This file provides documentation for this benchmark and its intended uses. All images, human behavioral data, code to evaluate model performance, and scripts used to visualize results are contained within a folder at this link**.** Below we outline information about images, data files, and analysis scripts in this folder, as well as information about licensing, hosting, and maintenance plans. We include metadata for this benchmark in croissant format. Finally, we include supplemental figures and descriptions of online/inlab data collection.

IMAGES ARE IN images/

images/

- Folder with all images in this benchmark
- completely flat organizational structure—no nested folders or directories
- images.zip is a zipped version of images/

DATA FILES ARE IN data/

data/benchmark.csv

- data file containing human/model behavioral data central to this benchmark
- Each row contains information for a single trial (i.e., an image triplet)
- Each column contains relevant behavioral and meta data for this triplet, including
    - ‘images’: list of all images in this trial (which are in /images)
    - ‘human_accuracy’: averaged human performance on this trial
    - ‘human_accuracy_sem”: standard error of the mean for trial
    - ‘human_rt’: averaged reaction time for this trial
    - 'human_rt_sem': standard error of the mean for human reaction times
    - 'human_rt_std': standard deviation of human reaction times
    - ‘condition’: name of image class (e.g., ‘abstract1 contains abstract objects)
    - 'dataset': dataset this trial comes from (e.g., barense, shapenet)
    - 'trial': name of this trial
    - 'n_subjects': number of participants who completed this specific trial
    - 'oddity_index': location of non-matching (i.e., the answer) object in image list
- Additionally, each row contains results from all models evaluated, e.g.
    - 'dinov2-giant_svm_avg': performance of dinov2-giant using linear
    - 'dinov2-giant_svm_std': standard deviation of dinov2-giant on this trial
    - 'dinov2-giant_svm_sem': standard error of the mean for dinov2-giant on trial
    - 'CLIP_ViT-B-32_svm_avg': as above, but for CLIP
    - …
    - 'Vit-mae-base_svm_avg': as above, but for MAE

data/df_behavior_subject.csv

- original human data used to estimate performance saved in ‘benchmark.csv’
- not averaged across trials, but preserves each participants choice behaviors
- necessary for reliability analyses

data/salience_maps.pickle

- dictionary with salience maps for all eye tracking data
- [‘population’][<imagename>] each image’s salience map averaged across the group
- [‘subject’][<imagename>] this images salience maps for each person who saw it

data/df_behavior_wdistance.csv

- Same structure/data as benchmark.csv but with columns for modeling results from distance metrics (instead of a weighted linear readout) using dinov2-base, e.g.:
    - dino_distance_avg: averaged performance of all distance metrics
    - 'dino_distance_std', std of all distance metrics
    - 'dino_distance_sem', sem of all distance metrics
    - 'dino_l1_avg': performance estimated using an l1 distance metric on this trial
    - …

data/croissant.json

- metadata for images/data in benchmark.csv using croissant format

MODELING AND VISUALIZATION SCRIPTS ARE IN scripts/

scripts/model_evaluation.ipynb

- main script for building the lightweight linear probe used in a model analysis
- requires path to df_behavior.csv
- requires path to images/

scripts/results.ipynb

- Main script to load data + generate all plots
- requires path to all data files

scripts/model_attention_analyses.ipynb

- scripts used to extract attention maps from dinov2 and relate to human gaze

scripts/visualize_data_and_singletrial.ipynb

- simple script to load data and visualize a single trial

scripts/relative_pose_analysis.ipynb

- script used to determine how pose variation relates to human/model performance

scripts/load_croissant.ipynb

- simple demo for loading metadata from croissant.json

LICENSING

The authors bear all responsibility in case of violation of rights, etc., and license these data under CC BY-NC-SA 4.0

HOSTING, LICENSING, AND MAINTENANCE PLAN

To ensure access to this benchmark we intend to host all data and images on Huggingface as a Huggingface dataset and will provide all necessary maintenance.

Upon acceptance, these data will be made accessible through Huggingface as well as a project page hosted on GitHub.

A Appendix / supplemental material

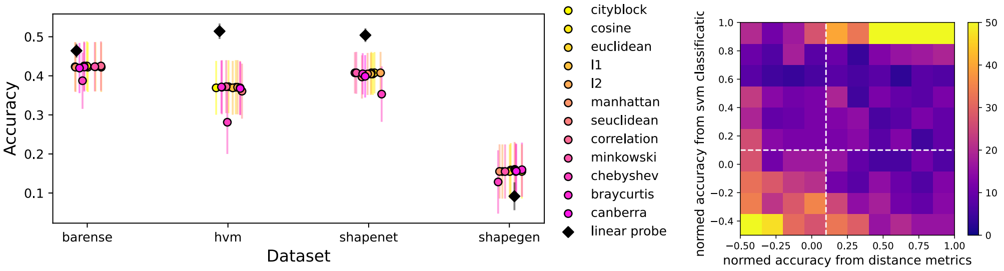

Figure 11:

