Evaluating Multiview Object Consistency in Humans and Image Models

Tyler Bonnen[1] **Stephanie Fu**[1] **Yutong Bai**[1] **Thomas O'Connell**[2] **Yoni Friedman**[2]
Nancy Kanwisher[2] **Joshua B. Tenenbaum**[2] **Alexei A. Efros**[1]

[1]University of California, Berkeley [2]Massachusetts Institute of Technology

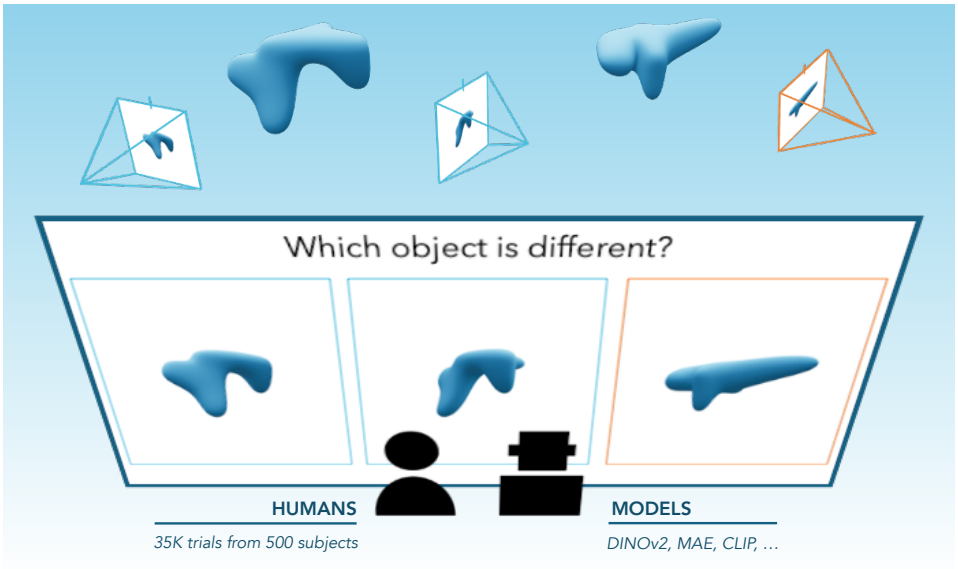

Figure 1: **How well do computer vision models represent the 3D structure of objects?** We develop a benchmark using a shape inference task from the cognitive sciences: Multiview Object Consistency in Humans and in Image models (MOCHI). Given three images of objects from random viewpoints, observers must identify which image depicts the object that is different. We compare human performance (35K trials from over 500 subjects, including accuracy, reaction time, and gaze data) against a number of standard computer vision models.

Abstract

We introduce a benchmark to directly evaluate the alignment between human observers and vision models on a 3D shape inference task. We leverage an experimental design from the cognitive sciences: given a set of images, participants identify which contain the same/different objects, despite considerable viewpoint variation. We draw from a diverse range of images that include common objects (e.g., chairs) as well as abstract shapes (i.e., procedurally generated 'nonsense' objects). After constructing over 2000 unique image sets, we administer these tasks to human participants, collecting 35K trials of behavioral data from over 500 participants. This includes explicit choice behaviors as well as intermediate measures, such as reaction time and gaze data. We then evaluate the performance of common vision models (e.g., DINOv2, MAE, CLIP). We find that humans outperform all models by a wide margin. Using a multi-scale evaluation approach, we identify underlying similarities and differences between models and humans: while human-model performance is correlated, humans allocate more time/processing on challenging trials. All images, data, and code can be accessed via our project page.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

https://doi.org/10.52202/079017-1379

1 Introduction

Do computer vision models represent the 3D structure of objects? Answering this question is more difficult than it might appear. Many tasks that seem to require explicit 3D representations can be performed directly from 2D visual features. Depth estimation, for example, is commonly used to evaluate a model's 3D 'awareness' (Banani et al., 2024), even though depth can be predicted using cues such as texture gradients (Malik and Rosenholtz, 1997), relative size (Cutting and Vishton, 1995), shading (Ramachandran, 1988), and camera blur (Mather, 1996). '3D understanding' may simply be a composite of such 2D features (e.g., the Aspect Graph model; Koenderink and Van Doorn, 1979), or it might require representations not directly accessible in these lower-level visual statistics (Marr, 1982). As large-scale vision models become more capable, these questions take on new implications. What methods might help us understand how vision models represent objects?

The cognitive sciences have grappled with questions about 3D perception for decades. To understand human visual abilities, classic work in this field (Shepard and Metzler, 1971) proposed a simple task: subjects were asked to determine whether two images contained the same or different objects, in spite of considerable viewpoint variation. Viewpoint variation was *intended* to prohibit low-level strategies (e.g., correspondence between 2D image features) and reveal veridical 3D shape understanding. Remarkably, the time needed to solve this task increases linearly with the angle of rotation between the two images, leading the authors to interpret these data as evidence for 'mental rotation.' However, there are many non-3D strategies that might be equally effective (e.g., Bülthoff and Edelman, 1992; Rock and DiVita, 1987; Tarr and Bülthoff, 1998; Ullman, 1998). To evaluate these competing claims, the field has developed increasingly refined tasks to probe human visual abilities (e.g., Fig. 1).

How do computer vision models compare to humans on 3D tasks from the cognitive sciences? While humans are able to dramatically outperform contemporary vision models (Bonnen et al., 2021; Bowers et al., 2022; O'Connell et al., 2023), there's a catch: when images are presented briefly (e.g., $<$100ms), humans achieve roughly the same accuracy on 3D shape inferences as vision models; it is only when humans are given more time that we outperform these models (Bonnen et al., 2023; Ollikka et al., 2024). Time is critical, in part, because it enables us to move our eyes around an image, sequentially attending to task-relevant features. Remarkably, when the neural structures responsible for integrating across visual sequences are damaged, human performance again resembles vision models (Bonnen et al., 2021). These data suggest that 3D shape perception in humans is a dynamic process with temporally and neurally distinct stages; contemporary vision models simply capture the feedforward components of this process (Jagadeesh and Gardner, 2022; Renninger and Malik, 2004).

Two aspects of these experiments are especially relevant for evaluating contemporary vision models. First, these tasks are agnostic as to how shape information is represented, focusing instead on visual abilities. This is in contrast to current computer vision methods which make strong assumptions about how 3D representations are formatted (e.g., depth, point clouds). Second, human-model alignment provides a better guarantee that model performance might relate to more general notions of '3D understanding'. That is, models that best predict human behaviors are those models most likely to exhibit similar choice behaviors in related tasks (Cao and Yamins, 2021). This requires going beyond determining if models achieve human-*level* performance (i.e., average accuracy) and evaluating whether models exhibit human-*like* performance (e.g., correlated choice behaviors).

In this paper, we construct a benchmark to evaluate the correspondence between humans and vision models using experimental tasks from the cognitive sciences. We integrate traditional experiments characterizing the object-level visual inferences of humans with recent benchmarking approaches (e.g., Rajalingham et al., 2018). The human behaviors in this dataset includes explicit measures, such as choice behavior, as well as intermediate measures, such as reaction time and gaze patterns. We develop an approach to evaluate the performance of several standard computer vision methods (e.g., DINOv2, MAE, CLIP), and then compare humans and models. Critically, instead of evaluating human-model alignment using a single measure, we use a series of increasingly granular metrics. Coarse-grained human-model comparisons (e.g., average performance across all trials) enables us to evaluate whether models achieve human-level shape inferences, while more fine-grained metrics (e.g., patterns of choice behaviors or attention) enable us to determine how human-like models are.

