DOES HUMAN-ALIGNMENT BENEFIT INTERPRETABILITY?

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

Aligning deep neural networks with human perception has been shown to benefit visual representations across a wide range of settings, from depth estimation to generalization capability, motivating the hope that injecting human knowledge into models may also improve their interpretability. However, whether and how alignment with human perception improves the interpretability of visual features remains unclear.

In this work, we introduce an experimental protocol to quantify the interpretability of visual features in deep neural networks and use it to probe the relationship between alignment and interpretability by comparing aligned and non-aligned models of the same family. In particular, we compare the interpretability of three transformer-based models: DINOv2, a variant of DINOv2 that has been aligned with human similarity judgments, and DINOv3. Our results illustrate a trend whereby alignment benefits the interpretability of learned representations, with the aligned model being significantly more interpretable than both non-aligned counterparts. Visual inspection further highlights a profound qualitative shift in the learned features: aligned features tend to be more spatially localized, yet qualitatively less visually rich. Together, these findings provide the first concrete evidence that alignment with human perception can enhance interpretability, while also underscoring that the mechanisms by which alignment reshapes visual representations remain only partially understood.

1 INTRODUCTION

Recent years have seen growing interest in aligning vision-based deep neural networks (DNNs) with human perception (Sucholutsky et al., 2023). This line of research is motivated by the observation that human perception provides a strong inductive bias for learning efficient and semantically meaningful representations—*i.e.*, representations that are robust and generalize well. Human similarity judgments are of particular interest, as they play a central role in cognition—supporting categorization, memory, and reasoning—and can be seen as a more general objective than object classification (Roads & Love, 2021; 2024). Accordingly, aligning DNN representations with human similarity judgments has been shown to improve few-shot learning (Sucholutsky & Griffiths, 2023; Muttenthaler et al., 2023), robustness, and the transferability of features across tasks (Sucholutsky & Griffiths, 2023). More recently, alignment with human similarity judgments has been shown to improve representation quality over a wide range of downstream tasks, suggesting that incorporating human perceptual knowledge can lead to better representations (Sundaram et al., 2024).

At the same time, the benefits of human alignment are not uniform. While it can improve several downstream tasks—such as depth estimation, instance-retrieval or counting, other tasks appear to benefit less, and some tasks may even suffer from alignment, for example, natural image classification (Sundaram et al., 2024). These trade-offs have been attributed not only to increased feature localization, but also to shifts in representational structure that favor human-like similarity at the expense of class-discriminative signals. These findings suggest that alignment does not universally