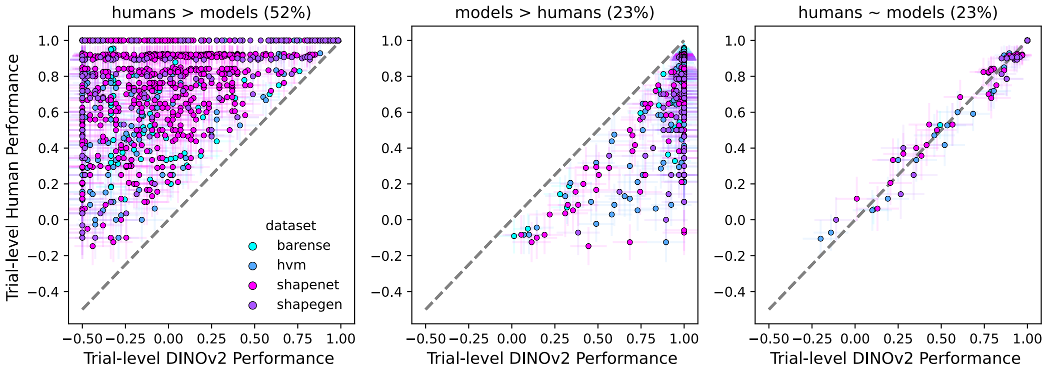

Figure 12: **Direct comparison of trial-level human accuracy to the performance of DINOv2**

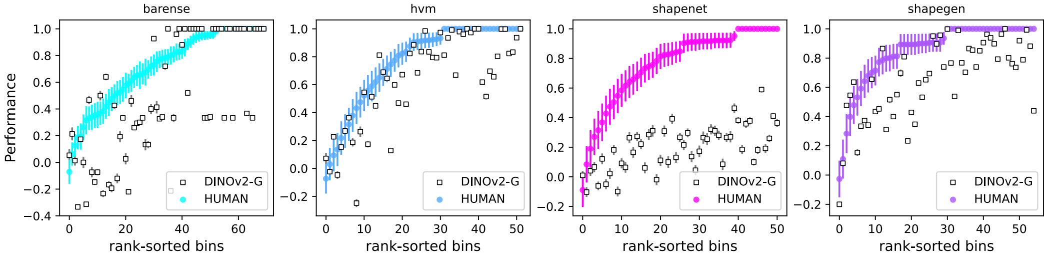

Figure 13:

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,

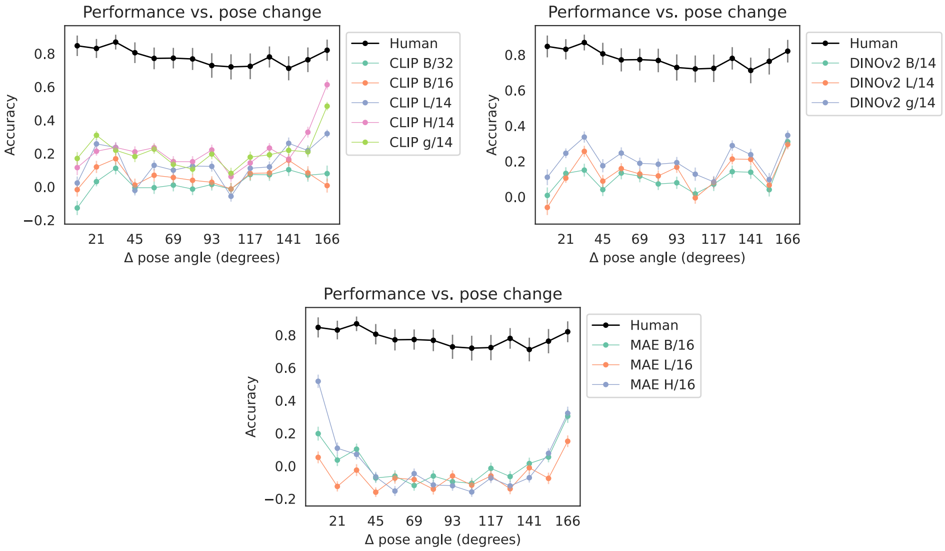

Figure 14: **Visualizing viewpoint tolerance in humans and models across stimuli in shapenet**

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.