2 Experimental design, data collection, and model evaluation methods

Here we outline the task used to probe 3D vision, images in this benchmark, and human data collection. Finally, we summarize the evaluation methods we use to determine model performance.

https://doi.org/10.52202/079017-1379

2.1 Behavioral task

Our experimental design requires zero-shot visual inference about object shape: given three images, participants identify which contain the same/different objects. This design enables us to parametrically vary trial difficulty by changing the relative similarity between different objects (i.e., between object A and object B), and the viewpoint variation between images of the same object (i.e., between object A and the same object from a different viewpoint, A′). Moreover, this task imposes minimal verbal demands, enabling us to focus on perceptual processes and not language or semantic knowledge. We employ two variants of this task. In the 'odd-one-out' task (Bussey and Saksida, 2002) participants must identify the image within a triplet that contains an object that is different from the other two (i.e., select B given A, A′, and B). In the 'match-to-sample' task, participants are presented a 'sample' image (A), and then must select the 'match' image (A′) and not a lure (B).

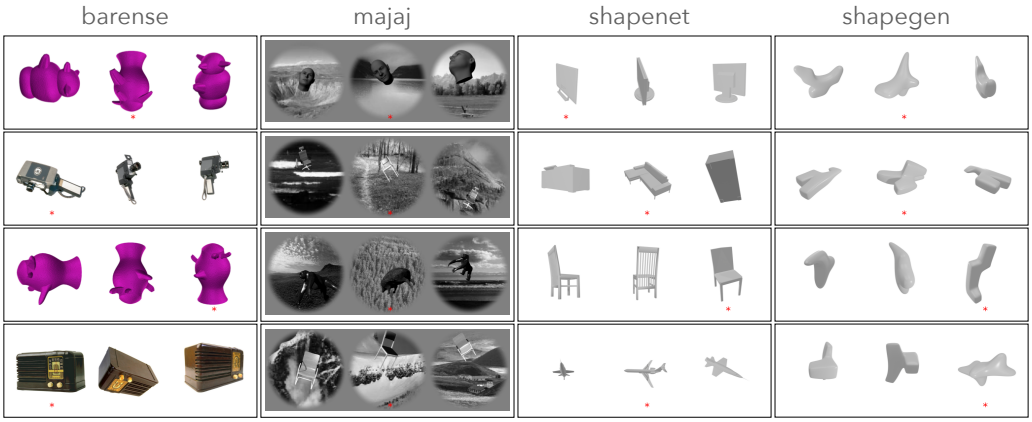

Figure 2: **Example stimuli from the four datasets in MOCHI.** Each trial is composed of a triplet of images containing two objects: one from two different viewpoints (A and A′), and another object (B). Depending on the experiment, participants either infer the matching/non-matching object (pairing A-A′, or identifying B). For illustrative purposes, B is marked in each trial above with a *. Descriptions and examples of all categories in this benchmark can be found in the appendix (25).

2.2 Stimuli

We integrate experimental stimuli from four datasets in Bonnen et al., 2021 and O'Connell et al., 2023. These images contain diverse object types and perceptual demands, which we organize into four distinct datasets. First, 'barense' contains color photographs of real-world objects (e.g., chairs) and abstract shapes (i.e., 'greebles' originally from Gauthier and Tarr, 1997) on a white background (Fig. 2 far left). These stimuli are separated into 'high similarity' and 'low similarity' conditions; 'high similarity' trials are thought to rely on understanding the 'compositional' 3D structure of objects while 'low similarity' trials are thought to rely on simpler visual features (e.g., color, texture). Images from 'majaj' contain four categories of objects (animals, chairs, planes, and faces) rendered in black and white and are superimposed onto randomized backgrounds (e.g., a chair floating in the sky above a mountain range; Fig. 2 middle left). These stimuli are designed to make object segmentation and representation challenging, given the object-irrelevant distractors. We describe these as 'barense' and 'majaj' according to the original source of these images (Barense et al., 2007; Majaj et al., 2015). Images from 'shapenet' contain everyday objects from multiple classes (e.g. cabinet, car, lamp) rendered without surface textures from random views sampled from a sphere (Fig. 2 middle right). The two objects selected for a given trial were always drawn from the same class. Images from 'shapegen' contain procedurally-generated objects using the ShapeGenerator extension for Blender (Fig. 2 far right). The result is a dataset which allows us to procedurally target a range of human behaviors when making zero-shot object-level inferences. None of these stimuli contain any personal identifying information or offensive content. We license all assets under CC BY-NC-SA 4.0.

2.3 Human data collection

After constructing over 2000 unique image triplets, we present these tasks to human participants, collecting 35K trials of behavioral data from over 500 participants via online and in-lab studies.

https://doi.org/10.52202/079017-1379

Experimental data were collected online using Prolific. Participants were each paid $15/hr for participating, and were free to terminate the experiment at any time. Experiments were approved by the MIT Committee on the Use of Humans as Experimental Subjects (Protocol #1011004131). Experiments consisted of an initial set of instructions, 6 practice trials with feedback, and 150 main trials with no feedback. The 150 main trials were constructed such that no objects were repeated across trials, to avoid learning effects, and the ordering of the correct choice was randomized, to control for ordering effects. Given the performance we estimate for each image triplet, we normalize human accuracy to lie between zero (chance) and 1 (ceiling). This enables us to compare odd-one-out tasks and match-to-sample tasks in the same metric space. We collect eyetracking data on a subset of images and outline the data collection and analysis procedures for these data in the appendix.

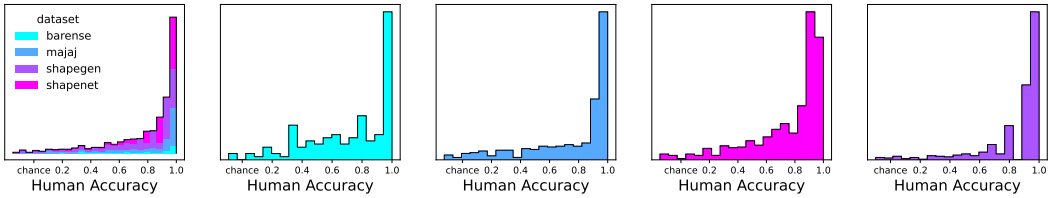

Figure 3: **Distribution of human accuracy across trials in each dataset.** For each of the four datasets, the average performance of each trial is plotting along the x axis, ranging from chance (left) to ceiling (right). While humans are reliably accurate on trials in each each dataset, there is a long-tailed distribution; this is by design, as it provides more challenging behavioral targets to model.

2.4 Establishing a suitable model evaluation method for this benchmark

We evaluate the performance of vision models optimized via contrastive (DINOv2) and autoencoding (MAE) self-supervision objectives. Furthermore, we analyse the performance of vision-language models trained via a contrastive image-text objective (CLIP) and with visual instruction tuning (LLaVA). Given this representative subset of available models, we evaluate multiple instance of each model class (e.g., DINOv2-Base, DINOv2-Large, and DINOv2-Giant) in order to determine how model scale relates to performance on this benchmark as well as human-model alignment. Given an image triplet, composed of two images of object A from different viewpoints (i.e., A and A$'$) and a third image of object B, we extract model responses independently for each image. For all DINOv2 models we use pooled features after the final hidden layer. For all CLIP models we extract features from the final layer of the vision encoder. For MAEs we extract features from the [CLS] token. To estimate model performance we use several evaluation metrics on each trial, given the feature vectors in response to A, A$'$, and B. We model all trials as odd-one-out tasks and determine model performance as the accuracy the model achieves in identifying the non-matching object (B).

