MODELING DYNAMIC SOCIAL VISION HIGHLIGHTS GAPS BETWEEN DEEP LEARNING AND HUMANS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning models trained on computer vision tasks are widely considered the most successful models of human vision to date. The majority of work that supports this idea evaluates how accurately these models predict behavior and brain responses to static images of objects and scenes. Real-world vision, however, is highly dynamic, and far less work has evaluated deep learning models on human responses to moving stimuli, especially those that involve more complicated, higher-order phenomena like social interactions. Here, we extend a dataset of natural videos depicting complex multi-agent interactions by collecting human-annotated sentence captions for each video, and we benchmark 350+ image, video, and language models on behavior and neural responses to the videos. As in prior work, we find that many vision models reach the noise ceiling in predicting visual scene features and responses along the ventral visual stream (often considered the primary neural substrate of object and scene recognition). In contrast, vision models poorly predict human action and social interaction ratings and neural responses in the lateral stream (a neural pathway theorized to specialize in dynamic, social vision), though video models show a striking advantage in predicting mid-level lateral stream regions. Language models (given human sentence captions of the videos) predict action and social ratings better than image and video models, but perform poorly at predicting neural responses in the lateral stream. Together, these results identify a major gap in AI's ability to match human social vision and provide insights to guide future model development for dynamic, natural contexts.

1 INTRODUCTION

Over the past decade, significant advances have been made in understanding the computations underlying both biological and artificial vision, in large part due to deep learning models that now provide the best match to human visual behavior and neural responses. However, most research has focused exclusively on static scene and object recognition, neglecting the rich, dynamic interactions that characterize real-world vision. Human vision is tuned to process dynamic social scenes from only a few months of age (Hamlin et al., 2007), yet social vision remains a substantial open challenge in artificial intelligence (AI) (Bolotta & Dumas, 2022), where current AI models struggle to match even human infants in their ability to understand social scenes (Gandhi et al., 2022; Shu et al., 2021). Many have argued that incorporating insights from cognitive (neuro)science may improve AI models' performance on social tasks (Zhou et al., 2019; McMahon & Isik, 2023; Malik & Isik, 2023). However, AI vision and language models are rapidly evolving, and current models have never been comprehensively tested against humans in dynamic, social vision.

The human brain processes dynamic social scenes in regions that are distinct from those involved in classical object perception (Tarhan & Konkle, 2020; Wurm et al., 2017; McMahon & Isik, 2023; Lee Masson & Isik, 2021). These regions form the recently proposed lateral visual stream (Pitcher & Ungerleider, 2021; Wurm & Caramazza, 2022), specialized for dynamic social perception and distinct from the classical ventral "what" and dorsal "where" streams (Ungerleider & Mishkin, 1982; Goodale & Milner, 1992). While AI models effectively predict the ventral stream's computations, little is known about the lateral stream's computations. Recent work suggests computational similarities between the ventral and lateral streams, proposing that lateral stream representations may also be hierarchical, with each computational stage yielding increasingly abstract representations (McMahon et al., 2023). Some work has even suggested that lateral stream computations are not

distinct from those in the ventral stream (Finzi et al., 2022; 2023). However, this claim is based on visually-evoked responses to static images, while lateral stream regions respond primarily to dynamic stimuli (Pitcher et al., 2011; Pitcher & Ungerleider, 2021).

Here, we use a large-scale benchmarking approach to investigate the computational principles underlying human social vision and identify areas of critical need for AI model development (Figure 1). Using over 350 image, video, and language models, we predict human behavioral ratings and fMRI responses to a dataset of publicly-available natural videos depicting human social actions (McMahon et al., 2023). We find that language models based on sentence captions of the videos are best at predicting human social ratings and that video models are best, on average, at predicting brain responses in the lateral visual stream. However, the performance in lateral stream regions is substantially lower than in the ventral visual stream, and no model is able to match both human behavior and brain data. Together, these results highlight a critical need for image-computable models of social perception that match human abilities.

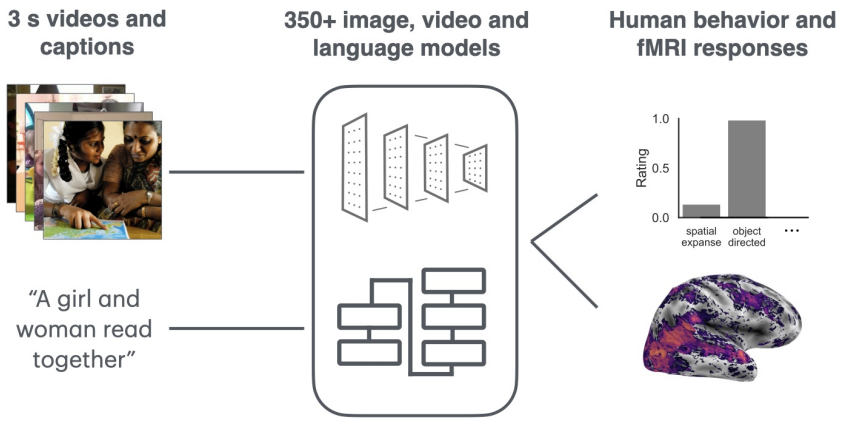

Figure 1: A summary of our overall approach. We extract representations from 350+ image, video, and language models based on 3s videos of human social actions or their captions. We then use DNNs to predict human behavioral ratings and fMRI neural responses to the videos.

2 Related Work

Our approach builds on the NeuroAI benchmarking approach that has been popularized by others (Schrimpf et al., 2018; 2020; Gifford et al., 2023; Willeke et al., 2022; Conwell et al., 2023b; Elmoznino & Bonner, 2024; Chen & Bonner, 2023). For cognitive neuroscience, NeuroAI benchmarking can elucidate the computational factors needed to match the human brain. For AI, these benchmarks can reveal whether different algorithms are human-aligned, and suggest avenues for future model development. The aim of NeuroAI benchmarking is often either to identify the single best model of the brain or behavior (Schrimpf et al., 2018; 2020; Gifford et al., 2023; Willeke et al., 2022), or to understand the computational principles underlying human-model alignment (Conwell et al., 2023b; Elmoznino & Bonner, 2024; Chen & Bonner, 2023).

Here, we take the latter approach, but rather than aiming to understand static scene responses in the ventral visual stream, which are well modeled by most current image models (Conwell et al., 2023b), we aim to understand dynamic visual responses across the brain, focusing on the lateral visual stream and human annotations of the features of interest for these regions. In addition to benchmarking vision models, we also use language models to predict behavioral ratings and visually-evoked neural responses as has been done in prior work with static scenes (Conwell et al., 2023a; Doerig et al., 2022; Geirhos et al., 2021; Linsley et al., 2023; Fel et al., 2022; Lahner et al., 2023). As in Conwell et al. (2023a), we use multiple language models as predictors and selectively perturb the sentence captions of the stimuli to provide insights into the kind of linguistic features that are predictive of visual responses. While modeling dynamic visual events is a growing area of interest (Lahner et al.,

2023), this is the first investigation of benchmarking many models in response to naturalistic videos of human actions.

3 METHODS

3.1 CODE AND DATA AVAILABILITY

All code used in this paper and our sentence captions are publicly available: (*redacted for anonymity, see supplemental files*). The social action ratings and fMRI responses from the original dataset (McMahon et al., 2023) are publicly available on OSF (`https://osf.io/4j29y/`) with a CC-By Attribution 4.0 International license. The videos shown to participants and used here to extract model activations are from the Moments in Time (MiT) dataset (`http://moments.csail.mit.edu`). The MiT license restricts public release of videos from the dataset, but the videos are provided for review and instructions to obtain the videos are available with the original dataset.

3.2 SOCIAL ACTION DATASET

Here, we model behavioral ratings and neural responses from a publicly available dataset of human social actions (McMahon et al., 2023). All experimental details are included in the original publication. Briefly, participants in the fMRI saw the stimuli at approximately 20 degrees of visual angle and were allowed to freely view the videos. Behavioral ratings were collected online from separate participants.

The dataset includes 250 three-second videos of social actions that are divided into 200 videos for training and 50 videos for evaluation. While this dataset is too small for training most AI models, it is on par with other NeuroAI benchmarking studies Conwell et al. (2023b). Each video includes human behavioral ratings of the visual and social scene features. The rated dimensions include descriptions of the visual scene such as how large the scene is (spatial expanse), physical relations between people such as how close together they are (interagent distance), the extent to which they are facing one another (agents facing), and the extent to which an action is object-directed (object directed); and social descriptions including the extent to which people are jointly engaged in an action (acting jointly), the extent to which they are communicating (communicating) and affective features (valence and arousal). Ratings were collected on a Likert scale by at least ten subjects, and we use the average rating for each feature. Inter-rater variability is quite high and used as the noise ceiling against which our models are evaluated.

The dataset also includes voxel-wise fMRI neural responses to each of the videos (beta values) in four participants, and an estimate of the explainable variance determined as the test-retest reliability of responses to the same videos. To restrict our analyses to reliable voxels, we use the reliability mask for each subject's data provided with the original paper. The fMRI data also includes anatomically- (early visual cortex, EVC, and motion-selective middle temporal area, MT) and functionally-defined regions of interest (ROIs) in the ventral and lateral streams. The ventral functional ROIs include the face-selective fusiform face area (FFA) (Kanwisher et al., 1997) and place-selective parahippocampal place area (PPA) (Epstein & Kanwisher, 1998), and the lateral ROIs include the extrastriate body area (EBA), which processes bodies and relations between bodies (Downing et al., 2001; Abassi & Papeo, 2020), the lateral occipital cortex (LOC) which is object selective (Grill-Spector et al., 2001) and is involved in processing object-directed actions (Wurm et al., 2017), and posterior and anterior social-interaction selective regions in the STS (pSTS and aSTS) (Isik et al., 2017; Walbrin et al., 2018; Lee Masson & Isik, 2021; McMahon et al., 2023).

