Does Human-Semantic Alignment Make Model Features Easier to Learn?

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

Thomas Serre*
Brown University, USA

Nuria Oliver*
ELLIS Alicante, Spain

Abstract

Different neural networks can develop similar representations of the world, and dimensions shared across vision models often capture concepts that humans can recognize and understand. But what does similarity to human representations enable? Prior work suggests that representational alignment between a teacher and a human learner improves learning, and that models more aligned with human semantics tend to have more interpretable features. This motivates the hypothesis that aligning models toward human semantic structure might make their internal features easier for humans to learn, and thus to interpret. We test this hypothesis by fine-tuning five differently pretrained vision transformers toward human semantic similarity judgments. We measure *feature learnability* with a few-shot learning task: observers see a few examples illustrating a feature, then predict where it activates on a novel image. Alignment consistently increases agreement with human semantic judgments and makes the models more similar to one another. Yet feature learnability moves in different directions depending on the model. Thus, alignment can help, but not reliably. These results raise the question of the interplay between human semantic similarity alignment, human individual feature learning and interpretability.

1 Introduction

Vision models trained with different data and objectives can develop similar representations of the world [1]. This shared structure also connects to humans: dimensions that recur across vision models tend to align with human visual cortex activity [2], capture conceptual properties, predict human similarity judgments, and be interpretable to human observers [3]. Human similarity judgments, in turn, can be used to reshape model representations [4, 5]. Aligning different models toward this common target can thus make them more similar both to humans and to one another [6]. This raises a functional question: what would such similarity enable?

One candidate benefit concerns learning. When a teacher's representations are aligned with a learner's, teaching can become more effective, including for human learners, although the benefit depends on the task [7]. Across pretrained vision models, those more aligned with *coarse* human semantic judgments also tend to have more interpretable features [8]. Together, these findings suggest that a model whose representations are more aligned with human semantic structure might have features that humans find easier to learn, and thus to interpret.

We call this property *feature learnability*: how well humans can infer a feature's behavior from a few examples and generalize to novel images. It provides a functional measure of interpretability, consistent with a view that understanding models involves communicating machine concepts to humans [9]. We extend the question of whether representational alignment facilitates learning to

*co-principal investigators

Preprint.

the model's own internal features: does aligning how a model relates whole images with human judgments make its individual features easier for humans to learn and thus to interpret? The answer is not obvious. A model can agree with human similarity judgments while relying on different dimensions, for instance, more visual and less semantic ones [10]. Moreover, the link between alignment and interpretability has so far been observed only across pretrained models that differ in many other ways [8]. Addressing the question requires inducing similarity to humans directly.

Human semantic-similarity judgments offer a natural target for this test, as they reveal meaningful dimensions of conceptual organization [11]. We fine-tune five differently pretrained vision transformers toward these judgments. We then measure their similarity to humans and to one another, and their feature learnability, using a few-shot learning task in which observers predict where a feature activates on a novel image. While alignment consistently increases both forms of similarity, it does not increase feature learnability: alignment can help, as for DINOv2, but not reliably, and feature learnability spreads further apart across models even as they become more similar in how they relate images. These results raise the question of *which aspects of a representation should become more similar to human ones* for its features to become easier to learn by humans.

2 Experimental Setup

Models and alignment. We build on the five ViT-B/16 backbones evaluated by Colin et al. [8]: supervised ViT-B, DINOv2, DINOv3, CLIP, and SigLIP. We fine-tune each model toward human semantic structure using odd-one-out judgments from THINGS [12] and the contrastive triplet objective of Muttenthaler et al. [4]. Low-Rank Adaptation (LoRA) [13] is applied to attention projections while the remaining backbone weights are frozen. A temporary projection head is used during training and discarded before evaluation (App. A).

Measuring representational similarity. We evaluate agreement with human similarity judgments using THINGS odd-one-out accuracy and the held-out Levels dataset [6], which assesses semantic similarity at different levels of abstraction, including between-class, within-class, and class-boundary relationships. To assess similarity between models, we use representational similarity analysis (RSA) [14]. For each model, we construct a representational dissimilarity matrix using pairwise cosine distances between image representations, then compute the Spearman correlation between model pairs before and after alignment. This measures similarity in how models organize images, rather than correspondence between their individual features.