2.4.1 Distance metrics

We define an analytic approach that generates an estimate of the non-matching object in each trial: given model responses to A, A$'$, and B, we compute the pairwise similarity between items, then determine the 'oddity' to be the item with the lowest off-diagonal similarity to the other images. We estimate the similarity between items using multiple distance metrics (cosine, euclidean, etc.). A one-to-one comparison with human performance requires that our we have a continuous estimate of model performance on each trial (i.e., not a single 1 or 0). To achieve this, we determine model performance for each triplet in the following way: for each iteration, we apply random in-plane rotations to the original images, estimate model performance, and average performance on this image across all iterations. The in-plane rotations we apply to each images are drawn from the same distribution of rotations used to generate the stimuli presented to humans. Thus, for each triplet we have an estimate of each models' averaged zero-shot performance, as well as estimates of the variance in this trial (e.g., standard error of the mean, SEM, and standard deviation, STD).

2.4.2 Linear probes

We design two independent linear probes to evaluate model performance on each triplet. We first design a standard linear probing strategy following the default setting in Caron et al., 2021. A trainable MLP layer is applied to a frozen visual backbone. We formulate this task as a multi-way

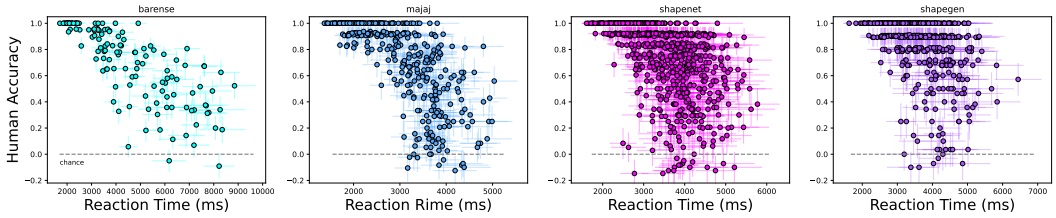

Figure 4: **Accuracy and reaction time distributions for human participants across datasets.** Across datasets we observe a clear relationship between accuracy (y axis) and processing time (x axis); as trials become more difficult, participants allocate more attention/time. Critically, the distribution of human behavior ranges from chance to ceiling, indicating that we have a suitable estimate of the full range of human visual abilities. This and all subsequent error bars are SEM computed over trials.

classification problem. Additionally, we design a lightweight linear probe hand-tailored for our experimental procedure: for a given triplet, T, we train a linear classifier to perform a 'same-different' classification on similar (not T) triplets in each dataset, and evaluate on triplet T. Concretely, in each condition (e.g., 'planes' in the Majaj dataset), for each triplet, we compute pairwise difference vectors between all images (i.e., $A - A' = \Delta_{A-A'}$) and label this vector appropriately (i.e., label($A - A'$) = 1, label(A - B) = 0). This results in a series of difference vectors and their corresponding labels. Critically, this enables us to learn a same-different decision boundary that generalizes to different decisions within this same condition. To this end, for each triplet T, we train an SVM to perform this classification using a subset of triplets from the same condition (e.g., 75% of 'planes' discriminations if T is in the 'planes' condition) and evaluate on trial T. We repeat this procedure 100 times for each trial, resulting in a continuous measure of model performance include SEM and STD.

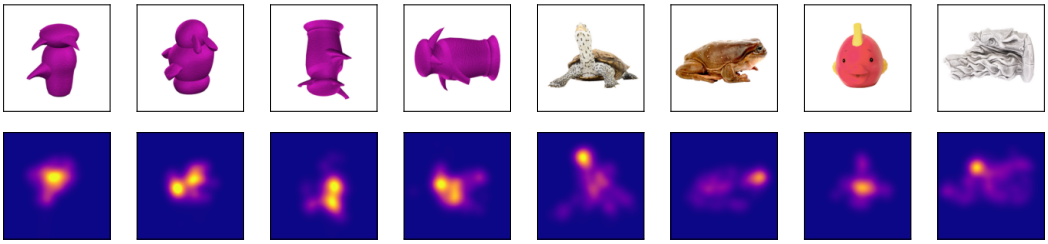

Figure 5: **Human gaze patterns measured via an in-lab eye tracking study.** We measure human eye movements during performance on one dataset in MOCHI (barense). This enables us to identify what features humans attend to, and whether this pattern of attention is reliable across participants. For illustrative purposes, here we visualize the saliency maps (bottom) for several images (top).

2.4.3 A lightweight linear probe exhibits the most promising performance on this benchmark

Distance metrics and linear probes each have desirable properties; distance metrics provide a 'zero-shot' estimate of model representations, while linear probes offer a more expressive readout tailored to a specific task. Both have undesirable properties as well; distance metrics typically require task-relevant information to be uniformly represented in model features, while linear probes require task-relevant training data and design choices (e.g., defining a suitable train-test split). Here we simply choose the evaluation approach that performs best. Our same-different classifier has a clear advantage over standard linear probes and distance metrics. Training a standard linear probe is unstable, leading to catastrophic performance, perhaps due to the relatively small dataset size. Distance metrics exhibit much better performance without any specific fine-tuning. However, the same-different linear classifier performs significantly better than both the average performance of the distance metrics (paired t-test $t(2018) = 12.83, P = 2.74\text{x}10^{-36}$) as well as all individual metrics (e.g., $l2$, paired t-test $t(2018) = 12.40, P = 4.3\text{x}10^{-34}$). We visualize this comparison across datasets in Fig. 14. For subsequent analyses we report the results of model performance using the lightweight same-different classifier. As a final step, we normalize the accuracy to be between 0 (chance) and 1 (ceiling).

https://doi.org/10.52202/079017-1379

3 Results

3.1 Human accuracy, reaction time, and gaze behaviors are reliable across participants.

Humans behaviors are both accurate and reliable, exhibiting considerable variation across conditions in this benchmark. Average human accuracy is 78%, with performance ranging between 40% and 100% for different conditions. To determine the noise ceiling for these behaviors—i.e., how much variance we might hope to explain with vision models—we compute the split-half reliability of these behaviors over 1000 permutations and compare this distribution to an empirically estimated null (i.e., generated by shuffling the order of the split-half correlation). We find that human performance is reliable at the level of averaged performance across datasets (r_{median}=.93, paired t-test $t(999)$ = $49.1, P = 2.24\text{x}10^{-268}$, Fig. 6), averaged performance across conditions (r_{median}=.95, paired t-test $t(999)$ = $48.993, P = 6.7\text{x}10^{-268}$), and averaged performance across trials (r_{median}=.93, paired t-test $t(999)$ = $958.17, P = 0$). Beyond performance itself, we observe a significant relationship between accuracy and reaction time across all trials in this dataset; as accuracy decreases, reaction time increases (r=-.52, $F(1, 2017) = -27.44, P = 4$ x 10^{-141}). This suggests that more difficult trials require more processing time for relatively accurate performance. Reaction time is also reliable across datasets (r_{median}=.99, paired t-test $t(999)$ = $52.68, P = 1.27$ x 10^{-290}), across conditions (r_{median}=.98, paired t-test $t(999)$ = $152.9, P = 0$) and across trials (r_{median}=.69, paired t-test $t(999)$ = $903.4, P = 0$). Finally, we turn our attention to the most fine-grained behavioral measure we collected on a subset of the experimental data (all stimuli in barense). We estimate the split-half reliability of human gaze patterns by determining how correlated salience maps are across participants who viewed the same image. These measures are reliable across participants (r_{median}=.94, paired t-test $t(83998)$ = $430.9, P = 0$, Fig. 12). That is, for a given image, different people attend to similar object locations/features. Taken together, these results suggest that while humans are highly accurate, their performance exhibits considerable variance. Critically, these behavioral data are reliable across participants, suggesting that there is variance to explain at multiple levels of granularity.