*Equal senior contribution. E-mail: julien_colin@brown.edu

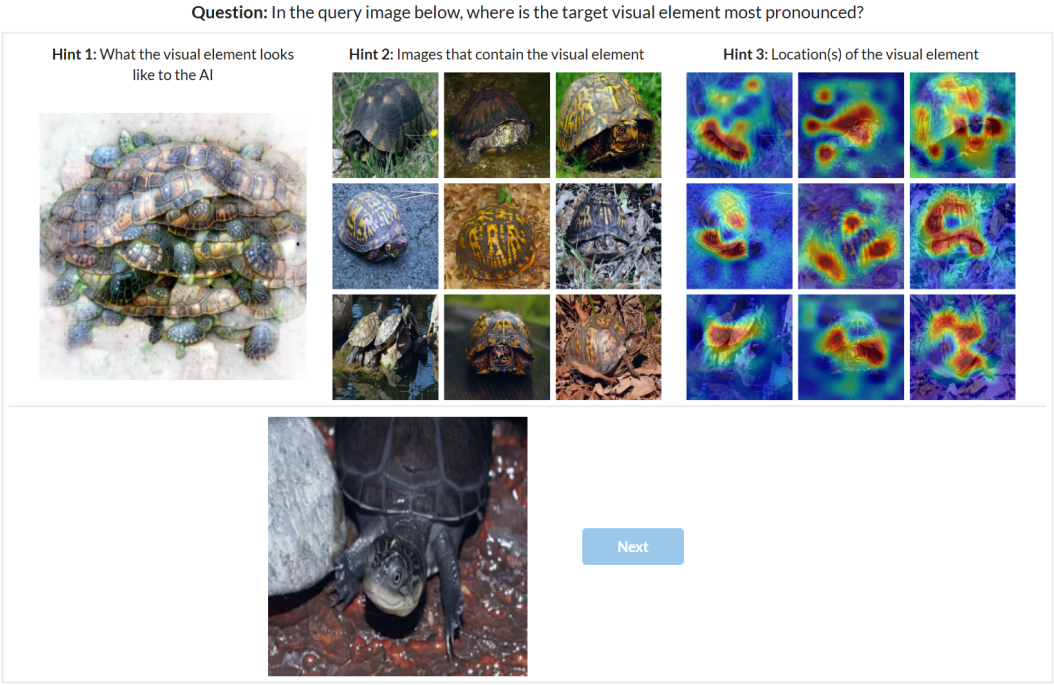

Figure 1: **Psychophysics experiment.** Example of a trial for DINOv3. A feature is explained through 3 panels at the top: (left) feature visualization by means of a maximally activating synthetic image, (middle) a set of images that highly activate the features, with their associated heatmaps (right) 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.

improve representation quality and that different tasks may rely on different representational properties. While prior work has shown that aligning representations with similarity judgments impacts the structure of the representation (Muttenthaler et al., 2023)—capturing typically local, human-like similarities but failing to reflect the global similarity structure— it remains unclear whether, and in what ways, alignment qualitatively alters the underlying visual features.

A common assumption underlying alignment research is that aligning models with human perception should lead to more human-like features, which in turn would yield more interpretable representations (Fel et al., 2022; Sundaram et al., 2024; Muttenthaler et al., 2025). Yet, direct evidence that alignment improves interpretability remains scarce. As the factors that determine the interpretability of DNNs—and their internal representations—remain an open research question, understanding whether and how alignment influences interpretability is of substantial practical importance.

While interpretability admits many definitions, we focus on two aspects that are particularly relevant: understanding (a) the computations underlying a model's decision and (b) the individual representational elements upon which those computations are based. In this paper, we concentrate on the latter and investigate whether alignment makes the visual features learned by a DNN more interpretable to humans.

Concretely, we assess the interpretability of a visual feature by measuring the extent to which humans are able to identify its location within an image (see Fig. 1 for an example trial). Just as interpretability can be defined in multiple ways, models can also be aligned with different aspects of human perception. In this paper, we focus specifically on alignment of their internal representations with human similarity judgments. We empirically evaluate the interpretability of 3 transformer-

based vision models: DINOv2, a variant of DINOv2 aligned with human similarity judgments, and DINOv3. Our contributions are as follows:

- We propose a novel experimental protocol to quantitatively assess the interpretability of individual visual features.
- We apply this protocol to compare the interpretability of aligned vs non-aligned models, directly investigating the relationship between alignment and interpretability.
- We find evidence of a trend whereby alignment with human perception improves interpretability of visual representations. Specifically, the aligned model exhibits significantly higher interpretability than both non-aligned counterparts.
- Beyond quantitative improvements, we identify a qualitative shift in the nature of the learned representations, whereby alignment yields features that are more spatially localized yet qualitatively less visually rich.

2 METHODS

2.1 MODELS

DINOv2 Oquab et al. (2023) introduces DINOv2, a self-supervised vision transformer that learns powerful visual representations, enabling it to support a wide range of downstream tasks. Owing to the quality of its learned representations, DINOv2 has become widely adopted as a general-purpose vision backbone (Liu et al., 2024; Yu et al., 2024; Lin et al., 2025; Carion et al., 2025). In addition to its robustness, DINOv2 produces visually rich representations, making it a natural choice for evaluating the human interpretability of visual features (Dreyer et al., 2024).

Aligned DINOv2 For direct comparison, we include a human-aligned variant of DINOv2 ($DINOv2_{aligned}$), created by fine-tuning the original model using human perceptual similarity judgments (Fu et al., 2023). The authors introduce the NIGHTS dataset, which consists of human similarity preferences collected through three-alternative forced-choice (3AFC) experiments over image triplets. They fine-tune DINOv2 on the NIGHTS dataset using a contrastive objective designed to align embedding-space distances with human similarity judgments. By more closely approximating human perceptual similarity, this alignment procedure is expected to suppress variations that humans are less sensitive to—such as image distortions or fine-grained texture—potentially yielding more interpretable visual features than those of the original DINOv2.