3.3 SENTENCE CAPTIONING OF VIDEOS

In order to evaluate language model prediction of behavioral and neural responses, we collected sentence captions of the videos from 150 online participants using the Prolific platform, and in accordance with our Institutional Review Board. Eligibility criteria included having completed at least 50 tasks with an 85% approval rate, having normal vision (or corrected-to-normal), and being native English speakers. All participants were 18+ (M = 39.72 years old, SD = 13.24) and reported gender and race/ethnicity were as follows: 63 female, 87 male; 114 white, 14 black, 10 asian, 9

mixed race, 2 other, 3 declined to report. Participants were compensated approximately $12 per hour on average for their participation.

Following informed consent, each participant captioned 12 videos presented in a random order: 10 videos from the dataset used in McMahon et al. (2023) and 2 additional catch videos that were the same across all participants. Each video appeared with a text box with grayed out text that read "Description of the actions and interactions of the people in the video in a single sentence..." and disappeared when the subject began typing their caption. Because the study was conducted online and we did not restrict the device or browser used by participants, the visual presentation likely varied among participants.

We collected at least five unique captions for each video in the main dataset. Captions were cleaned by removing participants whose captions on either of the catch videos was more than 2.5 standard deviations away from the mean of other participants in the embedding space of Hugging Face's fine-tuned all-MiniLM-L12-v1 (Wolf et al., 2020; Wang et al., 2020). All-MiniLM-L12-v1 was not reused in subsequent analyses.

3.4 DNN MODEL SELECTION

We used an "opportunistic" modeling approach used in other recent NeuroAI benchmarking research Conwell et al. (2023b). We selected a large set of publicly available models with a variety of modalities, architectures, training sets and objectives, and evaluated their performance along each of these dimensions. Image and video models were selected to represent a comprehensive cross-section of high-level visual tasks, including category supervision, self-supervision, and multimodal (image-language) training, and also include convolutional and transformer architectures. In total, we tested 348 image models from collections including Torchvision (maintainers & contributors, 2016) and Pytorch-Image-Models libraries (Wightman, 2019), VISSL (Goyal et al., 2021), OpenAI CLIP (Radford et al., 2021), and Dectectron2 (Wu et al., 2019), and 8 video models, including Facebook's SlowFast (Feichtenhofer et al., 2018) and TimeSformer (Bertasius et al., 2021) models. We also tested 15 language models, focusing on sentence-transformers, including GPT-2 (Radford et al., 2019) and BERT variants (Devlin et al., 2019). For a full model list and the corresponding license information, please see the supplemental files. The vision and language embeddings for multimodal models were considered separately in model evaluation (e.g., CLIP image encoder is grouped with image models and CLIP's text encoder with language models). We tested fewer video and language models relative to image models due to their availability and computational requirements, respectively. However, this only strengthens our conclusions when either model class outperforms image models.

3.5 MODEL ALIGNMENT WITH BEHAVIORAL AND NEURAL RESPONSES

3.5.1 MODEL FEATURE EXTRACTION

We utilized DeepJuice (Conwell et al., 2023b), a python package in alpha-release shared with us by the authors, that allows for memory efficient feature extraction from each layer of a DNN. We extracted the intermediate representations from every unique computational submodule (referred to here as layers) of every model. We then used GPU-optimized sparse random projection (SRP) implemented in the python package to project the activations in an approximately 4732-dimensional feature space based on the Johnson–Lindenstrauss lemma with $\epsilon = 0.1$ (Larsen & Nelson, 2014).

All model inputs were preprocessed in a model-specific manner. For image models, we extracted activations for seven evenly-sampled frames across the three-seconds of each video and then averaged the activations across frames from the same video. In preliminary analyses, we found that this produced almost identical results to using activations from only a single frame or concatenating activations across the seven-frames. Similar to image models, for language models, we extracted the activations for each caption, and then averaged the activations for the captions from the same video.

3.5.2 LINEAR MAPPING

Before fitting the linear mapping, we first Z-scored the model-SRP feature space across the samples independently for each feature in the 200-video train set defined in the original dataset (McMahon

et al., 2023) and then normalized the held-out data from 50 videos by the mean and standard deviation from the train set. We normalized the behavioral and neural data using the same procedure.

We performed linear mapping between the normalized model-SRP feature space and the normalized behavioral or neural response using leave-one-out ridge regression optimized for the GPU as implemented in DeepJuice (Conwell et al., 2023b). Our α-penalty search space was seven values sampled from a logspace of $10e^{-2}$ to $10e^{5}$. In the training set, we performed 4-fold cross validation in a full sweep of the model to determine the layer that produced the highest performance on the held-out data. Performance was measured as the Pearson correlation between the predicted behavioral or neural response and the true response.

We selected the optimal model layer based on this cross-validation in the training set, and evaluated each models' performance for the optimal layer in the test set. The optimal model layer was selected separately for every behavioral rating and voxel in the brain. For the brain data, we only predicted responses in voxels that were determined to have high test-retest reliability in the original dataset.

3.6 Caption perturbation experiment

To gain insight into what aspect of the captions produced high alignment between the language models and the brain/behavior, we performed selective perturbations to the captions as in (Conwell et al., 2023a; Kauf et al., 2024), but instead of deletion, we performed masking to keep the overall syntactic structure of the sentence intact. We used the spaCy (Honnibal & Montani, 2017) package and the en_core_web_sm model in particular to identify the parts of speech for masking. All other aspects of the experiment followed our model-brain/behavior mapping procedure described above.

3.7 Statistical analysis

To determine whether there was a difference in mean performance between classes of models (e.g., image relative to language), we first computed a null distribution for each model by correlating the permuted predicted response and actual response over 5000 iterations. The same shuffling procedure was used across all models. We compared the true mean difference between model classes to a null distribution of mean differences. The p-value was determined by performing a two-tailed test of the true value against the permuted distributions.

The same procedure was used to determine whether a perturbation on the sentence captions significantly decreased alignment between the language models and the brain/behavior, except using a one-tailed test to compare true degradation (performance on the "orginal" minus performance on perturbed sentence) to chance.

3.8 Compute specifications

The ridge regression for the encoding models required a substantial amount of computational resources. We used an institutional high performance computing cluster equipped with 31 A100 GPU nodes (with a mix of 40 and 80 GB memory). On average, each set of model regressions took approximately 0.38 core hours for the whole-brain results and 0.39 core hours for behavior rating results, for a total of 0.77 core hours per model. To run the full suite of 365 image, video and language models for the reported results took approximately 280 core hours. Full computational resources for the research project (including failed experiments and experiments not reported here) required approximately 1600 core hours.

4 Results

4.1 Behavioral ratings

4.1.1 Language models capture human visual social ratings better than vision models

In evaluating the models, we can compare both how the different model classes (image, video, or language) perform on average at predicting each caption (Figure 2) and the top performing model

for each rating (Supplemental Table 1). We note that due to the larger number of image models tested, the best performing model is biased towards the image models. Despite this, we see a large amount of similarity between the two metrics.

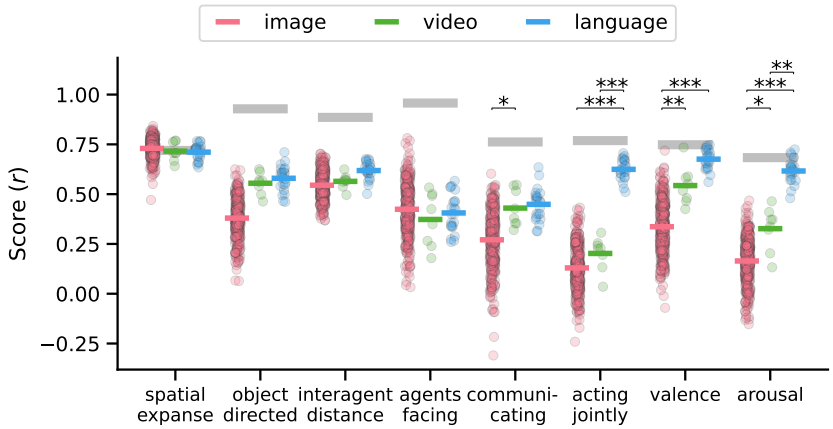

Figure 2: Prediction performance of all models in predicting behavioral responses. Each dot is the performance of a single model. The lines indicate the mean performance for image (pink), video (green), and language (blue) models. The horizontal gray lines are the inter-subject agreement, which is approximately the maximal level that any model could be expected to perform. Brackets and asterisks indicate significantly different performance between different classes of models ($p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***).

We find that for the visuospatial ratings (spatial expanse, interagent distance, and agents facing), no model class is substantially better on average ($p > 0.05$), but for each rating the top performing model is an image model. In contrast, for all social and action ratings, the best performing model is a language model despite the over-representation of image models in our model set. For predicting ratings of object directed actions and communicating, the mean difference between language and vision models is not significant ($p > 0.05$). For ratings of agents acting jointly and affective features (valence and arousal), language models perform better on average than image models ($ps < 0.01$) and video models ($ps < 0.001$), except for valence ($p < 0.05$).