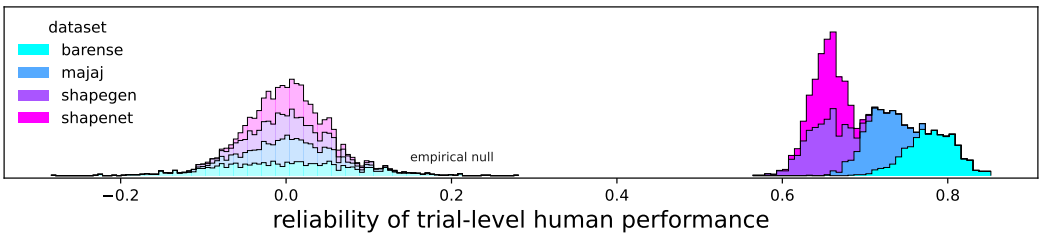

Figure 6: **Distribution of split-half reliability scores for human performance.** In order to estimate the reliability of human accuracy on this benchmark, we first estimate the average performance of two random halves of participants on each trial. After estimating performance for all trials, we determine the correlation between these two splits. We repeat this process for 1000 iterations, resulting in a distribution of split-half reliability scores (right), contrasted with the empirical null (left). These data suggest that this benchmark offers a reliable estimate of human object-level shape inference.

3.2 Humans outperform vision models by a wide margin, regardless of model scale

Humans outperform models on this benchmark by a wide margin. The best performing model (DINOv2-G) achieves 44% accuracy while humans achieve 78%. On average, DINOv2 models achieve an average accuracy of 0.43%, while CLIPs have an average accuracy of 0.29%, and MAEs have an average accuracy of -0.03%. DINOv2 outperforms all other models on this benchmark (Fig. 7, all DINOv2s vs all CLIPs, paired t-test $t(16150)$ = $14.31, P = 3.17$ x 10^{-46}, vs all MAEs, paired t-test $t(12112)$ = $47.23, P = 0$). Conversely, MAE's performance hovers around chance for all model sizes ($t(6056)$ = $.11, P = 0.91$). Increased model size leads to improved performance on this benchmark for both CLIP and DINOv2, while the performance of MAEs does not improve. Insofar as object-level shape inferences are concerned, not all self-supervision objectives benefit from increased model scale. Moreover, a thirteen-fold increase in the number of model parameters from DINOv2-Base to DINOv2-Giant only leads to an increase from 32% to 44%.

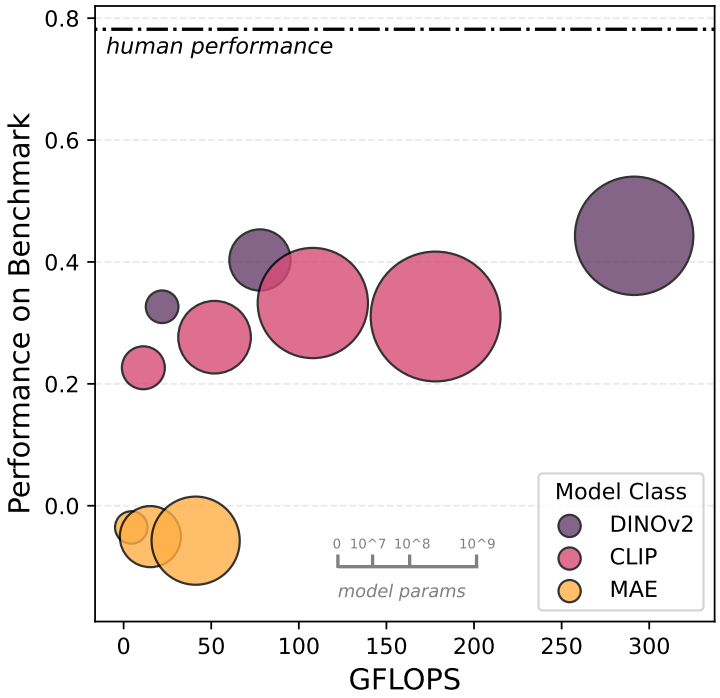

Figure 7: **Relationship between performance on 3D shape inferences and model size/FLOPS.** For each model, we determine how performance (y axis) relates to the number of floating-point operations per second (FLOPS, x axis) as well as model size (visualized as point size). We find that model performance improves with increased scale for some model types (e.g., DINOv2, purple) and not others (MAE, yellow). Nonetheless, humans (dashed line) outperform all models by a wide margin.

	Human	DINOv2-G	DINOv2-B	DINOv2-L	CLIP-B	CLIP-L	CLIP-H	MAE-B	MAE-L	MAE-H
Accuracy	**0.78**	0.44	0.33	0.40	0.23	0.28	0.33	-0.04	-0.05	-0.06

Table 1: Numerical values for performance of humans and vision models on our MOCHI benchmark. Absolute accuracy is normalized in relation to chance (zero) and ceiling (one).

3.3 While humans outperform models, their performance is correlated.

When we move beyond averaged accuracy of humans and models, we find that human and model performance is correlated. We begin by noting that the gap between human and model performance varies considerably across datasets (average difference 38% $\pm$ 30% STD). We first report some qualitative patterns observed across different conditions, in relation to their stimulus properties. The best performing model we evaluated (DINOv2) approaches human-level performance on a relatively diverse subset of experimental conditions: in the 'animals' and 'planes' conditions in the majaj dataset, which have greyscaled objects superimposed onto random backgrounds (Fig. 8, second row), 'abstract4' in the shapegen dataset, which has abstract objects removed of color- or texture-level visual identifiers (Fig. 8, fourth row, left), and in 'familiar_lowsim' and 'novel_lowsim', two conditions in barense (Fig. 8, top left). Nonetheless, there many more conditions where model performance drops to chance as human performance is relatively intact (e.g., 'familiar_hisim' in barense, Fig. 8, top row, and 'abstract0' in shapegen, bottom row). Returning to quantitative measures, we find that while there is a considerable performance gap between humans and DINOv2, their behaviors are nonetheless correlated. We observe this human-model correlation across datasets ($r = .42$, not reporting statistics because of low sample size, n=4), conditions ($r = .58$, $F(1, 23) = 3.40, P = 0.002$), and trials ($r = .35$ $F(1, 2017) = 16.86, P = 9$ x 10^{-60}). When comparing to the reliability ceiling estimated for human behaviors across these different resolutions, DINOv2 can predict about 0.61% and 0.37%

https://doi.org/10.52202/079017-1379

of the explainable variance across conditions, and trials, respectively (Table 2) . We visualize this relation in Fig. 9, binning across trials. Taken together, these data suggest that task difficulty is shared across humans and models, despite considerable differences in accuracy.

