elements are not necessarily about the main object in the image**. Participation in this study involves completing a **10-15 minute** task.

The task consists of **several independent trials**, each inquiring about a different target visual element. **You will be asked to indicate where the target visual element is most pronounced in an image**. The AI will never tell you in words what the target visual element is. Instead, it will give you three hints.

Each time, one of the hints will be a computer-generated visualization. These visualizations can look a little bizarre or unnatural.

Localizability

Welcome

This is a computational neuroscience study looking to discover how artificial intelligence (AI) "thinks". Specifically, we are interested in AI trained to recognize images. We wish to understand the visual elements (e.g., information in the image like color, simple or complex shapes, texture, abstract concepts, *etc.*,) it focuses on. **Beware, those visual elements are not necessarily about the main object in the images**. Participation in this study involves completing a **10-15 minute** task.

The task consists of **several independent trials**, each inquiring about a different target visual element. **You will be asked to give a short one sentence description of what the target visual element is about**. You will also rate how confident you are in your description on a scale from 1 to 5. The AI will never tell you in words what the target visual element is. Instead, it will give you three hints.

Each time, one of the hints will be a computer-generated visualization. These visualizations can look a little bizarre or unnatural.

Nameability

Figure A.11: Welcome screens shown to participants before the psychophysics experiments. Left: instructions for the *localizability* task. Right: instructions for the *nameability* task.

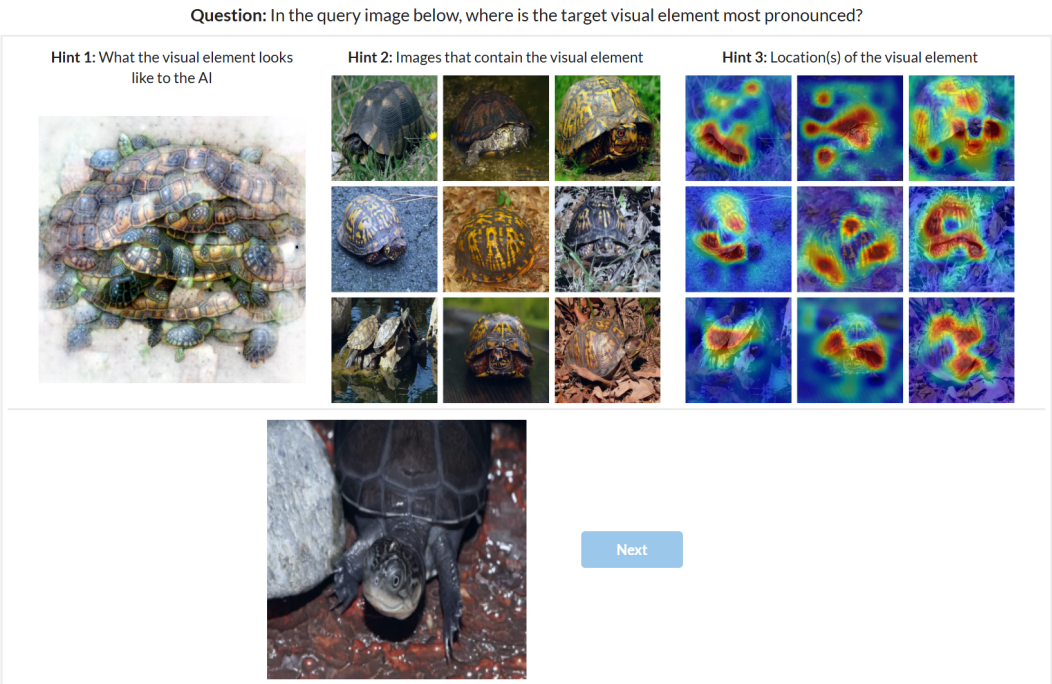


Figure A.12: Screenshot of a trial for the *localizability* experiment for DINOv3. A feature is explained through 3 panels at the top: (left) feature visualization by means of a maximally activating synthetic image, (middle) a set of images that highly activate the features, with their associated heatmaps (right) highlighting where the feature is located on those images. Participants were asked to click on a location where they expected the feature to be present on a new image (bottom). The more interpretable the feature, the more likely they are to correctly identify an area where the feature is present in the image.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope and contributions: a two-protocol framework for human interpretability, a comparison of six vision transformers, and analyses relating interpretability to downstream performance and several representational properties. These claims are supported by the Methods, Results, Discussion, and Limitations section.