DINOv3 As an additional baseline, we also consider DINOv3 (Siméoni et al., 2025). Building on DINOv2, DINOv3 introduces a larger-scale training regime, improved data filtering, and enhanced architectural and optimization strategies, resulting in more robust and localized visual features. As a result, DINOv3 provides a strong non-aligned baseline whose representations are expected to be of higher quality than DINOv2, while incorporating properties that may independently benefit interpretability.

2.2 METHODOLOGY

Experimental protocol We subscribe to the view that interpretability is intrinsically a human property, and therefore requires human evaluation. Within this line of research, Borowski et al. (2021) investigated the interpretability of DNN representations by presenting participants with sets of images that elicited maximal or minimal activation of a given feature, and subsequently asked them to identify which of two query images would also elicit maximal activation of that feature. Zimmermann et al. (2021) adapted this protocol by modifying the task so that participants were asked to predict the effect of an intervention (*e.g.*, occluding a specific image region) on the activation of a feature. Building upon this body of work, we further improve the protocol by placing participants in a setting where they directly indicate the location of the visual feature on the query image, enabling a more precise assessment of the understanding that humans have of a specific visual feature. This improvement addresses the limitation of previous protocols, where the task could be trivially solved if the two query images were semantically very different. Figure 1 illustrates the task at hand.

Feature extraction Interpreting individual neurons in modern vision models is often challenging, as single neurons can encode multiple unrelated features—a phenomenon commonly referred to as superposition (Elhage et al., 2022). A promising alternative for extracting interpretable features consists of training a sparse autoencoder (SAE) over layer activations. This approach recovers a dictionary of sparse latent features, which are expected to be easier to interpret than the activations of individual neurons. Thus, we train SAEs on the activations from the entire ImageNet training split, using approximately 1.28M images and all corresponding patch tokens—256 for DINOv2 and 196 for DINOv3. The SAEs are trained with an expansion factor of $\times 10$, yielding a dictionary of $7{,}680$ latent features (10x the 768-dimensional input).

Feature relevance and linear probing We train a linear classification head on ImageNet on top of the frozen self-supervised backbone of each of the models under consideration. This linear probe serves two complementary purposes. First, it provides a quantitative measure of how alignment with the learned representations impacts downstream classification performance. Second, it allows us to compute feature importance scores, which quantify the contribution of each feature to the classifier's predictions. We then leverage these scores to identify features that are both strongly activated and influential for the model's decision-making process.

Feature selection To obtain a representative subset of features, we aim to select features that are both diverse and relevant to the model's decisions. To this end, we randomly sample one image for each of the 559 superordinate classes underlying the 1,000 ImageNet labels, yielding a fixed set of 559 images. For each image, we select the three visual features that rank as the most important for the model's prediction on that image. This procedure yields a total of $559 \times 3 = 1{,}677$ trials per model.

Trials If a visual feature is interpretable, we expect participants to be able to identify where it appears in an image. To evaluate that, we illustrate a feature through (a) a feature visualization, *i.e.,* a synthetic image that maximally activates the feature, generated using MACO (Fel et al., 2023) (left-most image in Fig. 1); (b) a set of 9 images where the feature is highly activated (central panel in Fig. 1); and (c) their corresponding heatmaps highlighting where the feature is located on those images using RISE (Petsiuk et al., 2018) (right-hand panel in Fig. 1). Participants are asked to click where they expect the feature to be present on a new image (bottom image in Fig. 1).

Scoring We score each participant's response using the heatmap of the query image, smoothed with a Gaussian filter so that each location reflects the importance of a local region rather than an individual pixel. We then extract the importance value v at the click location and evaluate how it ranks relative to all pixels in the heatmap using its empirical cumulative distribution function (ECDF). To enable fair comparison across features and models, we apply a mean-aware normalization of this rank, such that chance-level performance corresponds to selecting locations with typical (mean-level) activation. Full details of the scoring procedure are provided in Appendix D.

3 Results

3.1 Alignment & Classification

First, we measure the alignment of all models on the NIGHTS dataset used to finetune DINOv2, as well as their classification performance on the ImageNet validation set. Consistent with Sundaram et al. (2024), we find that increased alignment with human judgments is associated with reduced downstream classification performance: the alignment of DINOv2 comes with a 7% drop in classification performance. Interestingly, DINOv3 achieves higher alignment than DINOv2 while maintaining a comparable classification performance.

3.2 Interpretability

To measure the interpretability of each of the three models, we carried out a psychophysics experiment with the previously described experimental protocol.