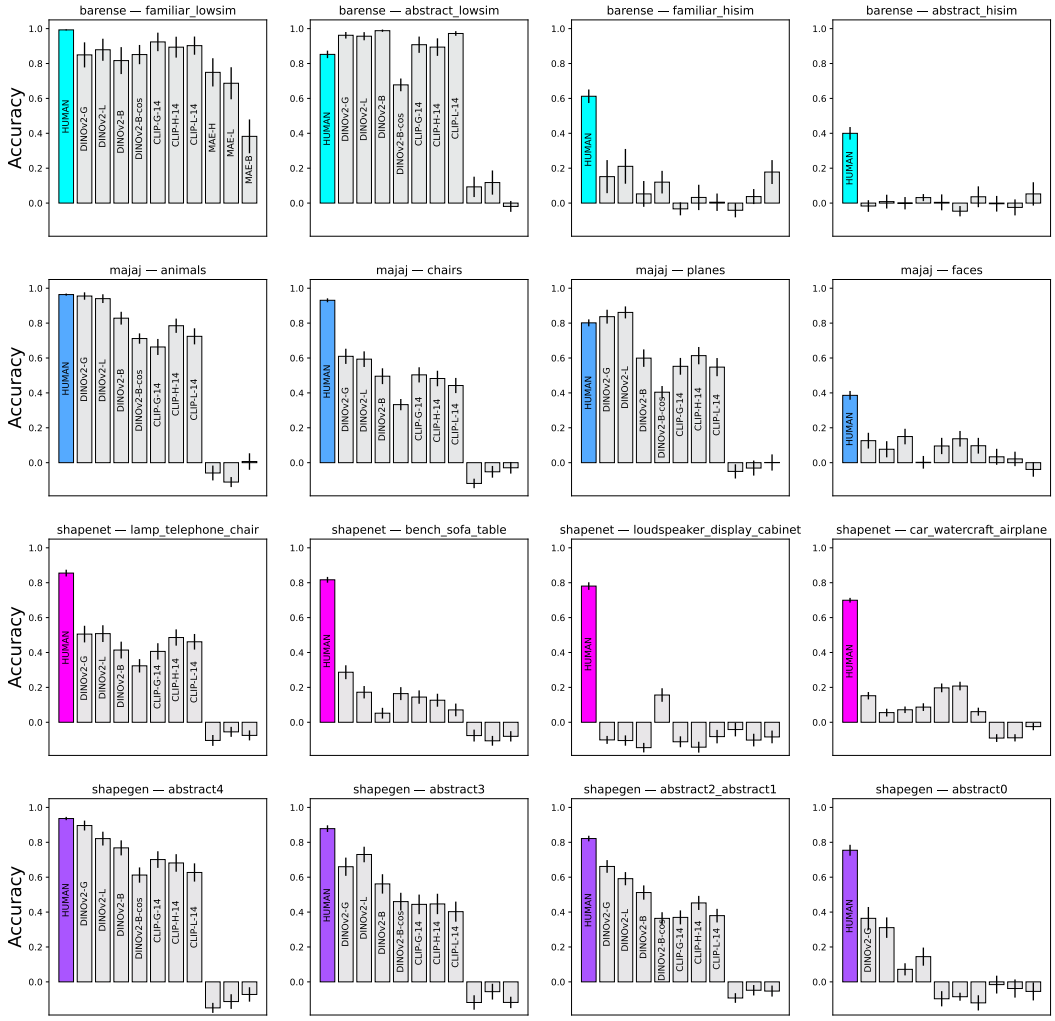

Figure 8: **Average performance of of humans and multiple vision models across conditions.** For each condition in this dataset, we compare human performance (left, color) to DINOv2s, CLIPs, and MAEs across multiple scales (greys). While humans outperform models by a wide margin, there is considerable variance in the gap between humans and models across conditions.

Level	DINOv2-B	DINOv2-L	DINOv2-G	CLIP-B	CLIP-L	CLIP-H	MAE-B	MAE-L	MAE-H
Condition	0.52%	**0.60%**	**0.60%**	0.56%	0.54%	0.53%	-0.09%	0.12%	0.08%
Trial	0.19%	0.18%	**0.28%**	0.18%	0.14%	0.22%	0.02%	0.01%	0.06%

Table 2: **Alignment between human and model performance across levels of granularity.** We determine the correlation between humans and vision models across coarse-grained (conditions, top) to fine-grained (trials, bottom) measures. We scale this correlation by the median split-half reliability of human behavior at each resolution, such that a correlation of 1 represents model-human correlation that is equal to human-human correlation. These results suggest that DINOv2 largely exhibits the coarse patterns in human behavior (60%) but not necessarily fine-grained choice behaviors (28%).

https://doi.org/10.52202/079017-1379

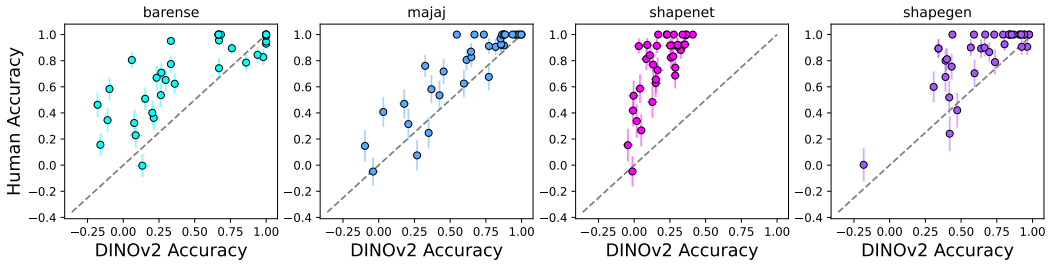

Figure 9: **Human and model performance is correlated.** While humans outperform vision models on MOCHI by a wide margin, there is still a correlation between their performance. We visualize this relationship by binning across trials in each dataset, then and plotting the performance of the best performing models (DINOv2-G) along the x-axis and human performance along the y-axis.

3.4 Human-model divergence may be explained by increased processing time and attention

To gain further insights into the superior performance of humans compared to vision models, we turn to additional measures of human behavior: reaction time and gaze dynamics (i.e., attention). Reaction time is commonly used as a measure for the amount of processing needed for a given behavior, and gaze behaviors are a direct measure of the visual features that humans attend to when viewing an image. These behavioral intermediates provide clues about the algorithmic basis of human visual abilities. First, we directly compare the relationship between human reaction time and model performance. We find that the trials for which model performance is degraded are the same trials where humans allocate more time (i.e., a significant correlation between human reaction time and model performance, r=-0.29, $F(1, 2017) = -13.72, P = 5 \text{ x } 10^{-41}$, Fig. 9). Given that human performance on these trials is significantly greater than model performance (i.e., significant above-diagonal variance observed in Fig. 9), it appears that humans allocate more processing time for those trials that are more challenging. If this were true, we might expect this processing time to be evident in some way in their attention patterns. Specifically, it has been hypothesized that human viewing behaviors (i.e., sequentially sampling visual images by moving our gaze to different visual features) enables us to 'compose' flexible representations of objects using a finite set of visual representations (Bonnen et al., 2023; Ullman, 1987). As we outlined in the human behavioral results, these gaze dynamics are highly reliable across participants; when two people encode a visual stimulus, they tend to look at the same features. Are vision models attending to the same feature that humans are, during these extending viewing periods? We extract attention measures from intermediate model layers, across all models layers, and we find that DINOv2 features do not predict human visual attention any better than would be expected from chance (human-human reliability scores visualized in Fig. 12, left, alongside model-human attention scores; random subset of attention maps shown in Fig. 12 right). The patterns observed in human attention are not evident in model attention.