Feature learnability. Following Colin et al. [8], we extract informative features from models using independently trained TopK sparse autoencoders (SAEs) [15] (App. C) and evaluate their learnability through *localizability*. Human observers inspect a synthetic feature visualization, highly activating images, and corresponding importance maps, then predict where the feature will activate in a novel image (see Fig. A.1 for an illustration of a trial). We score each prediction using the empirical cumulative distribution function (ECDF) rank of the feature's activation at the selected location, using the smoothed feature importance map. Scores are normalized relative to the mean activation, allowing comparisons across features and models. We apply the same evaluation pipeline to the aligned models and compare their scores with the pretrained baselines from Colin et al. [8]. Details on participant recruitment and filtering are provided in App. D.

3 Results

Alignment increases similarity to humans and between models. Fine-tuning increases THINGS odd-one-out accuracy of the five backbones, from 44.4–50.7% to 60.1–60.7%. This improvement extends to between-class agreement on the held-out Levels benchmark, which rises from 47.6–59.7% to 66.3–70.7% (Fig. 1B). However, within-class and class-boundary agreement decrease for every backbone (App. B), indicating that alignment selectively improves coarse-grained semantic structure rather than all levels of human similarity. At the same time, the models become more similar to one another: cross-model RSA increases for all ten model pairs, with a mean change of $\Delta\rho = +0.14$ (Fig. 1A).

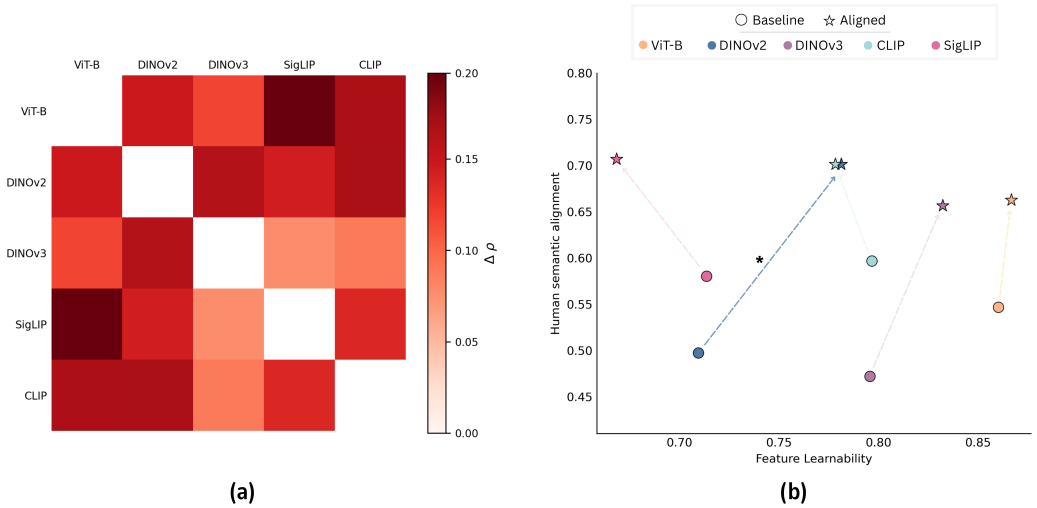

Figure 1: **Human-semantic alignment increases representational similarity but does not reliably improve feature learnability. (a)** Change in pairwise representational similarity (RSA, $\Delta\rho$) between the five vision models after alignment with human judgments. All ten model pairs become more similar. **(b)** Feature learnability *vs.* human semantic alignment (Levels [6]; between-class), before ($\circ$) and after ($\star$) fine-tuning. While alignment with human judgments consistently increases, feature learnability changes in different directions, with only DINOv2 showing a statistically significant improvement ($^*p = 0.009$).

