Incorporating natural language into vision models improves prediction and understanding of higher visual cortex

Aria Y. Wang[1,2], **Kendrick Kay**[3], **Thomas Naselaris**[3,4], **Michael J. Tarr**[1,2,5], **and Leila Wehbe**[1,2,5,*]

[1]Neuroscience Institute, Carnegie Mellon University
[2]Machine Learning Department, Carnegie Mellon University
[3]Center for Magnetic Resonance Research (CMRR), Department of Radiology, University of Minnesota
[4]Department of Neuroscience, University of Minnesota
[5]Department of Psychology, Carnegie Mellon University
[*]*corresponding author*: lwehbe@cmu.edu

ABSTRACT

We hypothesize that high-level visual representations contain more than the representation of individual categories: they represent complex semantic information inherent in scenes that is most relevant for interaction with the world. Consequently, multimodal models such as Contrastive Language-Image Pre-training (CLIP) which construct image embeddings to best match embeddings of image captions should better predict neural responses in visual cortex, since image captions typically contain the most semantically relevant information in an image for humans. We extracted image features using CLIP, which encodes visual concepts with supervision from natural language captions. We then used voxelwise encoding models based on CLIP features to predict brain responses to real-world images from the Natural Scenes Dataset. CLIP explains up to $R^2 = 78\%$ of variance in stimulus-evoked responses from individual voxels in the held out test data. CLIP also explains greater unique variance in higher-level visual areas compared to models trained only with image/label pairs (ImageNet trained ResNet) or text (BERT). Visualizations of model embeddings and Principal Component Analysis (PCA) reveal that, with the use of captions, CLIP captures both global and fine-grained semantic dimensions represented within visual cortex. Based on these novel results, we suggest that human understanding of their environment form an important dimension of visual representation.

Introduction

Recently the ability to account for neural responses associated with high-level vision has rapidly advanced due to the use of features derived from state-of-the-art neural networks[1]. Heretofore unaccounted for neural responses associated with tasks in both visual and semantic processing can now be well predicted by deep neural networks[1,2]. As suggested by Yamins and DiCarlo[3], these dramatic improvements in prediction performance may be driven by the fact that models sharing task goals with natural systems learn representations also shared with such systems. This is true not only at the highest levels of behavior, but also for specific mid-level tasks within our perceptual systems[4]. However, almost all neural networks for vision are trained on purely *visual* tasks. In contrast, human vision is an active process that, to support complex behaviors such as scene interpretation and navigation, incorporates information from diverse sources, for instance, conceptual knowledge or verbal descriptions. In this context, one impediment to unraveling the representational basis of visual pathways in the brain may be the failure to consider complex training signals that capture human-relevant information in most vision models. One way to capture this information is to learn from multiple modalities simultaneously, as the confluence of information from different sources can help determine what is important. This is especially true if one of the modalities is language, since language is behaviorally generated by humans and highlights aspects that are important. In this paper we investigate this hypothesis and demonstrate that state-of-the-art neural network models trained on more multimodal human-like task goals yield further improvements in predicting high-level visual regions, indicating that these regions represent complex, behaviorally important semantics.

Higher-level visual representations are thought to reflect the structure of the visual world and semantics beyond object identity; for example, non-perceptual associations such as function or linguistic meaning[5,6]. Similarly, in ongoing work[7], a searchlight analysis using an embedding model based on text captions for viewed images suggests that higher-level visual cortex represents semantic information related to those images. Supporting this point, the influence of language on vision is evident in the acquisition of visual categories during development where visual learning occurs concurrently with language and conceptual learning[8,9]. In the early development of object perception during infancy, the presence of language labels has been found to be crucial in the emergence of the holistic perception of objects, as well as in the differentiation of objects parts and objects themselves[10]. Similarly, there is evidence that language also plays an important role in the acquisition of semantics[9,11]. Thus, under the view that language and semantics influence the high-level organization of visual information, we propose that multimodal neural-network models incorporating visual *and* linguistic inputs will better predict neural responses to semantically complex visual inputs such as real-world scenes.

One attractive multimodal neural network for testing this prediction is a state-of-the-art model with "Contrastive Language-Image Pre-training" or "CLIP"[12]. CLIP successfully leverages supervision from natural language (image captions) for vision and from vision for language. The CLIP model, trained with real-world image/associated caption pairs, learns separate image and text encoders that encode each image/caption pair of training data with similar representations at the final layer. Different than earlier multimodal models (e.g., VisualBERT[13], LXMERT[14]), multimodal loss signals in the final layer of CLIP are propagated through all earlier layers of both the visual and language encoders, and, therefore, learning in CLIP may be more similar to human visual learning, where top-down knowledge has been found to influence even the earliest layers of the visual pathway[15,16]. Of particular interest, as a multimodal model, CLIP excels at current zero shot benchmark tests in computer vision – outperforming vision models that do not include natural language supervision. At the same time, CLIP relies on a model architecture that is similar to prior successful vision and language models used for brain activity prediction[1,2] and is trained on a similar distribution of natural images and language data. As such, brain activity prediction using CLIP representations is expected to be at least as good as earlier single-task neural-network models and, based on our characterization of high-level visual tasks, better than earlier models due to CLIP's multimodal structure.