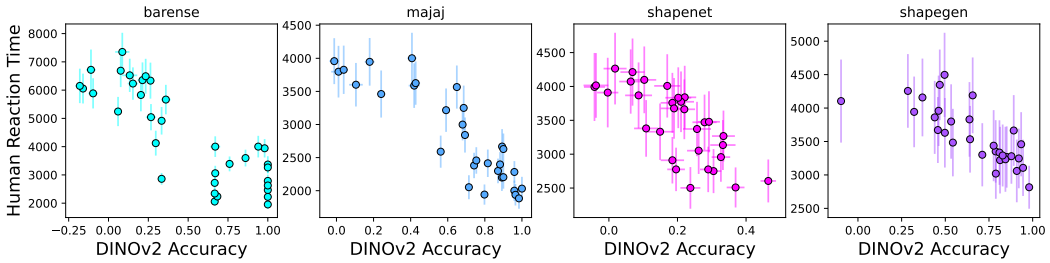

Figure 10: **Unlike models, humans dynamically process images in response to task difficulty.** How do humans maintain our advantage over vision models? Here we visualize the relationship between model performance (DINOv2-G, x-axis) and human reaction time (y-axis). Across all four datasets, as trials become more difficulty, model performance degrades, humans reaction time increases. That is, while vision models allocate the same amount of processing for each image, humans allocate processing time adaptively, in a manner that scales with the difficulty of each trial.

https://doi.org/10.52202/079017-1379

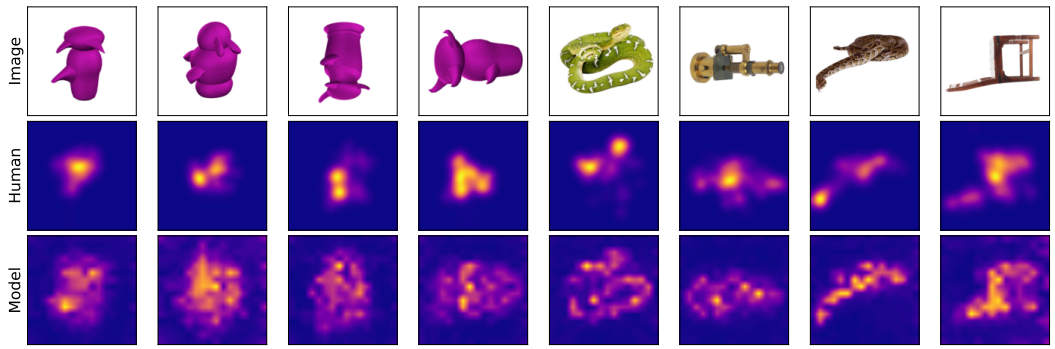

Figure 11: **Humans attend to images in a manner not captured by attention in vision models.** When presented within an image (top row), we find that humans attend to a relatively focal task-relevant features, as is evident in the saliency maps derived from their eye movements (middle row). In contrast, model attention (e.g., extracted from DINOv2) is distributed more uniformly across each object. These focal attention patterns may help explain how humans outperform vision models.

4 Limitations

While the objects in this benchmark are relatively diverse (e.g., semantically familiar objects like chairs, as well as abstract shapes) we have only a small sample of possible shapes. We leave it to future work to incorporate more diverse 3D objects, including a more thorough exploration of properties such as semantics, texture, and spatial frequency. Similarly, our current evaluation metrics are only a starting point for understanding the visual abilities of large vision models. For example, it is possible that DINOv2 performs best under the current evaluation scheme because these methods well aligned with DINOv2's contrastive pretraining objective. As such, we might expect to see an improvement in the performance of other models (e.g., MAEs) when they are evaluated with methods better suited to each model class.

5 Conclusion

We introduce a object-level shape inference benchmark that enables us to formally evaluate vision models alongside human accuracy, reaction time, and gaze data. Humans outperform all vision models by a wide margin on this benchmark. Nonetheless, human and model performance is correlated, indicating that similar trials are difficult for both systems. What accounts for this human ability to outperform models? Unlike models, humans process each trial in a way that depends on task difficulty. For example, those trials where models perform poorly are the same trials where human reaction time is longer (i.e., more processing allocated). Similarly, there is reliable variance in the human attention dynamics, as measured via eyetracking, which is unexplained by model attention. These data corroborate decades of research in the neuroscience and cognitive science of human vision, indicating that there are complementary neural systems and cognitive processes that support object-level shape inferences. We hope this benchmark will serve as an independent validation set, showcasing how evaluation approaches from the cognitive sciences can help evaluate increasingly capable vision models, and point towards novel modeling strategies to resolve their limitations.

6 Data and code availability

We license all assets (data, code, and images) under CC BY-NC-SA 4.0. These resources can be found at our project page. Images are available on huggingface and all code is available on github.

7 Acknowledgements

We thank Devin Guillory for suggestions about initiating this project and Sophia Koepke for feedback on this manuscript. This work is supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health (Award Number F99NS125816) as well as the National Science Foundation (Grant 2124136) and ONR MURI N00014-21-1-2801.

References

Banani, M. E., Raj, A., Maninis, K.-K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., & Jampani, V. (2024). Probing the 3d awareness of visual foundation models. *arXiv preprint arXiv:2404.08636*.

Barense, M. D., Gaffan, D., & Graham, K. S. (2007). The human medial temporal lobe processes online representations of complex objects. *Neuropsychologia*, *45*(13), 2963–2974.

Bonnen, T., Wagner, A. D., & Yamins, D. L. (2023). Medial temporal cortex supports compositional visual inferences. *bioRxiv*, 2023–09.

Bonnen, T., Yamins, D. L., & Wagner, A. D. (2021). When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*, *109*(17), 2755–2766.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., et al. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 1–74.

Buckley, M. J., Booth, M. C., Rolls, E. T., & Gaffan, D. (2001). Selective perceptual impairments after perirhinal cortex ablation. *Journal of Neuroscience*, *21*(24), 9824–9836.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, *89*(1), 60–64.

Bussey, T. J., & Saksida, L. M. (2002). The organization of visual object representations: A connectionist model of effects of lesions in perirhinal cortex. *European Journal of Neuroscience*, *15*(2), 355–364.

Bussey, T. J., Saksida, L. M., & Murray, E. A. (2002). Perirhinal cortex resolves feature ambiguity in complex visual discriminations. *European Journal of Neuroscience*, *15*(2), 365–374.

Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 1–taking mechanistic abstraction seriously. *arXiv preprint arXiv:2104.01490*.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Cutting, J. E., & Vishton, P. M. (1995). Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion* (pp. 69–117). Elsevier.

De Leeuw, J. R. (2015). Jspsych: A javascript library for creating behavioral experiments in a web browser. *Behav. Res. Methods.*, *47*(1), 1–12.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision research*, *37*(12), 1673–1682.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259.

Jagadeesh, A. V., & Gardner, J. (2022). Texture-like representation of objects in human visual cortex. *bioRxiv*.

Koenderink, J. J., & Van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biological cybernetics*, *32*(4), 211–216.

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, *35*(39), 13402–13418.

Malik, J., & Rosenholtz, R. (1997). Computing local surface orientation and shape from texture for curved surfaces. *International journal of computer vision*, *23*(2), 149–168.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

Mather, G. (1996). Image blur as a pictorial depth cue. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *263*(1367), 169–172.

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior research methods*, *42*(1), 188–204.

O'Connell, T. P., Bonnen, T., Friedman, Y., Tewari, A., Tenenbaum, J. B., Sitzmann, V., & Kanwisher, N. (2023). Approaching human 3d shape perception with neurally mappable models.

Ollikka, N., Abbas, A., Perin, A., Kilpeläinen, M., & Deny, S. (2024). Humans beat deep networks at recognizing objects in unusual poses, given enough time. *arXiv preprint arXiv:2402.03973*.

 https://doi.org/10.52202/079017-1379

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Ramachandran, V. S. (1988). Perception of shape from shading. *Nature*, *331*(6152), 163–166.

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision research*, *44*(19), 2301–2311.

Rock, I., & DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive psychology*, *19*(2), 280–293.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. *Cognition*, *67*(1-2), 1–20.

Ullman, S. (1987). Visual routines. In *Readings in computer vision* (pp. 298–328). Elsevier.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition*, *67*(1-2), 21–44.