For most ratings, video models do not perform better than image models, except for predictions of communicating, valence, and arousal ($ps < 0.05$). For image models, we do not see different predictions based on a convolutional versus transformer architecture (Supplemental Figure 10), supervised versus self-supervised learning objective (Supplemental Figure 9), or based on language-aligned training (Supplemental Figure 8, 11), or training dataset (Supplemental Figure 12). Together, these results suggest that drastically different vision models perform similarly in predicting video ratings.

4.1.2 Human-Language Model alignment depends on both noun and verb content

To understand the features driving the relatively high performance of language models, we performed selective perturbations on sentence captions by removing nouns or verbs from the captions (Figure 3A). We calculate the degradation in performance as the score of the original, unperturbed input captions (r_o) minus the score of the perturbed input captions (r_p) divided by the score of the original (r_o).

We find that when predicting most ratings shuffling the input captions does not decrease performance, except spatial expanse and agents acting jointly ($p < 0.05$), suggesting that behavior-model alignment does not rely on linguistic compositional structure (Figure 3B).

We can group the remaining perturbations based on disruption to noun content (no nouns, only verbs, teals in Figure 3B) and disruptions to verb content (no verbs, only nouns, pinks in Figure 3B). Prediction of most ratings is degraded by disrupting noun content ($ps < 0.05$), except for agents

A

Category	Linguistic perturbation	Example sentence
Original	—	two men are outside and talking
Shuffled	compositional structure	two and outside talking men are
No nouns	nouns	two [MASK] are outside and talking
Only verbs	nouns + context	[MASK] [MASK] are [MASK] [MASK] talking
No verbs	verbs	two men [MASK] outside and [MASK]
Only nouns	verbs + context	[MASK] men [MASK] [MASK] [MASK] [MASK]

B

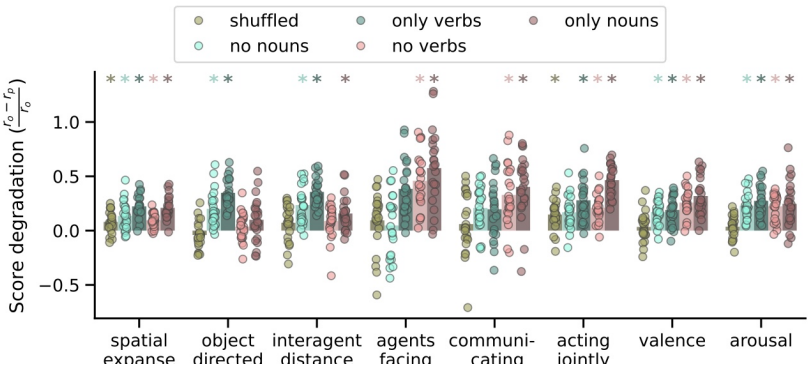

Figure 3: **A**. Example sentence perturbations. **B**. The performance of each language model (dots) in predicting human behavioral ratings following selective perturbation of the sentence captions. The bars indicate the mean performance across models for each condition and rating. Asterisks indicate that there is a significant degradation in model-behavioral alignment following perturbation relative to the unperturbed sentence.

facing and communicating. Prediction of most ratings is also degraded by both verb manipulations (no verbs and only nouns, ps < 0.05), except ratings of object directed actions ($p > 0.05$) and interagent distance (ps > 0.05). Together, these results suggest that the success of language models in predicting visual responses relies on both noun and verb content, with the exception of agents facing / communication and object-directedness, respectively.

4.2 Neural responses

4.2.1 Vision models best capture neural responses

As in behavior, we can compare both the average performance of the models, and the best model for each ROI (Supplemental Table 2). We evaluate performance in ROIs (Figure 4) and the whole brain (Figure 5). We find that for several mid-level ROIs (MT, EBA, and LOC) and a high-level ROI (pSTS), video models outperform image models on average (ps < 0.001). In both early visual cortex and aSTS, the quantitative performance gain is moderate and not significantly different (ps > 0.05), and the best performing model is an image model. Within image models, we we did not find a notable difference in performance based on architecture (Supplemental Figure 15) or training task (Supplemental Figures 16, 7, 18). While there is an improvement of kinetics-trained models over other models in mid-level regions, kinetics-trained models are all video models and the other datasets are all image models. These results further underscore the importance of video processing as a critical manipulation within vision models.

Contrary to what we see in behavior, language models do not outperform vision models on average in any region (ps > 0.05), but they dramatically underperform in EVC, FFA, and PPA (ps < 0.01), and there is no ROI in which the best model is a language model. Language perturbation experiments suggest both noun and verb content are important for prediction in most ROIs (Figure 17).

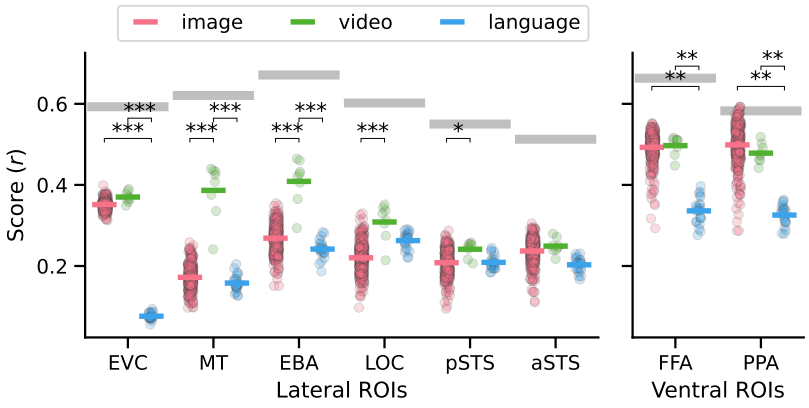

Figure 4: The performance of each model (dots) in predicting the average response in each ROI. The colored lines indicate the mean performance of the different classes of models, and the horizontal gray line is the split-half reliability of the voxel responses in each ROI averaged across participants. Brackets and asterisks indicate significantly different performance between different classes of models ($p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***).

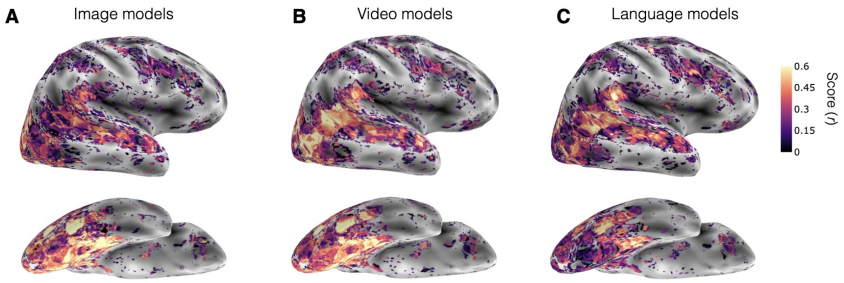

Figure 5: Visualization of the test set encoding performance of the best performing layer in the training set for each voxel from any (**A**) image, (**B**) video, and (**C**) language model. This is shown on the lateral and ventral surface in right hemisphere of one representative participant.

Despite the relatively higher performance by video models in predicting mid-level lateral regions, we still see a striking under performance of even the best models in lateral regions relative to ventral regions (Figures 4, 5).

4.2.2 Hierarchical alignment between models and brains

Previous investigations have found a hierarchical correspondence between vision models and the brain in the ventral temporal cortex in humans (Grill-Spector & Weiner, 2014) and inferiortemporal cortex in macaques (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014), where early and later model layers provide the best match to posterior and anterior brain regions, respectively. We investigated whether there was a similar hierarchical correspondence between models and responses in lateral visual regions by evaluating the relative depth of each voxel in the whole brain (Figure 6) and on average in our regions of interest (Figure 14). We find that though early visual cortex is best predicted by earlier layers in the models, all other regions are predicted by layers of approximately equal depth. In contrast, whole brain results do show a hierarchy along the ventral stream, thus highlighting an additional gap between models and lateral stream regions, in particular.

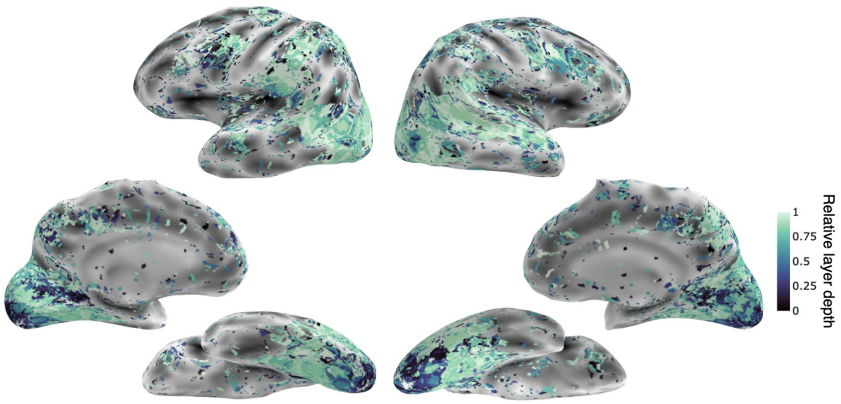

Figure 6: Relative depth of the best performing model layer across all vision models (image and video models) in the whole brain of one representative subject.

5 Discussion

We use a large-set of image, video, and language models to predict human behavioral ratings and and neural responses to dynamic social scenes. Overall, we found a notable gap in all models' ability to predict human responses. However, there are differences in the models that are best able to predict the brain versus behavior. In particular, language models tend to be the best models of human behavioral ratings, while video models best predict responses in lateral brain regions.