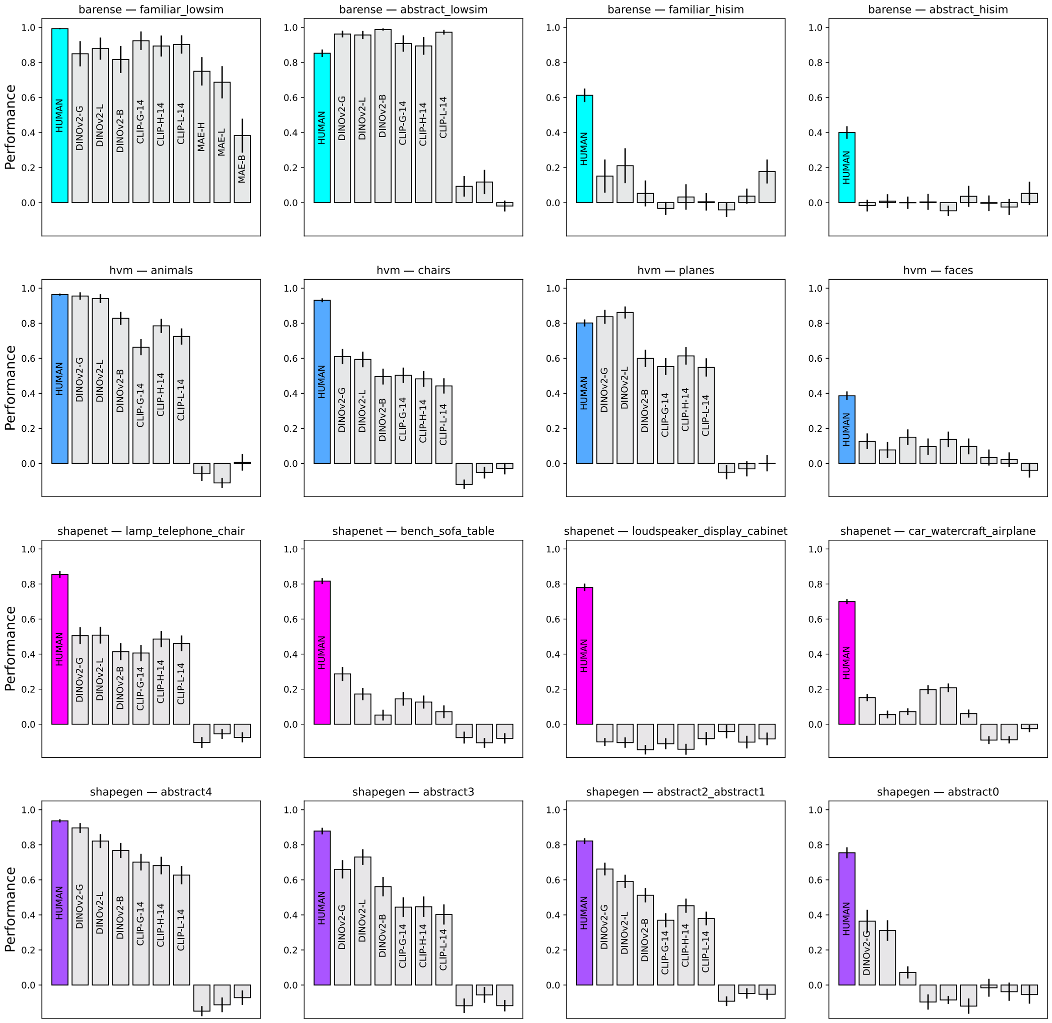

Figure 15: **Comparing human performance to multiple vision models across all conditions.**

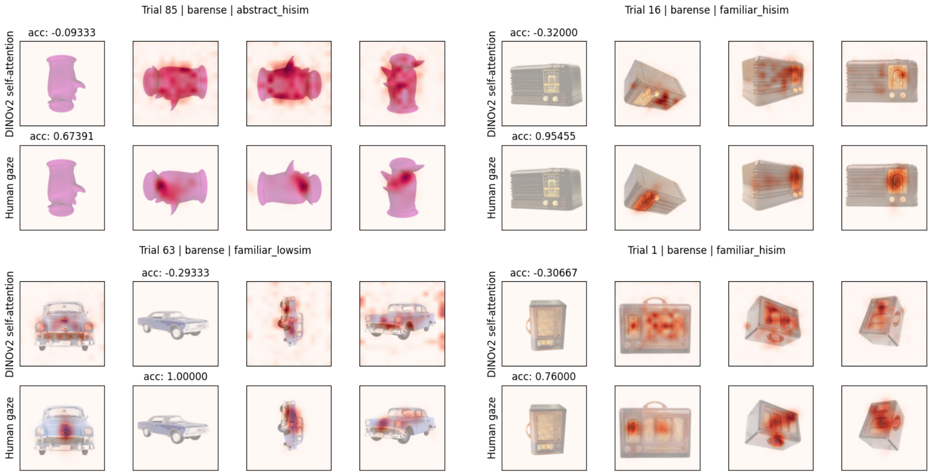
Trial 85 | barense | abstract_hisim
acc: -0.09333
acc: 0.67391
Trial 16 | barense | familiar_hisim
acc: -0.32000
acc: 0.95455
Trial 63 | barense | familiar_lowsim
acc: -0.29333
acc: 1.00000
Trial 1 | barense | familiar_hisim
acc: -0.30667
acc: 0.76000
DINOv2 self-attention
Human gaze

Figure 16: **Example comparisons between DINOv2 attention maps and human attention maps.**

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 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.