Can Human Perceptual Similarity Alignment Improve Interpretability in Visual Representations?

Julien Colin[1,2] **Nuria Oliver**[1*] **Thomas Serre**[2,3*]

[1] ELLIS Alicante, Spain
[2] Carney Institute for Brain Science, Brown University, Providence, USA
[3] Department of Cognitive and Psychological Sciences, Brown University, USA

Abstract

What makes the features learned by vision models interpretable to humans? We introduce a psychophysics protocol that directly measures feature interpretability—the degree to which a person, given a feature visualization and importance maps, can predict where that feature will activate in a new image. Applying this protocol to sparse autoencoder features from a range of vision transformers, we find that self-supervised and foundation models are consistently less interpretable than their supervised counterparts. While fine-grained alignment with human similarity judgments is uniformly high across models and does not predict interpretability, coarse-grained alignment—capturing semantic rather than perceptual structure—is a stronger correlate, suggesting that further improvements in interpretability depend less on perceptual fidelity than on whether a model's representations reflect the semantic organization of human perception.

Introduction

A growing hope in the field is that aligning vision models with human perception will make them more interpretable (Fel et al., 2022; Muttenthaler et al., 2025). Yet this hypothesis remains largely untested: we lack both a rigorous measure of the interpretability of individual model features and a systematic comparison across models with varying degrees of human alignment. Here we address both gaps.

We operationalize feature interpretability through a psychophysics protocol: participants are shown a feature visualization and importance maps, and are asked to localize the feature on novel images (Fig. 2). If a feature captures a recognizable visual concept, observers should be able to find it; if it does not, they will struggle. We apply this protocol to sparse autoencoder features extracted from a diverse set of vision transformers trained under different learning frameworks.

We find that foundation models are consistently less interpretable than their supervised counterparts. Critically, while fine-grained alignment with human similarity judgments does not predict interpretability, coarse-grained alignment does—suggesting that representations capturing the semantic structure of human perception yield features that are easier for people to understand.

Methodology

Features and models For each model, we train a sparse autoencoder (SAE) (Gao et al., 2024) to decompose the activations of the last layer into a dictionary of $K = 7{,}680$ (or $K = 3{,}840$ for ViT-S) individ-

*Equal senior contribution.
Correspondence: julien_colin@brown.edu

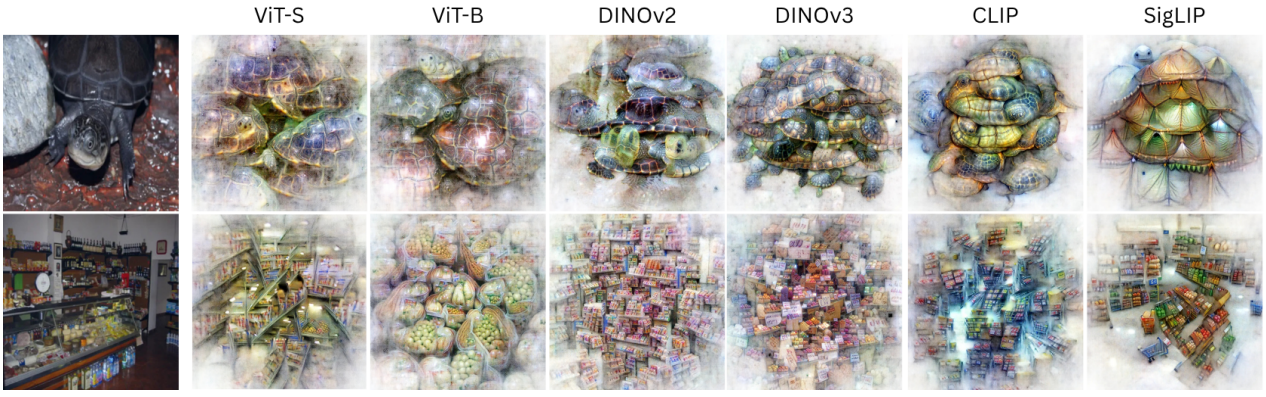

Figure 1: Most important feature for the prediction (top) *mud turtle* and (bottom) *grocery store*.

Extended abstract presented at the *9th Conference on Cognitive Computational Neuroscience*, New York, NY, USA, 2026.

ual features over the ImageNet training set (Deng et al., 2009). A linear probe is trained on top of the frozen representations to estimate feature importance. Figure 1 shows examples of extracted features for two representative classes. We evaluate a diverse set of vision transformers spanning supervised and self-supervised learning frameworks: `ViT-S/16` and `ViT-B/16` (supervised), `DINOv2 ViT-B/14`, `DINOv3 ViT-B/16`, `CLIP ViT-B/16`, and `SigLIP ViT-B/16`.

Psychophysics experiment For a given visual feature, we illustrate it via (i) a synthetic visualization (Fel et al., 2023), (ii) highly activating images, and (iii) corresponding importance heatmaps (Petsiuk et al., 2018), and we ask participants to indicate its location on a novel query image (Fig. 2). We select a representative set of features by sampling one image per superordinate class of ImageNet and retaining the top-3 most important features per image for the prediction of the model on that image (1,677 trials/model).