5.1 Language and video models for social scene understanding

The fact that language models align with human social ratings may suggest that humans rely on non-visual aspects of social interactions to rate social features (McMahon et al., 2023; Netanyahu et al., 2021). While this may be partially true, it is unlikely to be the whole answer because humans make many of these social judgements quickly and automatically (McMahon & Isik, 2023; Quadflieg & Penton-Voak, 2017). Further, these behavioral ratings strongly predict visual regions of the brain (McMahon et al., 2023; McMahon & Isik, 2023), which the language models cannot explain. Another reason the language models predict behavior so well may be due to the captioning prompt. By instructing participants to caption the "actions and interactions" in the videos, we might have biased the language model embeddings to better predict social action content. It is therefore somewhat surprising that language models do not perform better because, for example, high-communication videos are often captioned explicitly with verbs like"'talking" (e.g., example caption in Figure 3), and yet most language models still fall short of human agreement.

In contrast, video models provided a boost to prediction in mid-level lateral regions, which is surprising given the lack of available high-performing video models relative to image models, but they did not predict behavioral ratings or more anterior regions significantly better than image models. Together, this work reveals a significant gap in even state-of-the-art models' abilities to match human social action judgements and the underlying neural substrates. These results also provide insights into future directions for model development that can integrate both the relational structure that is readily present in language with dynamic visual information.

5.2 NeuroAI in dynamic social contexts

One major advantage of this work over most prior NeuroAI studies is the focus on dynamic, social scenes. In addition to the human-model gap, several other interesting findings come out of testing models in more ecologically valid conditions. First, prior work has suggested that affective features like valence and arousal can be extracted by relatively simple convolutional neural networks (Kragel et al., 2019; Conwell et al., 2021). While these features may be image computable, our work shows that in dynamic events, few existing vision models can match human social-affective ratings.

Further, unlike prior work with static scenes (Conwell et al., 2023a; Doerig et al., 2022), language models were not able to capture brain responses in the current dataset. They perform dramatically worse than image models in predicting responses in EVC, as has been documented elsewhere (Conwell et al., 2023a), but also significantly worse in high-level ventral regions (Figures 4, 5). This result calls into question strong ideas of "language alignment" in the visual cortex and highlights the importance of dynamic stimuli for studies of even the ventral visual stream (Haxby et al., 2020).

5.3 Limitations

One major challenge in NeuroAI is that most datasets available (Allen et al., 2022), though rich, are based on responses to static images with relatively little social content. This study aims to fill this gap with the current dataset of 250 naturalistic social videos. However, given the limited size of the dataset, there are inherent challenges in training novel deep learning models with these videos. These underscore the need for larger, more diverse datasets that can better represent the complexity of real-world social interactions. While larger dynamic vision neuroscience datasets exist (Lahner et al., 2023), they do not rival existing static image datasets like the Natural Scenes Dataset (Allen et al., 2022). Similarly, computer vision video datasets (Kay et al., 2017; Monfort et al., 2019; Thomee et al., 2016) lack the size and diversity of image datasets (Deng et al., 2009; Thomee et al., 2016).

Another limitation is that this work, as in most NeuroAI benchmarking studies, takes advantage of available models that differ along many factors, making it difficult to isolate the impact of any one computational factor on performance. This limitation may explain why we do not see significant differences for seemingly important model factors, such as convolutional versus transformer architectures (Supplemental Figures 10, 15). We aimed to overcome this by testing a wide array of models, but as noted above, our set includes relatively few video and language models compared to image models. Future work should investigate how larger language models compare to human on social visual tasks, though we note here that the largest language model tested (GPT-2) was far from the top performing model in most cases (Supplemental Tables 1, 2) and models with more tunable parameters did not always yield higher human data predictivity (Supplemental Figures 13, 20).

5.4 Future Directions

Moving forward, we suggest a couple avenues for advancing model development based on these results. The relative success of language and video models in matching behavior and brain respectively, suggest that models that explicitly represent agents, objects, and their relations over time will be critical to future modeling endeavors. While transformer models should be able to capture some of this relational information in theory, they still fall short in predicting the brain and behavior in the current dataset. This failure is likely due to the datasets and tasks they are trained on that do not lead to learning the kinds of agent and object-based representations that humans deploy in social reasoning tasks. It is possible that training on more human-aligned tasks will allow these large transformer models to pick up on relevant structure in social scenes.

Human-aligned DNNs may be a promising direction for dynamic social perception. For example, prior work has shown that implementing neuroscience-inspired circuits into transformers provides more human-aligned object tracking (Linsley et al., 2021). Other work has focused on training modern DNNs with more ecologically valid data. For example, Mineault et al. (2021) trained a 3D ResNet model on agent self-motion to model dorsal stream responses, while Orhan & Lake (2024) used infant headcam data to train models on high-level visual representations. We included the latter models in our set and while they performed relatively well in our benchmarks, they were still not able to reach human social ratings or neural responses (Supplemental Tables 1, 2).

Finally, efforts to merge structured cognitive modeling (Malik & Isik, 2023; Shu et al., 2021; Netanyahu et al., 2021) with image-computable DNNs could be another direction. One example baseline model, a combined GNN-RNN, tracks entities and their relations over time in social scenes on similar social datasets (Malik & Isik, 2023). Related prior work has also suggested that generative inverse planning models may be helpful in closing the gap between human social perception and current DNNs (Shu et al., 2021; Netanyahu et al., 2021). While most instantiations of these models are not image computable, this is an active area of model development that we believe is likely to yield advances in matching human social perception.

5.5 Conclusions

This work identifies human social vision as a key area of need for future AI research. It also offers some promising directions for future modeling endeavors. In particular, the relative success of language and video models over image models in predicting social behavior and brain responses suggests that models combining both compositional and dynamic information may be critical for more human-like social AI.

Reproducibility Statement

To ensure the reproducibility of our results we have included detailed descriptions of the dataset, preprocessing steps, model selection, feature extraction, linear mapping, and statistical analyses in the supplementary files. The supplemental files include the code, data repositories and model licenses for all models tested. The experiments are reproducible using the steps provided in the README file. The code is dependent on one additional Python library that is available upon request. Instructions to obtain this library are also included in the README file.

Ethics Statement

The research adheres to all relevant ethical guidelines. The human subjects experiments conducted for this paper were done in accordance with procedures approved by our institutional IRB. The original publicly available neuroimaging dataset was also conducted in accordance with IRB, and all human subjects were compensated for their time. The model benchmarking used open source models and all model repositories and license details are included with supplementary files.

References

Etienne Abassi and Liuba Papeo. The Representation of Two-Body Shapes in the Human Visual Cortex. *Journal of Neuroscience*, 40(4):852–863, 2020. doi: 10.1523/JNEUROSCI.1378-19.2019. Publisher: Society for Neuroscience.

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-021-00962-x. URL https://www.nature.com/articles/s41593-021-00962-x.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding?, June 2021. URL http://arxiv.org/abs/2102.05095. arXiv:2102.05095 [cs].

Samuele Bolotta and Guillaume Dumas. Social Neuro AI: Social Interaction as the "Dark Matter" of AI. *Frontiers in Computer Science*, 4, 2022. ISSN 2624-9898. URL https://www.frontiersin.org/articles/10.3389/fcomp.2022.846440.

Zirui Chen and Michael Bonner. Canonical dimensions of vision. In *2023 Conference on Cognitive Computational Neuroscience*, Oxford, UK, 2023. Cognitive Computational Neuroscience. doi: 10.32470/CCN.2023.1588-0. URL https://2023.ccneuro.org/view_paper.php?PaperNum=1588.

Colin Conwell, Daniel Graham, and Edward A Vessel. The perceptual primacy of feeling: Affectless machine vision models robustly predict human visual arousal, valence, and aesthetics. *PsyArXiv*, 2021. doi: https://doi.org/10.31234/osf.io/5wg4s.

Colin Conwell, Jacob S. Prince, George A. Alvarez, and Talia Konkle. Language Models of Visual Cortex: Where do they work? And why do they work so well where they do? *Journal of Vision*, 23(9):5653, August 2023a. ISSN 1534-7362. doi: 10.1167/jov.23.9.5653. URL https://jov.arvojournals.org/article.aspx?articleid=2792615.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?, July 2023b. URL https://www.biorxiv.org/content/10.1101/2022.03.28.485868v2. Pages: 2022.03.28.485868 Section: New Results.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. Semantic scene descriptions as an objective of human vision, September 2022. URL http://arxiv.org/abs/2209.11737. arXiv:2209.11737 [cs, q-bio].

Paul E. Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A Cortical Area Selective for Visual Processing of the Human Body. *Science*, 293(5539):2470–2473, 2001. doi: 10.1126/science.1063414. Publisher: American Association for the Advancement of Science.

Eric Elmoznino and Michael F. Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, 20(1):e1011792, January 2024. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011792. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011792. Publisher: Public Library of Science.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, April 1998. ISSN 0028-0836, 1476-4687. doi: 10.1038/33402. URL https://www.nature.com/articles/33402.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. *arXiv: 1812.03982*, 2018.

Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9432–9446. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3d681cc4487b97c08e5aa67224dd74f2-Paper-Conference.pdf.

Dawn Finzi, Daniel Yamins, Kendrick Kay, and Kalanit Grill-Spector. Do deep convolutional neural networks accurately model representations beyond the ventral stream? In *2022 Conference on Cognitive Computational Neuroscience*, San Francisco, 2022. Cognitive Computational Neuroscience. doi: 10.32470/CCN.2022.1219-0. URL https://2022.ccneuro.org/view_paper.php?PaperNum=1219.