Feature learnability does not reliably improve. Despite these consistent gains in representational similarity, feature learnability changes in different directions across models (Fig. 1B). Learnability scores increase for ViT-B ($+0.7$ points), DINOv2 ($+7.1$), and DINOv3 ($+3.6$), but decrease for CLIP (-1.9) and SigLIP (-4.5). Only DINOv2 shows a statistically significant improvement ($H(1) = 6.91$, $p = 0.009$). The sign of the change separates vision-only from vision–language backbones, but five backbones are too few to attribute this finding to pretraining. Learnability also spreads further apart, with its range across backbones growing from 15.0 to 19.8 points, even though the five backbones reach nearly identical agreement with human judgments (THINGS: 60.1-60.7%). The differences in learnability are thus unlikely to reflect backbones aligning to different degrees. Changes in learnability are also not simply explained by degraded downstream performance (App. F). Linear-probe ImageNet [16] accuracy decreases in the five backbones (by 0.6-3.4 points), so it cannot account for learnability changes in opposite directions, and across image categories the two changes are uncorrelated ($\rho = 0.014$, $n.s.$). In sum, increasing agreement with human similarity judgments, even when accompanied by greater similarity between models, does not reliably make their internal features easier for humans to learn.

What changes in the learned features? To understand these heterogeneous outcomes, we examine properties of the learned features (see App. E for details). To account for the pattern, a property must meet two conditions: alignment must change it in a way that tracks the changes in learnability across backbones, and it must itself be related to learnability. We test two candidate properties, locality and texture reliance, as well as an alternative mechanism, feature turnover.

Locality measures how spatially concentrated a feature's activation is within an image, computed as the Hoyer sparsity [17] of its attribution map. Concentrated features may be easier for observers to localize. Locality increases after alignment in exactly the three vision-only backbones (all $p \leq .001$), but not in CLIP ($p = .43$) or SigLIP ($p = .10$), matching the sign of the learnability change. It is also related to learnability: more local features tend to be easier to learn, both across all features ($\rho = 0.33$, $p < .001$, $n = 6695$) and within every backbone ($\rho = 0.17$–0.39). Locality thus meets both conditions, but leaves part of the pattern unexplained: SigLIP's learnability declines although its locality does not change significantly, and DINOv2 and DINOv3 gain locality similarly, yet only DINOv2's learnability improves significantly.

Texture reliance measures how much a feature depends on fine visual detail, estimated from how quickly its activation drops as images are blurred. Because the alignment that we perform favors coarse semantic structure, it might shift features away from such detail. Indeed, texture reliance decreases in the five backbones (all $p < .001$). But a change shared by every backbone cannot track learnability changes that go in opposite directions, and texture reliance is also not consistently related to learnability (pooled $\rho = 0.05$, with per-backbone associations of differing sign). It therefore fails both conditions.

Feature turnover. Alignment also reduces the number of informative features by 18–58% in every backbone. This suggests a different mechanism: learnability could improve if alignment drops hard-to-learn features and introduces easier ones. To test this hypothesis, we match features across the pretrained and aligned models using a dictionary (USAE [18]) trained jointly on both representations. We then compare pretrained features without an aligned counterpart (*dropped*) with aligned features without a pretrained counterpart (*introduced*). We find no evidence that introduced features are systematically easier to learn than dropped ones (all $p > .13$), offering no support for a simple feature-replacement explanation.

4 Discussion

Aligning models with human semantic judgments reliably made them more similar to humans and to one another, but did not reliably make their individual features easier for humans to learn. The association between semantic alignment and learnability observed across pretrained models [8] therefore does not reliably hold when alignment is induced directly. Similarly, the benefits of representational alignment for teaching [7] do not straightforwardly extend to a model's internal features. These findings suggest that representational alignment and feature learnability are related but different objectives.

Why does image-level alignment not reliably transfer to features? One possible explanation is a mismatch between the level of our intervention and the level at which learnability is evaluated. Our intervention reshapes how models relate whole images, whereas learnability depends on predicting individual feature activations. Although sparse autoencoders derive features from the structure of the representation space, reorganizing that space around human categories does not necessarily produce more learnable features. Indeed, models converged in their image-level judgments even while their feature learnability scores diverged. Alignment also improved coarse, between-class agreement while reducing within-class and class-boundary agreement, potentially weakening the finer distinctions that individual features capture. Consistent with evidence that similar image-level judgments can arise from different representational dimensions [10], our results suggest that image-level agreement does not guarantee shared or similarly learnable features.