A total of 162 participants were recruited through the Prolific [1] online platform. After applying quality and attentiveness criteria, we analyzed responses from 119 native English-speaking participants with normal vision, who completed the study on a desktop or laptop. Further details about the participants can be found in the Appendix.

Models	ImageNet Accuracy	Alignment	Interpretability
$DINOv2$	**84**	86.8	71.2
$DINOv3$	83.9	87.8	78.1
$DINOv2_{aligned}$	77	**95.5**	**86.8**

Table 1: ImageNet validation set accuracy, Alignment with the NIGHTS human similarity judgment dataset, and median interpretability score (across all 1,677 trials) as evaluated with the protocol proposed in this paper. Higher is better.

Responses were analyzed using linear mixed-effects models with the interpretability score as the dependent variable. Models included fixed effects for model identity and feature importance rank, and random intercepts for image and participant to account for shared stimulus difficulty and individual differences. The effects of model and rank were assessed using Type III Wald F-tests, and pairwise model comparisons were performed using estimated marginal means with Holm correction for multiple comparisons. Models were fit in R, with degrees of freedom estimated via Satterthwaite's approximation.

Based on our analyses, we identify the following three main findings.

Alignment benefits Interpretability The median interpretability score across all trials is 71.2% for DINOv2, 78.1% for DINOv3, and 86.8% for $DINOv2_{\text{aligned}}$. Although interpretability increases monotonically with alignment ($DINOv2 < DINOv3 < DINOv2_{\text{aligned}}$), our analysis reveals that only $DINOv2_{\text{aligned}}$ is significantly more interpretable than both DINOv2 and DINOv3 (F(2, 3713.8) = 74.09, $p < .001$), while no significant difference is observed between the non-aligned models. Interestingly, we also find weak but consistent evidence that model differences are concentrated among the most important features, as reflected by a significant effect of feature rank on interpretability (F(2, 3818.0) = 15.68, $p < .001$).

Alignment qualitatively alters visual features We illustrate representative trials for the different models in Figures 1, 2 and 3, and visualize the most important feature for 6 randomly sampled images using maximally activating stimuli in Figure 4. The most salient effect of alignment is a qualitative shift in the nature of the learned representations: aligned features are markedly more localized and appear to rely on less global context than those of either baseline model, including DINOv3, which was explicitly trained to promote spatially localized representations. Consistent with this observation, heatmaps for highly activating images of aligned features consistently highlight similar, spatially confined regions. This increased localization provides a plausible explanation for the observed drop in classification accuracy after alignment.

4 Conclusion

In this paper, we investigate how aligning visual representations with human similarity judgments affects interpretability. Through a psychophysics study involving 119 participants, we find that alignment significantly improves the interpretability of the representation, albeit at the cost of classification performance. Alignment also induces qualitative changes in the learned features, which become more spatially localized and less dependent on global context. This is reflected both in feature heatmaps that consistently highlight similar, spatially confined regions, and in maximally activating stimuli that appear visually less rich than those of the non-aligned models. Together, our results show that alignment reshapes not only the structure and functional properties of the learned representations, but also the nature of their visual features, underscoring the need for further research to understand the mechanisms through which alignment induces these transformations.

[1] www.prolific.com

ACKNOWLEDGEMENT

This work was funded by the ONR grant (N00014-24-1-2026), NSF grant (IIS-2402875) and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004). 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. have been partially supported by Intel Corporation and funding from the Regional Government of Valencia in Spain (Resolución de la Conselleria de Industria, Turismo, Innovación y Comercio, Dirección General de Innovación). J.C. has also been partially supported by a grant by Banco Sabadell Foundation and funding from the European Union's Horizon Europe research and innovation program (ELIAS; grant agreement 101120237).

REFERENCES

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

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

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*, pp. 8212–8217, 2024.

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

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*, 35:9432–9446, 2022.

Thomas Fel, Thibaut Boissin, Victor Boutin, Agustin Martin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom ROUSSEAU, Remi Cadene, Lore Goetschalckx, et al. Unlocking feature visualization for deep network with magnitude constrained optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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.

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

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

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*, 36:50978–51007, 2023.

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

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.

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.

Brett D Roads and Bradley C Love. Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 3547–3557, 2021.

Brett D Roads and Bradley C Love. Modeling similarity and psychological space. *Annual Review of Psychology*, 75(1):215–240, 2024.

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.