Dawn Finzi, Eshed Margalit, Kendrick Kay, Daniel L. K. Yamins, and Kalanit Grill-Spector. A single computational objective drives specialization of streams in visual cortex, December 2023. URL https://www.biorxiv.org/content/10.1101/2023.12.19.572460v1. Pages: 2023.12.19.572460 Section: New Results.

Kanishk Gandhi, Gala Stojnic, Brenden M. Lake, and Moira R. Dillon. Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others, February 2022. URL http://arxiv.org/abs/2102.11938. arXiv:2102.11938 [cs].

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision, October 2021. URL http://arxiv.org/abs/2106.07411. arXiv:2106.07411 [cs, q-bio].

A. T. Gifford, B. Lahner, S. Saba-Sadiya, M. G. Vilas, A. Lascelles, A. Oliva, K. Kay, G. Roig, and R. M. Cichy. The Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes, July 2023. URL http://arxiv.org/abs/2301.03198. arXiv:2301.03198 [cs, q-bio].

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. ISSN 0166-2236. doi: https://doi.org/10.1016/0166-2236(92)90344-8. URL https://www.sciencedirect.com/science/article/pii/0166223692903448.

Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. Vissl. https://github.com/facebookresearch/vissl, 2021.

Kalanit Grill-Spector and Kevin S. Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014. doi: 10.1038/nrn3747. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4143420/.

Kalanit Grill-Spector, Zoe Kourtzi, and Nancy Kanwisher. The lateral occipital complex and its role in object recognition. *Vision Research*, 41(10-11):1409–1422, May 2001. ISSN 00426989. doi: 10.1016/S0042-6989(01)00073-6. URL https://linkinghub.elsevier.com/retrieve/pii/S0042698901000736.

J. Kiley Hamlin, Karen Wynn, and Paul Bloom. Social evaluation by preverbal infants. *Nature*, 450 (7169):557–559, November 2007. ISSN 1476-4687. doi: 10.1038/nature06288. URL https://doi.org/10.1038/nature06288.

James V. Haxby, M. Ida Gobbini, and Samuel A. Nastase. Naturalistic stimuli reveal a dominant role for agentic action in visual representation. *NeuroImage*, 216:116561, August 2020. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2020.116561.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Leyla Isik, Kami Koldewyn, David Beeler, and Nancy Kanwisher. Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 114(43), October 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1714471114. URL https://pnas.org/doi/full/10.1073/pnas.1714471114.

Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, 17(11): 4302–4311, June 1997. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.17-11-04302.1997. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.17-11-04302.1997.

Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. Lexical-Semantic Content, Not Syntactic Structure, Is the Main Contributor to ANN-Brain Similarity of fMRI Responses in the Language Network. *Neurobiology of Language*, 5(1):7–42, April 2024. ISSN 2641-4368. doi: 10.1162/nol_a_00116. URL https://doi.org/10.1162/nol_a_00116.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *arXiv: 1705.06950*, 2017.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11): e1003915, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003915. URL https://dx.plos.org/10.1371/journal.pcbi.1003915.

Philip A. Kragel, Marianne C. Reddan, Kevin S. LaBar, and Tor D. Wager. Emotion schemas are embedded in the human visual system. *Science Advances*, 5(7):eaaw4358, July 2019. doi: 10.1126/sciadv.aaw4358. URL https://www.science.org/doi/10.1126/sciadv.aaw4358. Publisher: American Association for the Advancement of Science.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N. Apurva Ratan Murty, Kendrick Kay, Aude Oliva, and Radoslaw Cichy. BOLD Moments: modeling short visual events through a video fMRI dataset and metadata. *Nature Communications*, 15:6241–6267, March 2023. doi: 10.1038/s41467-024-50310-3. URL https://www.nature.com/articles/s41467-024-50310-3.

Kasper Green Larsen and Jelani Nelson. The Johnson-Lindenstrauss lemma is optimal for linear dimensionality reduction, November 2014. URL http://arxiv.org/abs/1411.2404. arXiv:1411.2404 [cs, math].

Haemy Lee Masson and Leyla Isik. Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage*, 245:118741, December 2021. ISSN 10538119. doi: 10.1016/j.neuroimage.2021.118741. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811921010132.

Drew Linsley, Girik Malik, Junkyung Kim, Lakshmi N Govindarajan, Ennio Mingolla, and Thomas Serre. Tracking without re-recognition in humans and machines, 2021. URL https://arxiv.org/abs/2105.13351.

Drew Linsley, Ivan F. Rodriguez Rodriguez, Thomas Fel, Michael Arcaro, Saloni Sharma, Margaret Livingstone, and Thomas Serre. Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. *Advances in Neural Information Processing Systems*, 36:28873–28891, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5bf234ecf83cd77bc5b77a24ba9338b0-Abstract-Conference.html.

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016.

Manasi Malik and Leyla Isik. Relational visual representations underlie human social interaction recognition. *Nature Communications*, 14(1):7317, November 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-43156-8. URL https://www.nature.com/articles/s41467-023-43156-8. Number: 1 Publisher: Nature Publishing Group.

Emalie McMahon and Leyla Isik. Seeing social interactions. *Trends in Cognitive Sciences*, 27(12):1165–1179, December 2023. ISSN 13646613. doi: 10.1016/j.tics.2023.09.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661323002486.

Emalie McMahon, Michael F. Bonner, and Leyla Isik. Hierarchical organization of social action features along the lateral visual pathway. preprint, PsyArXiv, March 2023. URL https://osf.io/x3avb.

Patrick J Mineault, Shahab Bakhtiari, Blake Aaron Richards, and Christopher C Pack. Your head is there to move you around: Goal-driven models of the primate dorsal pathway. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=6p-zJaheTW.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, and others. Moments in Time Dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–8, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2901464.

Aviv Netanyahu, Tianmin Shu, Boris Katz, Andrei Barbu, and Joshua B. Tenenbaum. PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception, March 2021. URL http://arxiv.org/abs/2103.01933. arXiv:2103.01933 [cs, stat].

A. Emin Orhan and Brenden M. Lake. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, pp. 271–283, 2024. doi: 10.1038/s42256-024-00802-0. URL https://www.nature.com/articles/s42256-024-00802-0#citeas.

David Pitcher and Leslie G. Ungerleider. Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, 25(2):100–110, February 2021. ISSN 13646613. doi: 10.1016/j.tics.2020.11.006. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661320302783.

David Pitcher, Daniel D. Dilks, Rebecca R. Saxe, Christina Triantafyllou, and Nancy Kanwisher. Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage*, 56(4):2356–2363, June 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.03.067.

Susanne Quadflieg and Ian S. Penton-Voak. The Emerging Science of People-Watching: Forming Impressions From Third-Party Encounters. *Current Directions in Psychological Science*, 26(4):383–389, August 2017. ISSN 0963-7214. doi: 10.1177/0963721417694353. URL https://doi.org/10.1177/0963721417694353. Publisher: SAGE Publications Inc.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? preprint, Neuroscience, September 2018. URL http://biorxiv.org/lookup/doi/10.1101/407007.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020. URL https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. AGENT: A Benchmark for Core Psychological Reasoning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9614–9625. PMLR, July 2021. URL https://proceedings.mlr.press/v139/shu21a.html. ISSN: 2640-3498.

Leyla Tarhan and Talia Konkle. Sociality and interaction envelope organize visual action representations. *Nature Communications*, 11(1):3002, June 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16846-w. URL https://www.nature.com/articles/s41467-020-16846-w.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 59(2):64–73, January 2016. ISSN 0001-0782, 1557-7317. doi: 10.1145/2812802. URL http://arxiv.org/abs/1503.01817. arXiv:1503.01817 [cs].

Leslie G. Ungerleider and Mortimer Mishkin. Two cortical visual systems. In David J. Ingle, Melvyn A. Goodale, and Richard J. W. Mansfield (eds.), *Analysis of Visual Behavior*, pp. 549–586. The MIT Press, 1982.

Jon Walbrin, Paul Downing, and Kami Koldewyn. Neural responses to visually observed social interactions. *Neuropsychologia*, 112:31–39, April 2018. ISSN 1873-3514. doi: 10.1016/j.neuropsychologia.2018.02.023.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, April 2020. URL http://arxiv.org/abs/2002.10957. arXiv:2002.10957 [cs].

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Pede, Max F. Burg, Christoph Blessing, Santiago A. Cadena, Zhiwei Ding, Konstantin-Klemens Lurz, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. The Sensorium competition on predicting large-scale mouse primary visual cortex activity, June 2022. URL http://arxiv.org/abs/2206.08666. arXiv:2206.08666 [cs, q-bio].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020. URL http://arxiv.org/abs/1910.03771. arXiv:1910.03771 [cs].

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Moritz F. Wurm and Alfonso Caramazza. Two 'what' pathways for action and object recognition. *Trends in Cognitive Sciences*, 26(2):103–116, February 2022. ISSN 13646613. doi: 10.1016/j.tics.2021.10.003. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661321002588.

Moritz F. Wurm, Alfonso Caramazza, and Angelika Lingnau. Action Categories in Lateral Occipitotemporal Cortex Are Organized Along Sociality and Transitivity. *Journal of Neuroscience*, 37(3):562–575, 2017. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1717-16.2016. URL https://www.jneurosci.org/content/37/3/562. Publisher: Society for Neuroscience _eprint: https://www.jneurosci.org/content/37/3/562.full.pdf.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1403112111. URL https://pnas.org/doi/full/10.1073/pnas.1403112111.