When does alignment help? DINOv2 illustrates that alignment *can* improve feature learnability, while differences between vision-only and vision–language backbones suggest that outcomes depend on the starting representation. In our exploratory analyses, locality —the property most strongly associated with learnability— increased only in the vision-only backbones, whereas texture reliance decreased across all backbones without corresponding improvements in learnability. However, none of the examined properties explained changes in learnability. The representational conditions under which alignment improves feature learnability therefore remain unclear.

Scope and limitations. We evaluated a single intervention (LoRA fine-tuning) on five backbones, limiting our ability to identify why models responded differently. Our alignment target, odd-one-out judgments, improved between-class agreement at the expense of within-class agreement, and other supervision targets may yield different outcomes. Finally, our learnability measure captures whether humans can predict feature activations, not whether humans and models share the underlying concept.

In sum, making representations more human-aligned is not sufficient to make their features more learnable. Our findings highlight the need to evaluate these objectives separately and motivate future work on alignment methods that target not only how models organize images, but also the properties of their internal features that make them easier for humans to learn and predict.

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, Direccion General de Innovacion). This work has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No: 101120237 (ELIAS). 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

[1] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the International Conference on Machine Learning*, 2024. URL `https://proceedings.mlr.press/v235/huh24a.html`.

[2] Zirui Chen and Michael F Bonner. Universal dimensions of visual representation. *Science Advances*, 11(27):eadw7697, 2025.

[3] Florian P. Mahner, Johannes Roth, Ka Chun Lam, Michael F. Bonner, Francisco Pereira, and Martin N. Hebart. Characterizing universal object representations across vision models. *arXiv preprint arXiv:2605.13675*, 2026. URL `https://arxiv.org/abs/2605.13675`.

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

[5] 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.

[6] 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.

[7] Ilia Sucholutsky, Katherine M. Collins, Maya Malaviya, Nori Jacoby, Weiyang Liu, Theodore R. Sumers, et al. Representational alignment supports effective teaching. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 146–173, 2025. URL `https://proceedings.mlr.press/v273/sucholutsky25a.html`.

[8] Julien Colin, Lore Goetschalckx, Nuria Oliver, et al. Capability $\neq$ Interpretability: Human Interpretability of Vision Foundation Models. *arXiv preprint arXiv:2605.20337*, 2026.

[9] John Hewitt, Robert Geirhos, and Been Kim. Position: We can't understand ai using our existing vocabulary. In *Proceedings of the International Conference on Machine Learning*, 2025. URL `https://proceedings.mlr.press/v267/hewitt25a.html`.

[10] Florian P Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6):848–859, 2025.

[11] 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.

[12] 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.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[14] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

[15] 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.

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

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

[18] Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Forty-second International Conference on Machine Learning*, 2025.

[19] 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.

[20] 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.

[21] Justin D Lieber, Gerick M Lee, Najib J Majaj, and J Anthony Movshon. Sensitivity to naturalistic texture relies primarily on high spatial frequencies. *Journal of vision*, 23(2):4–4, 2023.

[22] Ajay Subramanian, Elena Sizikova, Najib J Majaj, and Denis G Pelli. Spatial-frequency channels, shape bias, and adversarial robustness. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[23] Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, volume 2025, pages 53617–53642, 2025.

[24] Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

[25] Piotr Jedryszek and Oliver M Crook. Stable and steerable sparse autoencoders with weight regularization. *arXiv preprint arXiv:2603.04198*, 2026.

A Alignment implementation

Each backbone is fine-tuned on the full THINGS triplet training set using LoRA [13]. For supervised ViT-B/16, DINOv2, DINOv3, and CLIP ViT-B/16, the training representation concatenates the CLS token with the mean of patch tokens, producing a 1536-dimensional vector. SigLIP has no CLS token in the encoder used here, so we use its 768-dimensional mean-pooled patch representation. A temporary linear projection head maps these representations to 1024 dimensions during optimization and is discarded afterward.