A Appendix / supplemental material

A.1 Online human data collection

Human experimental data were collected online via Amazon Mechanical Turk and Prolific via experiments were implemented in JsPsych (De Leeuw, 2015). Each experiment began with an instruction phase, which introduced them to the task as well as provided 5 practice trials. This provided an opportunity for participants to acclimate themselves to the task and the controls. Once the experiment began, participants initiated the beginning of each trial with a button press (spacebar), such that they can (effectively) pause the experiment whenever they deem appropriate. This was designed to reduce environmental interference in the experiment. Experiments were designed to be completed in 10 minutes and participants were payed at a rate of roughly $16/hour. In addition, participants were awarded a bonus commensurate with their performance, enabling them to earn up to twice the base pay. In order to ensure that participants were fairly compensated for their time, even in the case of a crowdsourcing platform errors, trial-by-trial data were collected throughout the experiment and stored on a custom server built from a Digital Ocean 'droplet.'

We administer two related experimental designs. First, we use a 3-way concurrent visual discrimination task commonly used to evaluate the role of MTC in perception (Barense et al., 2007; Buckley et al., 2001; Bussey et al., 2002). This design enables us to determine visual inferences that are possible with unlimited viewing time, as all stimuli remain on the screen for the duration of the trial. On each trial, participants are presented with three images and must identify the image that does not match the other two in terms of object identity (i.e., the 'oddity'). Participants are given upwards of ten seconds to complete each trial. At any point in this duration, participants can select the oddity with a button press (right arrow, left arrow, or down arrow) corresponding to those locations on the oddity array. After this button press, participants are given feedback related to their performance on that trial, indicating whether their choice was correct or incorrect. If participants do no press a button in these ten seconds, the trial is marked as incorrect, feedback is given on the screen encouraging them to complete each trial within the allotted time.

A.2 Eye tracking data collection

Eye tracking was performed using an infrared video-based eye-tracker at 1000 Hz (Eyelink 1000; SR Research). Stimuli were displayed on a 22.5 inch VIEWPixx LCD display (resolution of 1900×1200, refresh rate of 120 Hz) and responses collected via keyboard. Other sources of light were minimized during data collection. The stimulus on the sample screen was presented at the central field of view and spanned up to 10 degrees of visual angle. This stimulus size was selected such that in order to collect high-acuity visual information from various stimulus locations, participants had to move their eyes (i.e., make a saccade). Stimuli on the match screen were the same size, but presented side by side, offset from the horizontal midpoint of the screen by 10 degrees of visual angle. Each experiment began with gaze calibration, then 5 practice trials to acclimate participants to the experimental setup. Each trial was initiated by the participants and began with participants maintaining fixation at the center of the screen (to perform drift correction at the beginning of each trial). Participants completed each trial at their own pace and there was a brief rest period every 5 minutes. This duration of this rest period was at the discretion of each participant. After this rest period, there was another gaze calibration, after which participants again completed a series of trials at their own pace as described above. For all gaze analyses (e.g., evaluating gaze reliability) we estimate gaze-related events (e.g. fixations) directly from the raw gaze data using a standard python library (REMoDNaV; Nyström and Holmqvist, 2010).

A.3 Estimating gaze reliability

We estimate the split-half reliability of in-lab gaze dynamics using the following protocol. First, for each trial, a subject-level salience map is generated from the raw gaze behaviors: a 2D histogram is generated from the raw time series, which is then smoothed with a Gaussian kernel. We note that the results reported in this manuscript are robust to the resolution of the 2D histogram and size of the smoothing kernel. This protocol yields a salience map for each image for each subject. We then generate a random split of subjects and partition the salience maps for a given image using this random subject split. We then average across participants in each random split, which results in two salience maps, each corresponding to the random split of participants allotted to that half. We

https://doi.org/10.52202/079017-1379

then estimate the correlation between the two (random split-half) salience maps associated with this image. We repeat this protocol for 100 random split-half permutations (i.e., generating a new shuffle of participants each iteration). For each image, we then have a distribution of split-half correlations which enables us to evaluate how similar participants viewed each image. To establish an empirical null we compute the correlation between random splits corresponding the different images within the same trial. Additionally, we estimate the bottom-up salience of each image (Itti et al., 1998) and compute the correlation between this bottom-up salience map and the random splits associated with each permutation of each image.

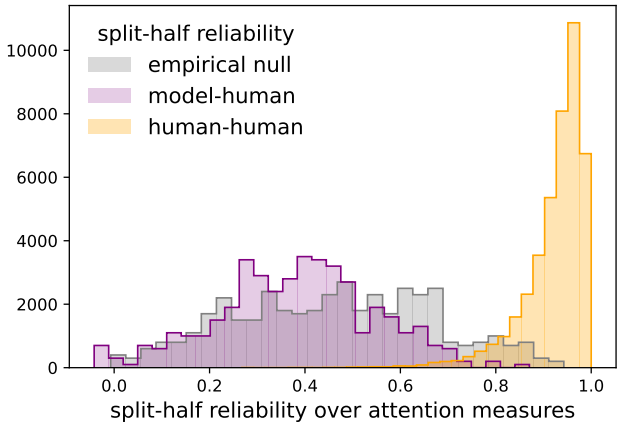

Figure 12: **Human gaze patterns are reliable but not predicted by model self-attention.** For each trial, we compute the split half reliability of human gaze patterns by assigning participants to two random splits, averaging the saliency maps for each split, flattening the saliency map, then determining the correlation between these two splits. We repeat this process for 100 random splits for each trial, then apply this same procedure to all trials. This give us an estimate of the reliability of human gaze patterns (yellow). We can shuffle the ordering of one of these splits at each iteration to estimate the empirical null (grey). Finally, for each image, we extract self-attention maps from a single model (DINOv2 B/14), perform PCA on the maps to 3 dimensions, then estimate the correlation between model attention and human gaze patterns (purple). Model attention does not capture any more variance in human gaze patterns then the null distribution, indicating that the features attended to by humans and models is quite different.

A.4 Stimuli

Data from Barense et al., 2007 include common, real-world objects (e.g., color photographs of chairs, tables) and abstract shapes (i.e., synthetic objects without semantic attributes) on a white background (Fig. 2 far left). These stimuli are separated into 'high similarity' and 'low similarity' conditions; 'high similarity' trials are thought to rely on understanding the 'compositional' 3D structure of objects while 'low similarity' trials are thought to rely on simpler visual features. Images from Majaj et al., 2015 belong to four categories of objects: animals (e.g., gorillas, lions), chairs, planes, and faces. Images are in black and white, all from multiple viewpoints, and are superimposed onto randomized backgrounds (e.g., a chair floating in the sky above a mountain range; Fig. 2 middle left). These stimuli are designed to make object segmentation and representation challenging, given the object-irrelevant distractors. The authors from Barense et al., 2007 and Majaj et al., 2015 were contacted directly and provided explicit consent for using these stimuli for all future modeling work. Data from O'Connell et al., 2023 includes two synthetic datasets rendered in Blender from manmade (ShapeNet) and abstract (ShapeGen) objects. The ShapeNet dataset uses objects from multiple classes (e.g. cabinet, car, lamp) rendered with surface textures removed from random views sampled from a sphere around the object (Fig. 2 middle right). The two objects selected for a given trial were always drawn from the same class. The ShapeGen dataset is composed of algorithmically-generated objects using the ShapeGenerator extension for Blender (Fig. 2 far right). These objects are created by iteratively extruding random mesh faces from a base shape, and applying a Catmull-Clark modifier to produce smooth edges on the final objects. This pipeline can generate infinitely many unique shapes, while simultaneously controlling for the similarity of any two objects. These objects were rendered