Chen Zhou, Ming Han, Qi Liang, Yi-Fei Hu, and Shu-Guang Kuai. A social interaction field model accurately identifies static and dynamic social groupings. *Nature Human Behaviour*, 3(8):847–855, August 2019. ISSN 2397-3374. doi: 10.1038/s41562-019-0618-2. URL https://www.nature.com/articles/s41562-019-0618-2. Publisher: Nature Publishing Group.

A SUPPLEMENTAL MATERIALS

Table 1: A table of the top-30 performing models averaged across features. Models are arranged in descending order of their overall average performance. Bold scores highlight the model that is the top performing model for a given feature.

Row number	Model name	Model class	spatial expanse	object directed	interagent distance	agents facing	communicating	acting jointly	valence	arousal
1	paraphrase-MiniLM-L6-v2	language	0.764	0.576	0.664	0.534	**0.635**	0.583	0.692	0.684
2	all-mpnet-base-v2	language	0.688	0.665	0.676	0.537	0.384	0.678	0.743	**0.725**
3	paraphrase-multilingual-MiniLM-L12-v2	language	0.725	0.609	0.621	0.523	0.595	0.625	0.684	0.647
4	all-mpnet-base-v1	language	0.766	**0.711**	0.615	0.424	0.313	0.643	**0.747**	0.709
5	FacebookAI_roberta-large-mnli	language	0.725	0.605	0.608	0.567	0.468	0.683	0.664	0.605
6	mixedbread-ai_mxbai-embed-2d-large-v1	language	0.766	0.520	0.670	0.541	0.431	**0.706**	0.686	0.588
7	paraphrase-multilingual-mpnet-base-v2	language	0.735	0.610	0.652	0.462	0.504	0.660	0.705	0.572
8	all-distilroberta-v1	language	0.722	0.582	0.605	0.458	0.465	0.603	0.709	0.638
9	stsb-distilroberta-base-v2	language	0.660	0.624	0.623	0.401	0.506	0.586	0.720	0.621
10	all-roberta-large-v1	language	0.760	0.548	0.621	0.428	0.400	0.691	0.634	0.613
11	distiluse-base-multilingual-cased-v1	language	0.740	0.630	0.669	0.266	0.489	0.582	0.656	0.651
12	clip-ViT-B-32-multilingual-v1	language	0.709	0.607	0.545	0.399	0.483	0.669	0.724	0.541
13	clip_vitl14	image	0.771	0.535	0.660	**0.781**	0.492	0.319	0.718	0.392
14	mixedbread-ai_mxbai-colbert-large-v1	language	0.660	0.540	0.570	0.432	0.528	0.670	0.652	0.599
15	mixedbread-ai_mxbai-embed-large-v1	language	0.690	0.558	0.633	0.465	0.461	0.630	0.631	0.570
16	all-MiniLM-L6-v2	language	0.700	0.627	0.579	0.343	0.481	0.529	0.661	0.636
17	FacebookAI_xlm-roberta-large	language	0.729	0.650	0.602	0.357	0.401	0.620	0.561	0.612
18	all-MiniLM-L6-v1	language	0.634	0.605	0.676	0.260	0.492	0.611	0.596	0.637
19	FacebookAI_xlm-roberta-base	language	0.657	0.494	0.641	0.357	0.374	0.610	0.732	0.633
20	multi-qa-MiniLM-L6-cos-v1	language	0.686	0.461	0.648	0.337	0.347	0.643	0.668	0.700
21	gpt2	language	0.730	0.599	0.637	0.340	0.315	0.555	0.624	0.631
22	LaBSE	language	0.648	0.550	0.563	0.274	0.477	0.674	0.625	0.568
23	timm_deit3_large_patch16_224_in21ft1k	image	0.820	0.504	0.663	0.628	0.371	0.340	0.629	0.300
24	FacebookAI_roberta-base	language	0.727	0.499	0.607	0.335	0.400	0.511	0.696	0.480
25	all-MiniLM-L12-v2	language	0.731	0.465	0.501	0.288	0.373	0.603	0.733	0.515
26	timm_beitv2_large_patch16_224	image	**0.842**	0.503	0.642	0.667	0.505	0.358	0.417	0.143
27	timm_deit3_base_patch16_224_in21ft1k	image	0.789	0.498	0.584	0.649	0.291	0.375	0.514	0.333
28	timm_beit_large_patch16_224	image	0.828	0.506	**0.703**	0.755	0.277	0.184	0.588	0.184
29	timm_beit_large_patch16_384	image	0.811	0.565	0.644	0.705	0.278	0.298	0.634	0.084
30	timm_deit3_large_patch16_384_in21ft1k	image	0.804	0.510	0.617	0.597	0.481	0.417	0.333	0.224

Table 2: A table of the top-30 models on averaged across all ROIs. Models are arranged in descending order of their overall average performance in all reliable voxels. Bold scores highlight the top score in each ROI. The bottom row indicates the performance in predicting the ROIs based on hand engineered features from McMahon et al. (2023).

Rank	Model name	Model class	EVC	MT	EBA	LOC	pSTS	aSTS	FFA	PPA
1	x3d_s	video	0.390	**0.440**	0.427	**0.352**	0.251	0.249	0.510	0.518
2	timm_beitv2_base_patch16_224	image	0.369	0.245	0.349	0.326	0.275	0.293	0.536	0.588
3	timm_beitv2_large_patch16_224	image	0.376	0.259	0.355	0.330	**0.287**	0.291	0.524	0.593
4	x3d_m	video	0.386	0.425	0.460	0.336	0.255	0.280	0.512	0.492
5	timm_beit_large_patch16_224	image	0.373	0.241	0.335	0.314	0.277	**0.297**	0.520	0.592
6	timm_beit_large_patch16_384	image	0.368	0.251	0.336	0.317	0.253	0.266	0.501	**0.593**
7	clip_vitl14	image	0.376	0.223	0.325	0.289	0.245	0.278	0.526	0.574
8	i3d_r50	video	0.378	0.433	**0.465**	0.345	0.254	0.239	0.516	0.442
9	timm_deit3_huge_patch14_224_in21ft1k	image	0.375	0.233	0.320	0.291	0.264	0.287	0.538	0.587
10	timm_deit3_large_patch16_384_in21ft1k	image	0.376	0.241	0.328	0.292	0.253	0.276	0.517	0.574
11	timm_beit_base_patch16_224	image	0.365	0.228	0.322	0.298	0.258	0.267	0.522	0.578
12	timm_deit3_large_patch16_224_in21ft1k	image	0.371	0.241	0.337	0.295	0.245	0.274	0.499	0.584
13	slow_r50	video	0.358	0.408	0.421	0.329	0.248	0.245	0.512	0.477
14	timm_convnext_large	image	0.373	0.230	0.318	0.284	0.252	0.273	0.502	0.559
15	timm_convnext_large_in22ft1k	image	0.373	0.230	0.318	0.284	0.252	0.273	0.502	0.559
16	timm_convnext_xlarge_in22k	image	0.370	0.219	0.303	0.283	0.256	0.265	0.515	0.555
17	c2d_r50	video	0.371	0.376	0.405	0.305	0.258	0.272	0.511	0.462
18	timm_deit3_medium_patch16_224_in21ft1k	image	0.363	0.228	0.348	0.284	0.249	0.270	0.522	0.568
19	timm_convnext_xlarge_in22ft1k	image	0.372	0.233	0.320	0.285	0.247	0.268	0.501	0.558
20	clip_rn50x4	image	0.376	0.229	0.316	0.277	0.237	0.272	**0.552**	0.568
21	timm_convnext_large_in22k	image	0.368	0.225	0.307	0.269	0.249	0.282	0.520	0.555
22	timm_mixer_b16_224_miil_in21k	image	**0.399**	0.219	0.304	0.254	0.238	0.292	0.542	0.559
23	timm_beit_base_patch16_384	image	0.358	0.230	0.331	0.292	0.235	0.266	0.512	0.568
24	timm_convnext_base_in22ft1k	image	0.364	0.225	0.324	0.275	0.240	0.258	0.517	0.549
25	timm_convnext_base	image	0.364	0.225	0.324	0.275	0.240	0.258	0.517	0.549
26	timm_deit3_base_patch16_224_in21ft1k	image	0.362	0.217	0.321	0.286	0.224	0.257	0.539	0.576
27	slowfast_r50	video	0.363	0.435	0.432	0.313	0.242	0.250	0.474	0.461
28	clip_vitb32	image	0.376	0.203	0.300	0.246	0.262	0.241	0.507	0.568
29	timm_deit3_base_patch16_384_in21ft1k	image	0.367	0.209	0.289	0.269	0.248	0.269	0.526	0.566
30	timm_convnext_base_in22k	image	0.362	0.217	0.308	0.276	0.228	0.250	0.515	0.559
-	McMahon et al. (2023)	hand engineered	0.347	0.392	0.365	0.336	0.261	0.292	0.418	0.435