LoRA adapters use rank $r = 8$ and scaling $\alpha = 1.0$ in query, key, and value attention projections throughout the transformer blocks. All other backbone parameters are frozen. We optimize the contrastive triplet objective of Muttenthaler et al. [4], with temperature $\tau = 1.0$, Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$), no weight decay, and a peak learning rate of 10^{-4}. The learning rate warms up over 500 steps before cosine decay, and training proceeds for $20{,}000$ steps.

B Full human-alignment results

Table A.1: Alignment accuracy (%) before $\rightarrow$ after LoRA fine-tuning. Between-class agreement improves while within-class and class-boundary agreement declines.

Metric	ViT-B	DINOv2	DINOv3	CLIP	SigLIP
THINGS	$48.6 \rightarrow 60.1$	$45.1 \rightarrow 60.6$	$44.4 \rightarrow 60.5$	$50.7 \rightarrow 60.6$	$48.8 \rightarrow 60.7$
Levels between-class	$54.7 \rightarrow 66.3$	$49.7 \rightarrow 70.1$	$47.6 \rightarrow 69.7$	$59.7 \rightarrow 70.1$	$58.0 \rightarrow 70.7$
Levels within-class	$57.8 \rightarrow 53.4$	$65.2 \rightarrow 57.2$	$64.3 \rightarrow 56.6$	$62.1 \rightarrow 56.7$	$63.5 \rightarrow 57.9$
Levels class-boundary	$94.9 \rightarrow 86.6$	$93.0 \rightarrow 86.1$	$90.9 \rightarrow 86.6$	$91.0 \rightarrow 85.9$	$92.9 \rightarrow 86.9$

C Sparse features and feature learnability

Sparse features. For each backbone, independently trained TopK SAEs decompose baseline and aligned representations into sparse features (dictionary size 7680, expansion factor 10). Following Colin et al. [8], an informative feature is among the three features contributing most strongly to the model's top prediction for at least one image in the psychophysics stimulus set.

Feature learnability. We measure feature learnability with the localizability task. Each feature is explained to observers through a MACO synthetic visualization [19], nine highly activating natural images, and corresponding RISE importance maps [20].

Observers click where they expect the feature to be most active in an unseen query image. Each click is scored against the feature's Gaussian-smoothed activation map on that image. Let v be the map value at the click and $p = \widehat{F}(v)$ its percentile rank under the empirical CDF of the map's pixels. Raw percentiles are not comparable across features, because some features activate focally and others diffusely. We therefore anchor chance at the mean activation μ, the expected value under a uniformly random click, and rescale each side linearly. With $p_\mu = \widehat{F}(\mu)$,

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

Scores range from 0 to 1, and a click above the mean activation scores above chance ($s = 0.5$), however focal or diffuse the feature. Model-level learnability (Table A.2) averages these scores and is reported in $\%$.

Table A.2: Feature learnability (%) before and after. Chance-level is 50%. DINOv2 is the only backbone with a significant change.

	ViT-B	DINOv2	DINOv3	CLIP	SigLIP
Baseline	86.0	71.0	79.6	79.7	71.4
Aligned	86.7	78.1	83.2	77.8	66.9
Change	$+0.7$	$+7.1$	$+3.6$	-1.9	-4.5