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

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

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*, 37:55314–55341, 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Roland S. Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas Wallis, and Wieland Brendel. How well do feature visualizations support causal understanding of cnn activations? *Advances in Neural Information Processing Systems*, 34:11730–11744, 2021.

A LIMITATION

This work represents a first step toward an empirical investigation of how aligning models with human perception impacts interpretability. To keep the study tractable, we focus on a restricted setting: alignment with human fine-grained similarity judgments, a single family of models (DINO), and a specific definition of interpretability based on feature localization. While this limits the scope of our conclusions, it nevertheless provides initial evidence that alignment induces a qualitative shift in the nature of the learned features, a phenomenon that remains only partially understood. Whether these effects extend to other forms of alignment, model families, or definitions of interpretability remains an open question for future work.

B HUMAN SIMILARITY JUDGMENTS AND INTERPRETABILITY

Using a large-scale dataset of human similarity judgments over image triplets, Hebart et al. (2020) learn a model that accurately predicts human judgments. They further examine the dimensions recovered by this model through a combination of qualitative and quantitative evaluations, providing evidence that these dimensions are meaningful and interpretable. Qualitatively, 20 participants are asked to assign semantic labels to each dimension while viewing images sorted by their values along that dimension; the responses are summarized as word clouds, and the most prevalent label is taken to characterize each dimension. Quantitatively, 20 participants are shown images spanning a given dimension and asked to indicate where novel query images fall along it; averaging these responses yields a human-predicted similarity matrix that strongly correlates with the model-derived one. Together, these findings suggest that the learned dimensions capture interpretable structure.

In contrast, subsequent work building on this line of research typically does not directly evaluate interpretability (Muttenthaler et al., 2023; 2025). Instead, improved prediction of human similarity judgments is used as a proxy for interpretability. However, there is limited empirical evidence directly assessing whether the resulting dimensions or features of these aligned models are actually understandable to humans.

APPENDIX

C PARTICIPANTS IN THE PSYCHOPHYSICS EXPERIMENT

A total of 162 participants were recruited through the Prolific [2] online platform. All participants were native English speakers who reported no visual impairments and completed the study on a laptop or desktop computer (not mobile devices). They provided informed consent electronically and were compensated \$2.95 for their time, corresponding to $\tilde{\$}$16 USD per hour (approximately 10–13 minutes). The protocol was approved by the Institutional Review Board (IRB) of an institution affiliated with the authors. We analyze the responses of 119 participants who (1) succeeded in at least 4 of the 6 practice trials, (2) correctly answered all 4 catch trials (attentiveness tests) randomly inserted throughout the experiment, and (3) completed the experiment within 3 standard deviations of the mean completion time for that experiment.

[2] www.prolific.com

D SCORING DETAILS

We assign each participant click a score $s \in [0, 1]$ that quantifies the degree to which the feature is present at the selected location, relative to all possible locations in the image.

Let x_i denote the normalized heatmap value at pixel i and n the total number of pixels. Given a participant click with associated value v, we first compute its empirical cumulative distribution function (ECDF):

$$p = \widehat{F}(v) \;=\; \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \leq v)\,.$$

This percentile measures how highly the clicked location ranks within the heatmap. However, because the distribution of heatmap values varies across features (e.g., sparse vs. dense), raw percentiles are not directly comparable across features and models.

To address this, we introduce a mean-aware normalization. Rather than interpreting percentiles in absolute terms, we measure their deviation from the typical activation level of the heatmap, as captured by its mean intensity μ. This ensures that selecting a location with average activation corresponds to chance-level performance.

Concretely, we obtain the score s by normalizing the ECDF value p around $p_\mu = \widehat{F}(\mu)$ so that the mean maps to 0.5, while the minimum and maximum map to 0 and 1:

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

E ILLUSTRATION OF TRIALS

We illustrate the same trial (image 24, the most important feature for the image classification) illustrated in Fig 1 but for DINOv2 and DINOv2 aligned.