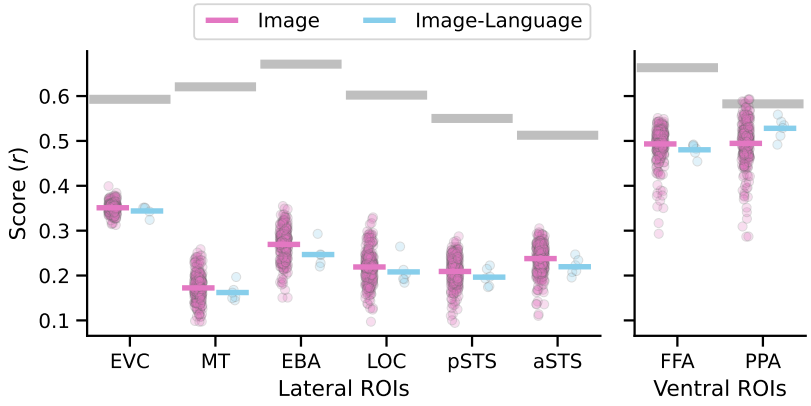

Figure 7: The performance of models in our set trained either with an image-based (image, pink) or multimodal (image-language, blue) objective functionin predicting neural responses in ROIs in the lateral and ventral visual streams. Plotting conventions are the same as Figure 4

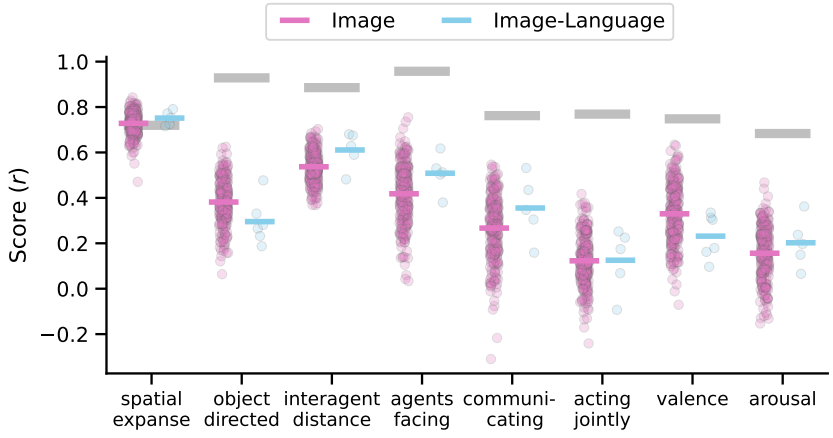

Figure 8: The performance of models in our set trained either with an image-based (image, pink) or multimodal (image-language, blue) objective functionin predicting behavior ratings. Plotting conventions are the same as Figure 2.

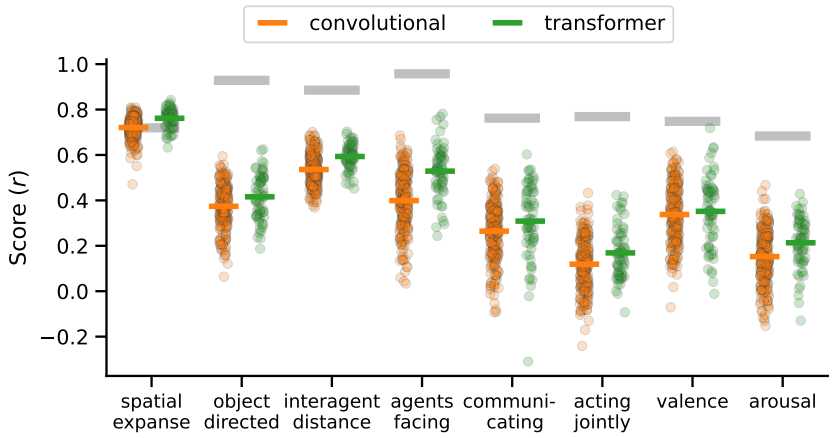

Figure 9: The performance of each image model in our set in predicting behavioral responses separated by whether the model has a convolutional (orange) or transformer (green) architecture. Plotting conventions are the same as Figure 2.

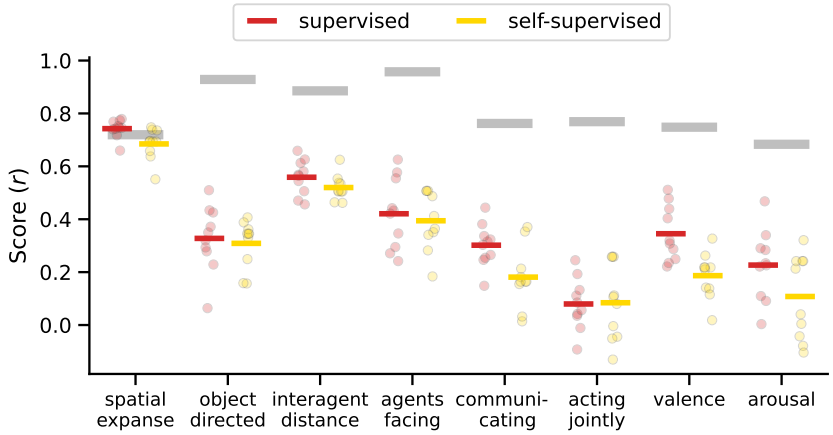

Figure 10: The performance of each image model in our set with a ResNet-50 backbone in predicting behavioral responses separated by whether the model uses a supervised (red) or self-supervised (yellow) learning objective. This analysis was restricted to a single architecture class (ResNet-50) to focus specifically on the learning objective. Plotting conventions are the same as Figure 2.

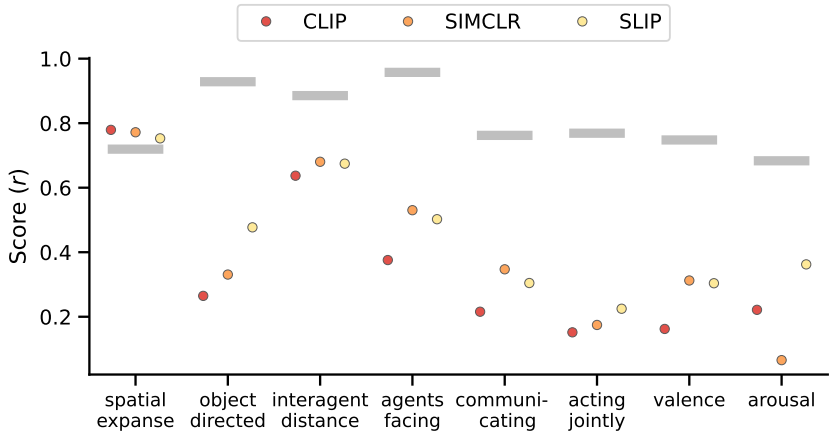

Figure 11: The performance of SLIP family models all with the same architecture (VIT-B) and training data, trained either with purely visual contrastive (SimCLR), vision-language contrastive (CLIP), or a combination (SLIP) in predicting behavioral responses. Models are grouped by CLIP (red), SimCLR (orange), and SLIP (yellow). Plotting conventions are the same as Figure 2.

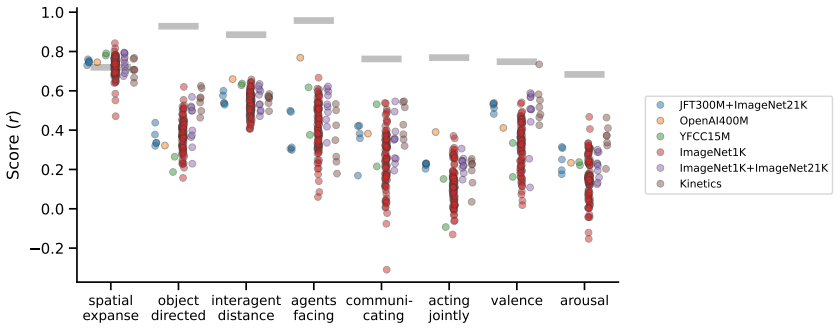

Figure 12: The performance of the vision models (image and video) at predicting the behavioral responses grouped according to their training data. Plotting conventions are the same as Figure 2.

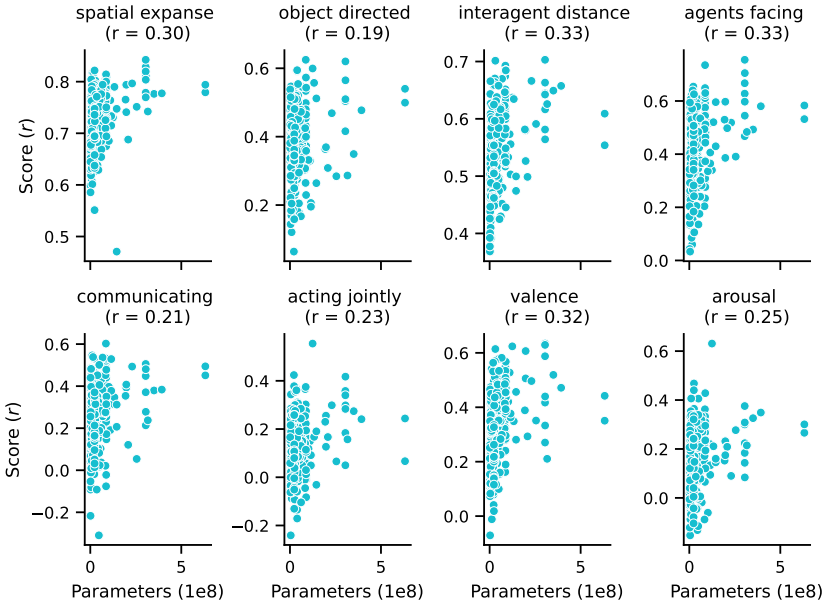

Figure 13: For each behavioral response, the model's prediction score (image and language models only) is plotted against the number of trainable parameters. The r-value below the behavioral response indicates the Pearson correlation between the score and number of trainable parameters.