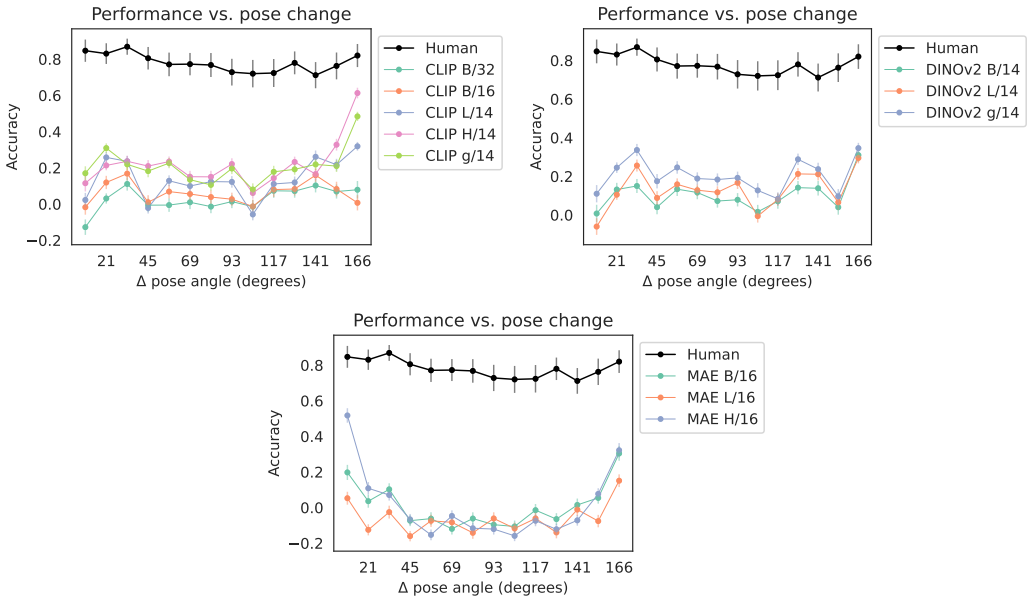

Figure 13: **Visualizing viewpoint tolerance in humans and models across stimuli in shapenet**. Here we determine how model and human performance varies as a function of viewpoint change. Given that we have ground truth viewpoint changes for all objects in shapenet, we determine the change in object pose (in geodesic distance) between the pair of two objects that are the same in each trial (i.e., given a trial with A, A', and B, we determine the viewpoint change between A and A'). We evaluate how viewpoint change effects performance in humans (black, all plots) as well as multiple CLIPs (top left), DINOv2s (top right), and MAEs (bottom). Beyond the mean shift between humans and models, we observe several interesting trends. MAEs across sizes perform well at either end of the viewpoint distribution, when there is very little viewpoint change or a near flip. Larger CLIPs show this same pattern, but only when images are flipped. This pattern is not evident in DINOv2s. It is possible that these patterns relate to the pretraining and data augmentation strategies in each model class; for example, DINOv2 is trained with aggressive local cropping, so it might be more invariant to the pose of the entire object, and not as sensitive to relative pose changes.

from a circular yaw sweep with the camera pointed 45 degrees down at the object. The result is a dataset which allows us to procedurally target a range of human behavior when making zero-shot 3D shape inferences.

A.4.1 Description of all stimulus conditions

Each condition has its uniqe properties and belongs to one stimulus set (either barense, majaj, shapenet, or shapegen). Each stimulus set has properties that are shared across all conditions (e.g., all majaj conditions contain greyscale objects on randomized backgrounds, all shapegen conditions contain procedurally generated non-semantic objects).

1. **Familiar objects low similarity (barense)**: Objects from familiar categories (e.g., cars) in color with a white background; little similarity between matching/non-matching object.
2. **Familiar objects high similarity (barense)**: Objects from familiar categories (e.g., cars) in full color on a white background; high similarity between matching/non-matching object.
3. **Novel objects low similarity (barense)**: Semantically meaningless objects (greebles) in full color on a white background; low similarity between matching/non-matching object.
4. **Novel objects high similarity (barense)**: Semantically meaningless objects (greebles) in full color on a white background; high similarity between matching/non-matching object.
5. **Animals (majaj)**: Eight different animals, greyscale, random category irrelevant background (e.g., mountains, beach).

https://doi.org/10.52202/079017-1379

6. **Chairs (majaj)**: Eight different chairs, greyscale, random category irrelevant background (e.g., mountains, beach).
7. **Planes (majaj)**: Eight different planes, greyscale, random category irrelevant background (e.g., mountains, beach).
8. **Faces (majaj)**: Eight different faces, greyscale, random category irrelevant background (e.g., mountains, beach).
9. **Lamp (shapenet)**: Lamps in greyscale, textures removed, grey background
10. **Telephone (shapenet)**: Telephones in greyscale, textures removed, grey background
11. **Chair (shapenet)**: Chairs in greyscale, textures removed, grey background
12. **Bench (shapenet)**: Benches in greyscale, textures removed, grey background
13. **Sofa (shapenet)**: Sofas in greyscale, textures removed, grey background
14. **Table (shapenet)**: Tables in greyscale, textures removed, grey background
15. **Loudspeaker (shapenet)**: Loudspeakers in greyscale, textures removed, grey background
16. **Display (shapenet)**: Display monitors in greyscale, textures removed, grey background
17. **Cabinet (shapenet)**: Cabinets in greyscale, textures removed, grey background
18. **Car (shapenet)**: Cars in greyscale, textures removed, grey background
19. **Watercraft (shapenet)**: Watercrafts in greyscale, textures removed, grey background
20. **Airplane (shapenet)**: Airplanes in greyscale, textures removed, grey background
21. **Abstract4 (shapegen)**: Procedurally generated non-semantic objects, most dissimilar match/non-matching object, greyscale, textures removed, grey background
22. **Abstract3 (shapegen)**: Procedurally generated non-semantic objects, second-most dissimilar match/non-matching object, greyscale, textures removed, grey background
23. **Abstract2 (shapegen)**: Procedurally generated non-semantic objects, intermediate dissimilarity between match/non-matching object, greyscale, textures removed, grey background
24. **Abstract1 (shapegen)**: Procedurally generated non-semantic objects, second-most similar match/non-matching object, greyscale, textures removed, grey background
25. **Abstract0 (shapegen)**: Procedurally generated non-semantic objects, most similar match/non-matching object, greyscale, textures removed, grey background

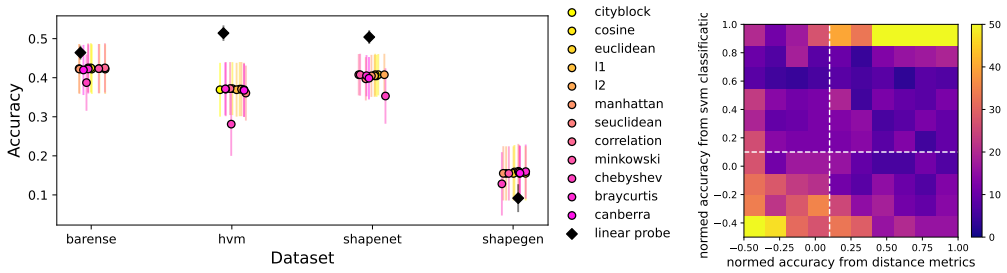

Figure 14: **Comparing distance metrics to a lightweight linear readout.** Left: we visualize the performance of multiple distance metrics (circles) across each datasets, as well as the performance of a lightweight classifier (black, diamond). On average, the classifier outperforms each of the distance metrics. Right: we bin performance and plot a 2D histogram showing the relationship between a cosine distance metric (x axis) and the classifier (y axis). When the distance metric performs relatively well (e.g., above 50% accuracy) the classifier performs near ceiling, but when the distance metric performs poorly (e.g., around chance) the classifier is reliably below chance.

https://doi.org/10.52202/079017-1379