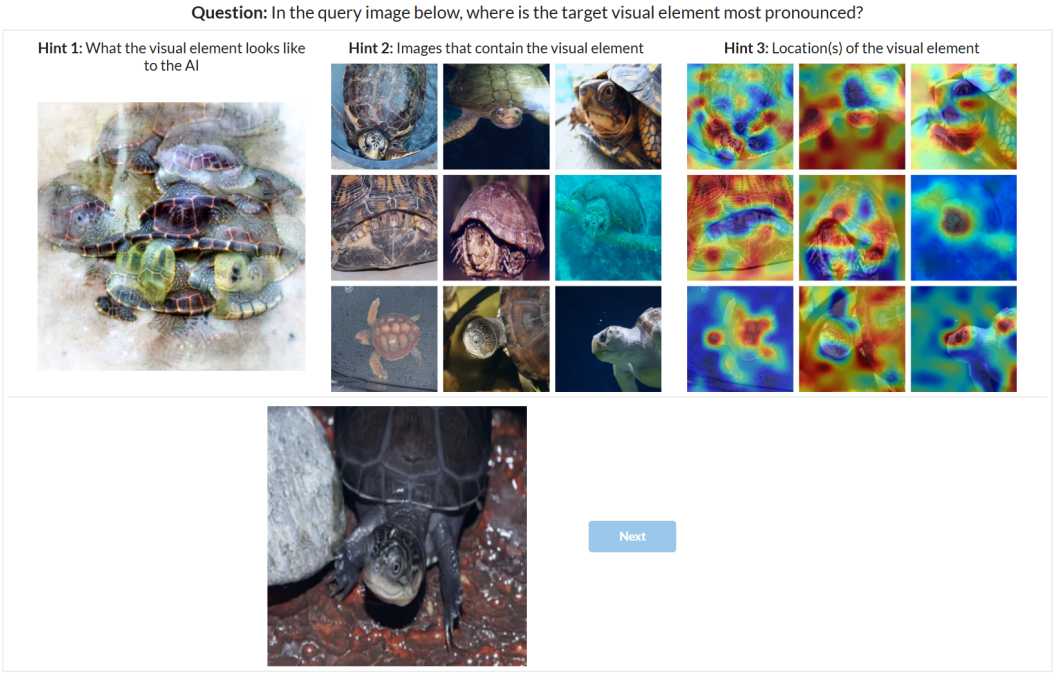

Figure 2: **Psychophysics experiment.** Example of a trial for DINOv2. A feature is explained through 3 panels at the top: (left) a feature visualization by means of a maximally activating synthetic image, (middle) a set of images that highly activate the features, with (right) their associated heatmaps 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.

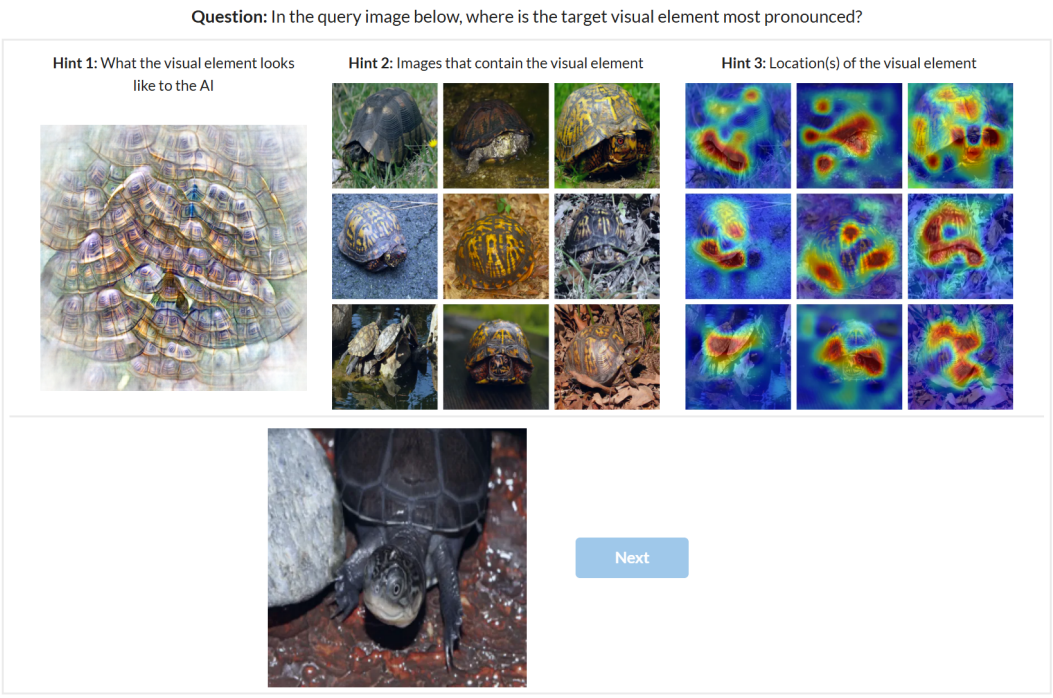

Figure 3: **Psychophysics experiment.** Example of a trial for $DINOv2_{aligned}$. A feature is explained through 3 panels at the top: (left) a feature visualization by means of a maximally activating synthetic image, (middle) a set of images that highly activate the features, with (right) their associated heatmaps 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.