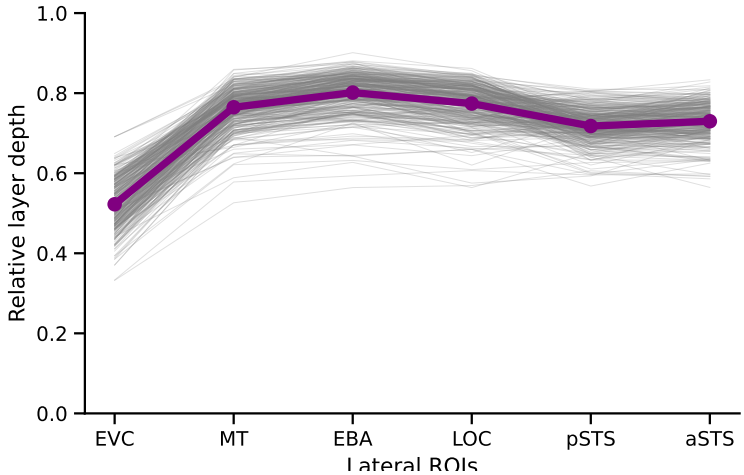

Figure 14: The relative layer depth of the best performing model layer for each image and video model (thin gray lines) and the average best layer depth across models (thick purple line) in each ROI along the lateral visual stream.

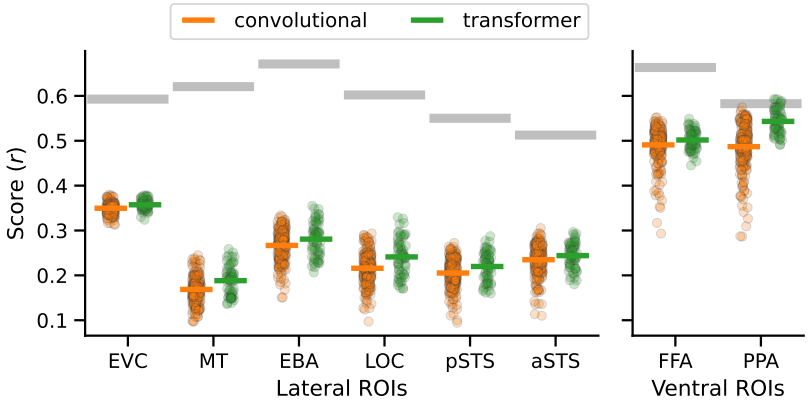

Figure 15: The performance of each image model in our set in predicting neural responses in ROIs in the lateral and ventral visual streams. Models are grouped by convolutional (orange) or transformer (green) architecture. Plotting conventions are the same as Figure 4.

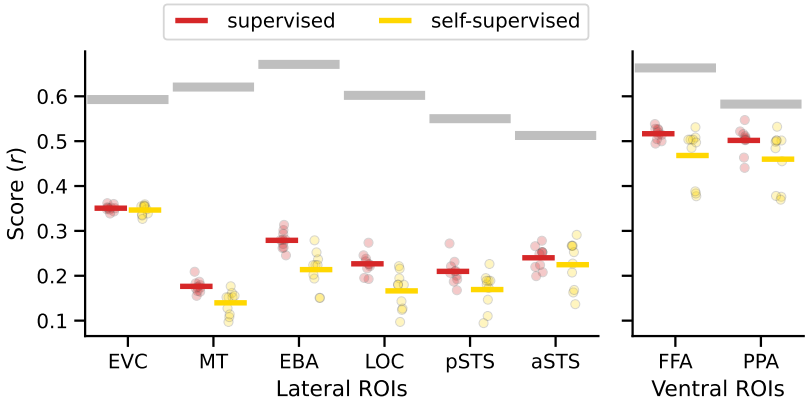

Figure 16: The performance by each image models with a ResNet-50 backbone in predicting neural responses in ROIs in the lateral and ventral visual streams. Models are grouped by supervised (red) or self-supervised (yellow) learning objective. This analysis was restricted to a single architecture class (ResNet-50) to focus specifically on the training objective. Plotting conventions are the same as Figure 4.

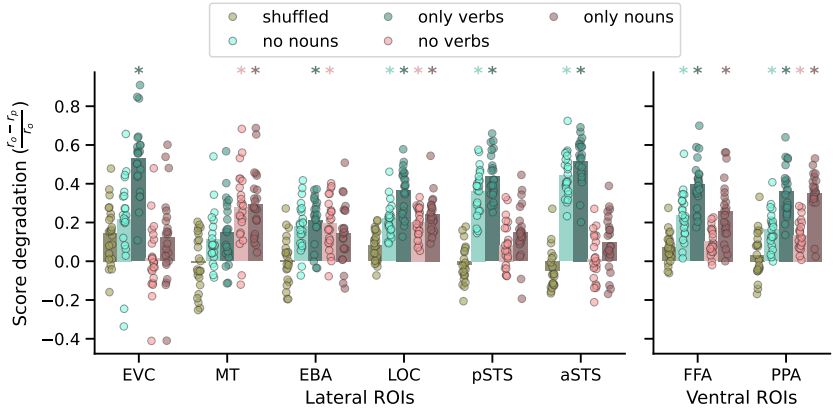

Figure 17: The average performance in ROIs of each language model (dots) in predicting neural responses following selective perturbation of the sentence captions. The bars indicate the mean performance across models for each condition and rating. Asterisks indicate that there is a significant degradation in model-neural alignment following perturbation relative to the unperturbed sentence.

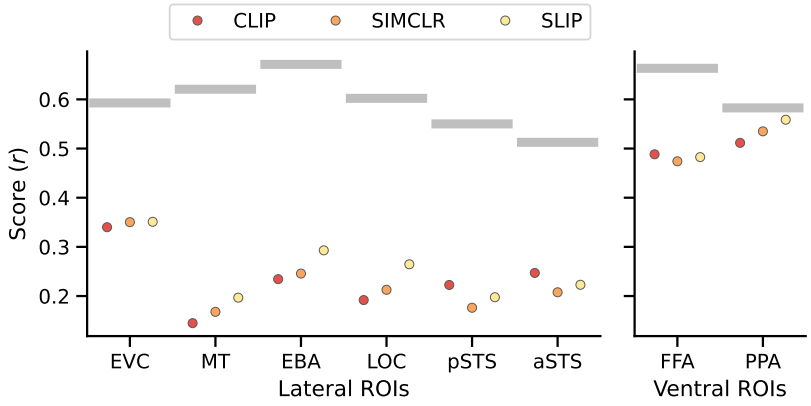

Figure 18: The performance of SLIP family models all with the same architecture (VIT-B) and training data, trained either with purely visual contrastive (SimCLR), vision-language contrastive (CLIP), or a combination (SLIP) in predicting neural responses in ROIs in the lateral and ventral visual streams. Models are grouped by CLIP (red), SimCLR (orange), and SLIP (yellow). Plotting conventions are the same as Figure 4.

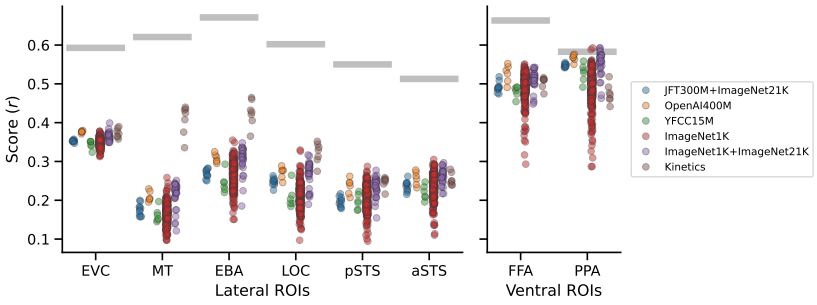

Figure 19: The performance of the vision models (image and video) at predicting neural responses grouped according to their training data. Plotting conventions are the same as Figure 4.

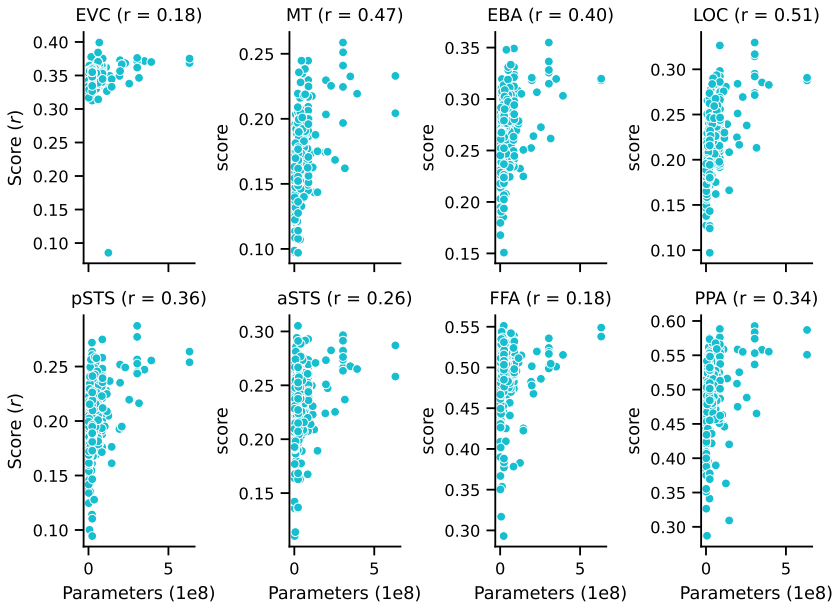

Figure 20: For each ROI response, the model's prediction score (image and language models only) is plotted against the number of trainable parameters. The r-value next to the ROI response indicates the Pearson correlation between the score and number of trainable parameters.