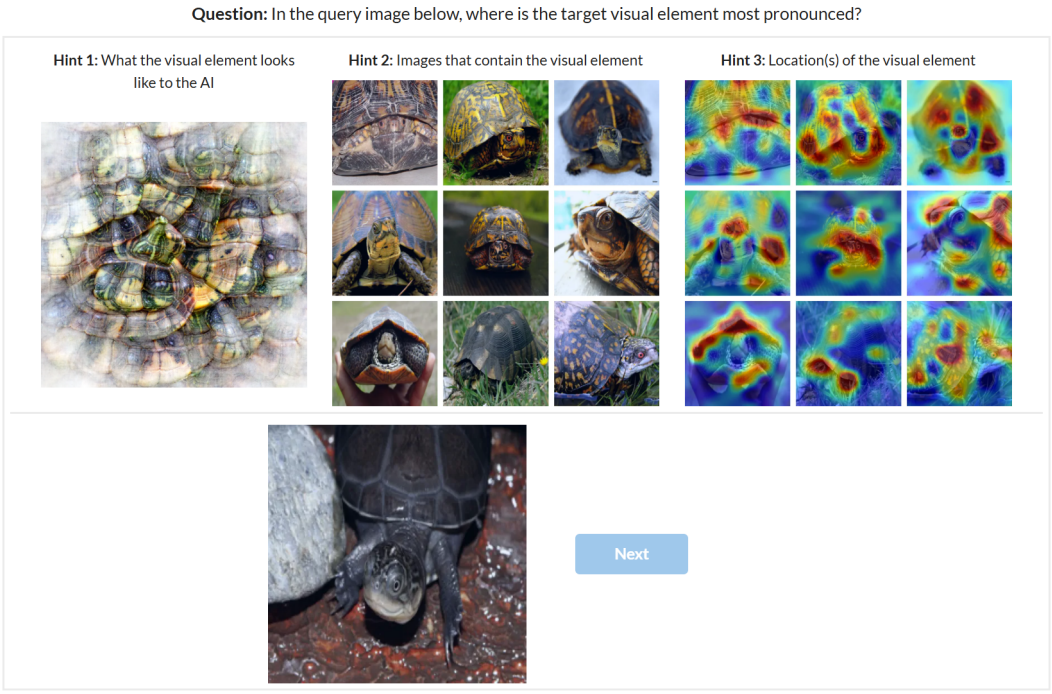

Figure A.1: Screenshot of a trial for the *localizability* experiment for the aligned DINOv2. A feature is explained through 3 panels at the top: (left) feature visualization by means of a maximally activating synthetic image, (middle) a set of images that highly activate the features, with their associated 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.

D Psychophysics experiment

Participants. To evaluate the aligned models, we recruited 212 participants via Prolific[2]. All were native English speakers without reported visual impairments and completed the study on a laptop 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 5,700 behavioral responses from the 147 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. We define one *behavioral response* as one participant's answer to one trial, i.e., a click in the localizability experiment.

We compare those responses to the 7,400 behavioral responses from the 190 participants collected by Colin et al. [8].

[2] www.prolific.com

E Exploratory analyses

These analyses characterize how alignment changes the informative features of each backbone (App. C). Associations with feature learnability are Spearman correlations computed over informative features.

Locality. For each informative feature, locality is the Hoyer sparsity of its RISE importance map, averaged over images on which the feature is active; higher values indicate more spatially concentrated features. Locality increases after alignment for ViT-B and DINOv2 ($p < .001$) and DINOv3 ($p = .001$), but not for CLIP ($p = .43$) or SigLIP ($p = .10$). Pooled across informative features, locality correlates with learnability ($\rho = 0.33$, $p < .001$, $n = 6695$), and the association is positive and significant within every backbone ($\rho = 0.17$–0.39, all $p < .001$). Two observations limit locality as an explanation. SigLIP's learnability declines without a significant change in locality. And DINOv3 shows a similar locality increase and an even stronger locality–learnability association than DINOv2 ($\rho = 0.36$ vs. 0.29), yet only DINOv2's learnability gain is significant. Figure A.2 illustrates a DINOv2 feature whose locality and learnability both increase after alignment.

Texture reliance. For each informative feature, we progressively blur its nine most activating images and track the peak activation retained relative to the unblurred image. The normalized area under this curve measures robustness to the removal of high-spatial-frequency content. Because sensitivity to high frequencies is closely tied to texture [21, 22], we treat lower robustness as a proxy for texture reliance, computed as one minus this area. Blur also removes fine contours, so this is not a direct measure of shape–texture bias. Texture reliance decreases after alignment in all five backbones (Kruskal–Wallis, all $H(1) > 37$, $p < .001$), most strongly for CLIP (-0.14). Yet texture reliance is essentially uncorrelated with learnability when pooling features ($\rho = 0.046$, $p < .001$, $n = 6695$), and per-backbone correlations range from -0.05 to $+0.20$ with no consistent sign.