F ILLUSTRATION OF FEATURES

We illustrate the qualitative differences between the features of the three models considered in Fig 4.

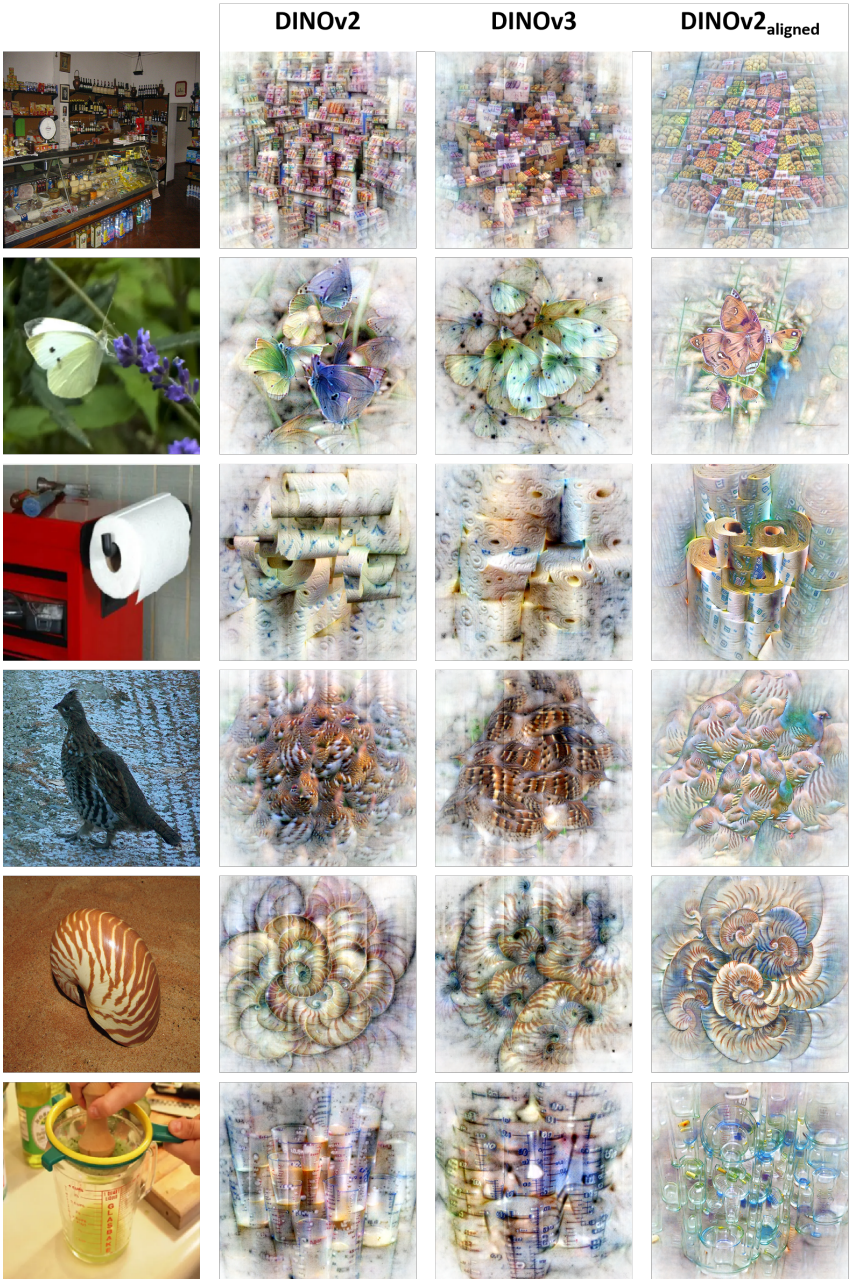

Figure 4: The most important features for 6 randomly sampled images for (left) $DINOv2$, (middle) $DINOv3$ and (right) $DINOv2_{aligned}$