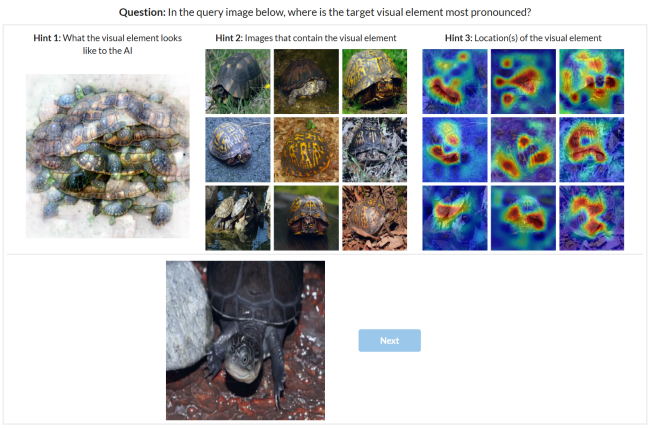

Figure 2: **Psychophysics experiment.** Example of a trial for DINOv3. The more interpretable the feature, the more likely participants are to correctly identify an area where the feature is present in the image.

Interpretability score Responses are scored against the heatmap of the query image. We compute the percentile rank of the clicked heatmap value among all pixels, then normalize it relative to the percentile of the mean heatmap activation, which defines chance performance. This maps clicks at the mean to 0.5, clicks below the mean below chance, and clicks above the mean above chance, yielding a comparable *localizability* score across features and models.

Alignment Human perceived similarity is commonly evaluated using an odd-one-out task (Fu et al., 2023; Hebart et al., 2020), where three images are presented to participants, and they must indicate which image from the triplet is least similar to the other two. Depending on how this triplet is constructed, different levels of similarity will be captured. To test whether alignment with human representations can benefit interpretability, we leverage the dataset of human coarse-grained similarity judgment from Hebart et al. (2020) collected on the THINGS dataset (Hebart et al., 2023), and we leverage the NIGHTS dataset from (Fu et al., 2023) for fine-grained similarity judgments.

Discussion

Table 1: Interpretability, accuracy and alignments score (in %) across models. Chance level is 50% for interpretability, 0.1% for accuracy, and 33% for alignments. **Bold**: best. Underline: second best.

	ViT-S	ViT	DINOv2	DINOv3	CLIP	SigLIP
Interpretability ↑	<u>80.3</u>	**86**	71	80	79.7	71.4
ImageNet Accuracy ↑	78.2	78.9	**84**	<u>83.9</u>	78.2	80
Fine-grained Alignment ↑	85.2	83.9	<u>86.8</u>	**87.8**	83.5	85.1
Coarse-grained Alignment ↑	47.5	<u>50.6</u>	41.9	43.3	**50.7**	40.8

We analyze responses from 228 participants ($\sim$40 per model). Interpretability differs significantly across models ($H(5) = 118.40$, $p < .001$), with supervised ViT-B achieving the highest score ($p < .001$, Dunn's test). Surprisingly, we find that foundation models are consistently less interpretable than their supervised counterparts ($p < .001$).

Interpretability is not predicted by classification performance ($r = -0.48, p = 0.33$), nor by fine-grained alignment with human similarity judgments ($r = -0.41, p = 0.42$). In contrast, coarse-grained alignment strongly correlates with interpretability ($r = 0.84, p = 0.04$).

This suggests that differences in interpretability are driven less by already-saturated perceptual fidelity and more by how representations organize semantic structure. While fine-grained similarity is largely shared across models (Gröger et al., 2026), coarse-grained structure reflects model-specific invariances. When these align with human perception, features appear more stable and easier to interpret (Fig. 1).

Conclusion

We introduced a human-centered measure of feature interpretability based on the ability to localize features in images. Using this framework, we find that foundation models are less interpretable than supervised models, and that interpretability is not explained by performance or fine-grained perceptual alignment.

Instead, interpretability is best predicted by coarse-grained alignment with human similarity judgments, suggesting that improving interpretability will require models to better capture the semantic organization of visual experience. These results provide initial evidence that aligning representations with human perception is a promising direction for improving interpretability and identify semantic structure as a key target for future work.

Acknowledgement

This work was funded by the ONR grant (N00014-24-1-2026), NSF grant (IIS-2402875), the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004), the Regional Government of Valencia in Spain (Resolución de la Conselleria de Industria, Turismo, Innovación y Comercio, Dirección General de Innovación) and the European Union's Horizon Europe research and innovation program (ELIAS; grant agreement 101120237). The computing hardware was supported in part by NIH Office of the Director grant S10OD025181 via the Center for Computation and Visualization (CCV) at Brown University. J.C. and N.O. are supported by Intel Corporation and J.C. is also supported by a grant by Banco Sabadell Foundation.

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fel, T., Boissin, T., Boutin, V., Picard, A. M., Novello, P., Colin, J., Linsley, D., ROUSSEAU, T., Cadene, R., Goetschalckx, L., et al. (2023). Unlocking feature visualization for deep network with magnitude constrained optimization. *Thirty-seventh Conference on Neural Information Processing Systems*.

Fel, T., Rodriguez Rodriguez, I. F., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in neural information processing systems*, *35*, 9432–9446.

Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., & Isola, P. (2023). Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*.

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

Gröger, F., Wen, S., & Brbić, M. (2026). Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., & Baker, C. I. (2023). Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, *12*, e82580.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Muttenthaler, L., Greff, K., Born, F., Spitzer, B., Kornblith, S., Mozer, M. C., Müller, K.-R., Unterthiner, T., & Lampinen, A. K. (2025). Aligning machine and human visual representations across abstraction levels. *Nature*, *647*(8089), 349–355.

Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. *Proceedings of the British Machine Vision Conference (BMVC)*.