Feature turnover. To identify features that alignment removes or introduces, we additionally train, for each backbone, a joint TopK SAE (USAE [18]) over the pretrained and aligned representations (dictionary size 6144), with a translation loss that encourages a shared basis. Each feature of the independently trained SAEs is assigned to its closest joint-dictionary feature by cosine similarity between decoder directions. The match is accepted when similarity exceeds 0.7, a common convention for deciding that two SAE features coincide [23–25], and the joint feature also activates in the other model. Pretrained features without a match are *dropped*; aligned features without a match are *introduced* (Table A.3). Alignment reduces the number of informative features in every backbone, and 24.5–71.4% of features lack a correspondent in the other model. Introduced features are not significantly more learnable than dropped ones in any backbone (Kruskal–Wallis, all $p > .13$), including SigLIP, which shows the largest learnability decline.

Table A.3: Informative features before and after alignment, and features without a cross-model correspondent. Percentages for dropped and introduced features are relative to the pretrained and aligned populations, respectively.

	ViT-B	DINOv2	DINOv3	CLIP	SigLIP
Informative (pretrained)	1259	1119	1018	551	896
Informative (aligned)	1029	667	426	417	561
Reduction	18%	40%	58%	24%	37%
Dropped	309 (24.5%)	365 (32.6%)	511 (50.2%)	270 (49.0%)	360 (40.2%)
Introduced	487 (47.3%)	253 (37.9%)	304 (71.4%)	230 (55.2%)	180 (32.1%)

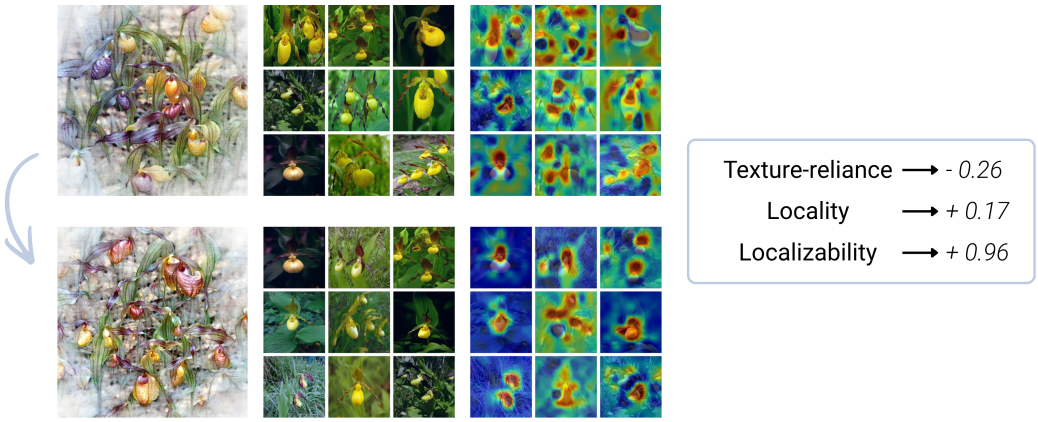

Figure A.2: **Illustration of a feature randomly sampled from the ones strongly affected by alignment.** A DINOv2 feature before (top) and after (bottom) alignment, shown through its synthetic visualization, top-activating images, and attribution maps. Its learnability (localizability) increases from 0.019 to 0.981, its locality from 0.063 to 0.229, and its texture reliance decreases by 0.264.

F Downstream-performance control

Linear probes on frozen backbone representations assess ImageNet-1k top-1 classification and ADE20K segmentation mIoU. Classification accuracy decreases in all five models by 0.6–3.4 points (mean -1.8); segmentation changes in both directions, including DINOv2 $+2.3$ mIoU and SigLIP -4.0 mIoU. Across image categories, the change in feature learnability is unrelated to the change in classification accuracy ($\rho = 0.014$, n.s.); within each backbone, associations are small ($|\rho| \leq 0.09$).