   Guidelines:

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

2. **Limitations**

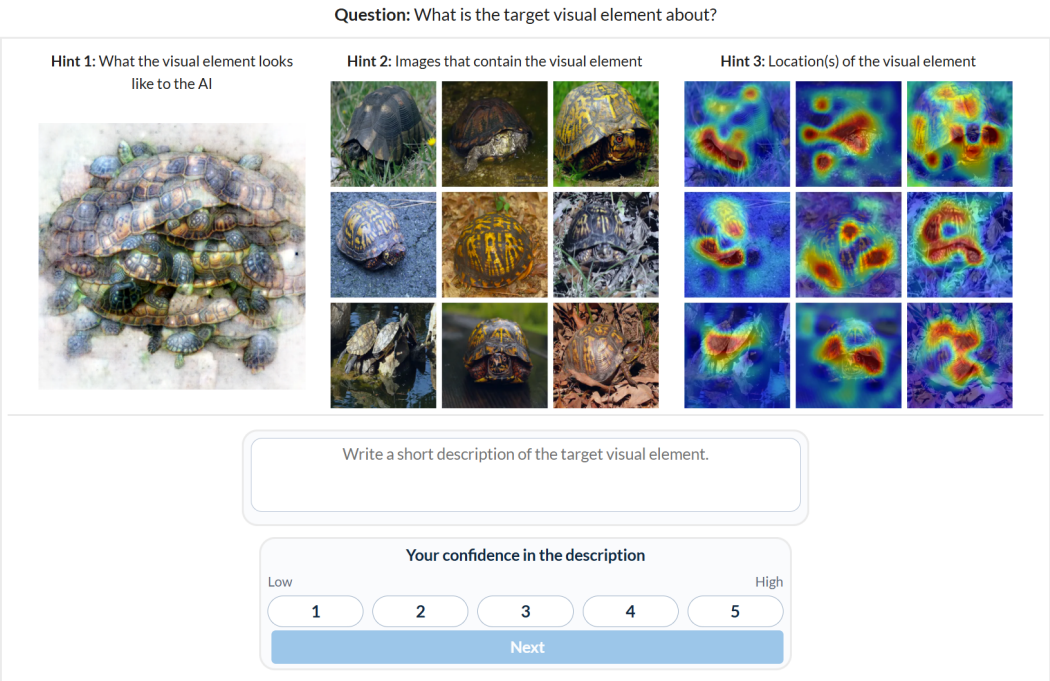


Figure A.13: Screenshot of a trial for the *nameability* experiment for DINOv3. A feature is explained through 3 panels at the top: (left) feature visualization by means of a maximally activating synthetic image, (middle) a set of images that highly activate the features, with their associated heatmaps (right) highlighting where the feature is located on those images. Participants were asked to give a text-free description of what the feature is about (bottom). The more interpretable the feature, the more accurate their description.

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

Answer: [Yes]

Justification: The paper includes a dedicated Limitations section in the Appendix discussing the limited number of models, restriction to ViT-B scale, the focus on a specific set of representational properties, the bias introduced by selecting decision-relevant features, and the limitations of the protocol.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper does not present theoretical results, theorems, or proofs. The equations define the proposed scoring procedures rather than establishing formal theoretical claims.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Because this work centers on human behavioral evaluation rather than model training, reproducibility mainly depends on the experimental protocol and data-processing pipeline. We describe the feature selection, stimulus generation, participant tasks, scoring functions, and filtering criteria, and will release processed data and analysis files to support figure reproduction and reuse.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: The submission does not yet include a public data or code release. Upon acceptance, we will make the behavioral responses, analysis scripts, and materials needed to regenerate the main results available to the research community.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Yes. Since the main contribution is a human behavioral evaluation rather than a new training method, the relevant details concern the evaluation pipeline more than optimizer or hyperparameter choices. The paper specifies the model set, feature extraction and selection procedure, stimulus construction, participant tasks, scoring functions, data filtering, and the benchmark evaluations used to contextualize the results.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The Results section reports statistical significance for the main model comparisons and correlation analyses, including Kruskal–Wallis/ANOVA tests, post-hoc comparisons, and Spearman correlations; Appendix references provide additional details where relevant.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: The paper does not yet report detailed compute requirements such as hardware type, memory, or runtime. Most experiments are human behavioral evaluations and offline analyses rather than large-scale model training, but we will add compute details for feature extraction, SAE training, and analysis where relevant.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research involves low-risk behavioral experiments with informed consent, compensation, participant filtering, and IRB approval reported in the Participants paragraph. No deceptive, harmful, or high-risk intervention is described.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The manuscript contains a broader impacts statement in the Appendix.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release high-risk models, image generators, scraped datasets, or other assets with a clear misuse risk. The evaluated models are existing public vision backbones, and the work primarily introduces an evaluation framework and human-study results.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The paper cites the main datasets, models, and methods used.

    Guidelines:

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

13. **New assets**

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

    Answer: [No]

    Justification: The paper introduces a new set of human behavioral data, which is described in the experimental sections but not yet released. We will provide the dataset and accompanying documentation upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [Yes]

Justification: Yes, details about the experiment are available through the main text or appendix.

Guidelines:

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

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

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

Answer: [Yes]

Justification: The Participants paragraph states that participants provided informed consent, were compensated, and that the protocol was approved by the IRB of an author-affiliated institution. The task is low risk and involves visual judgments and free-text descriptions of image features.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core methodology does not use LLMs as an important, original, or non-standard research component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.