To test this hypothesis, we extracted network representations from CLIP (using each image or its associated caption) and from several single modality task-optimized models: ImageNet[17] pretrained ResNet[18] (which we refer to as "ResNet$_I$") and BERT[19] (using the caption associated with each image). We then constructed voxelwise encoding models to explain whole brain responses arising from *viewing* natural images from Allen et al.'s Natural

Scenes Dataset (NSD)[20]. Our goal was to use this extensive brain activity dataset to evaluate and quantify the contribution of multimodal pre-training in generating more brain-like, semantically-grounded visual representations. Our results, through brain prediction, variance partitioning and representation visualization establish that CLIP is much more accurate than single modality models at predicting higher-level visual representations in the brain. Building on this result, we are able to successfully apply CLIP in conjunction with principal component analysis (PCA) to learned representations to tease apart important semantic dimensions of visual knowledge and, thus, provide insight into the fine-grained organization of knowledge in the human brain.

Results

Multimodal embeddings best predict high-level visual cortex

The central question of our current study is whether CLIP better predicts neural responses as compared to previous, vision-only models. To address this question, we extracted representations from the CLIP image encoder and used them to predict voxelwise responses (as measured by fMRI) across the brain. In Figure 1 we show the R^2 performances in the held out data set across the whole brain. For visualization purpose, in the flatmap we only plotted the voxels that are predicted significantly higher than chance ($p < 0.05$, FDR-corrected[21]). The encoding model built with the last layer of CLIP's visual encoder explains variances in voxels close to its noise ceiling (see Supplementary Fig. S1 for performance measured in r for Subject S5). As a reference, earlier papers using voxelwise encoding models for brain prediction report well below 0.7 in maximum correlation[22,23]. In Allen et al.[20] a brain optimized model of early visual cortex (V1-V4) explains up to 0.8 in R^2, similar to what we observe here in high-level visual cortex. However, directly comparing performance across wide range of models is challenging due to the fact that different studies are carried out with very distinct experimental designs and rely on different data preprocessing and fitting pipelines. Studies that report model performance in terms of averages within ROIs and representation similarity (RSA) scores are also difficult to compare to our present results. Importantly, the high level performance we observed was not idiosyncratic to a few subjects: both the overall level and the pattern of prediction performance were highly consistent across S1-S8 (results for S5 are shown in Fig. 1, results for S1-S8 are shown in Supplementary Fig. S2 and Fig. S3).

The CLIP encoder model's superior prediction performance provides compelling evidence that joint supervision from text information leads to representations that are better predictive of high-level visual cortex. We discuss this further in the *Discussion* section. From a theoretical point of view, these results suggest that the semantic information summarized in the image captions plays an important role in the organization of high-level visual knowledge in the human brain.

Beyond overall performance metrics, performance peaks in the brain prediction maps were aligned with common functionally-defined category-selective ROIs. In particular, peaks within regions implicated as scene-selective[24], body-selective[25], and face-selective[26,27] were sufficiently well defined so as to allow localization of these ROIs based solely on the prediction performance of CLIP. We speculate that these alignments signal the importance of semantic associations in scene understanding and person recognition.

In order to rule out performance improvements based on a specific network architecture, we extracted features from two available backbones for CLIP: visual transformer (ViT-32) and ResNet50. Differences in prediction performance were small (see Supplementary Fig. S6), indicating that the improvement provided by CLIP is not due to any particular neural-net architecture.

To explore whether captions associated with images could predict the brain activity in response to viewing the corresponding image, representations extracted from the CLIP text encoder were also used to predict voxelwise responses across the brain. To accomplish this we provided the CLIP text encoder with the captions of the images viewed in the scanner by each subject. The text encoder representation was then used to make voxelwise brain predictions. Somewhat surprisingly, in the absence of any image information, the model is still able to predict higher level visual cortex similar to that of the model based on CLIP's image encoder (Fig. 2), though the visual encoder still explains most of the unique variance throughout the cortex (see Supplmentary Fig. S7). This result indicates the efficacy of CLIP in capturing brain relevant visual-semantic information from the images and the captions. The

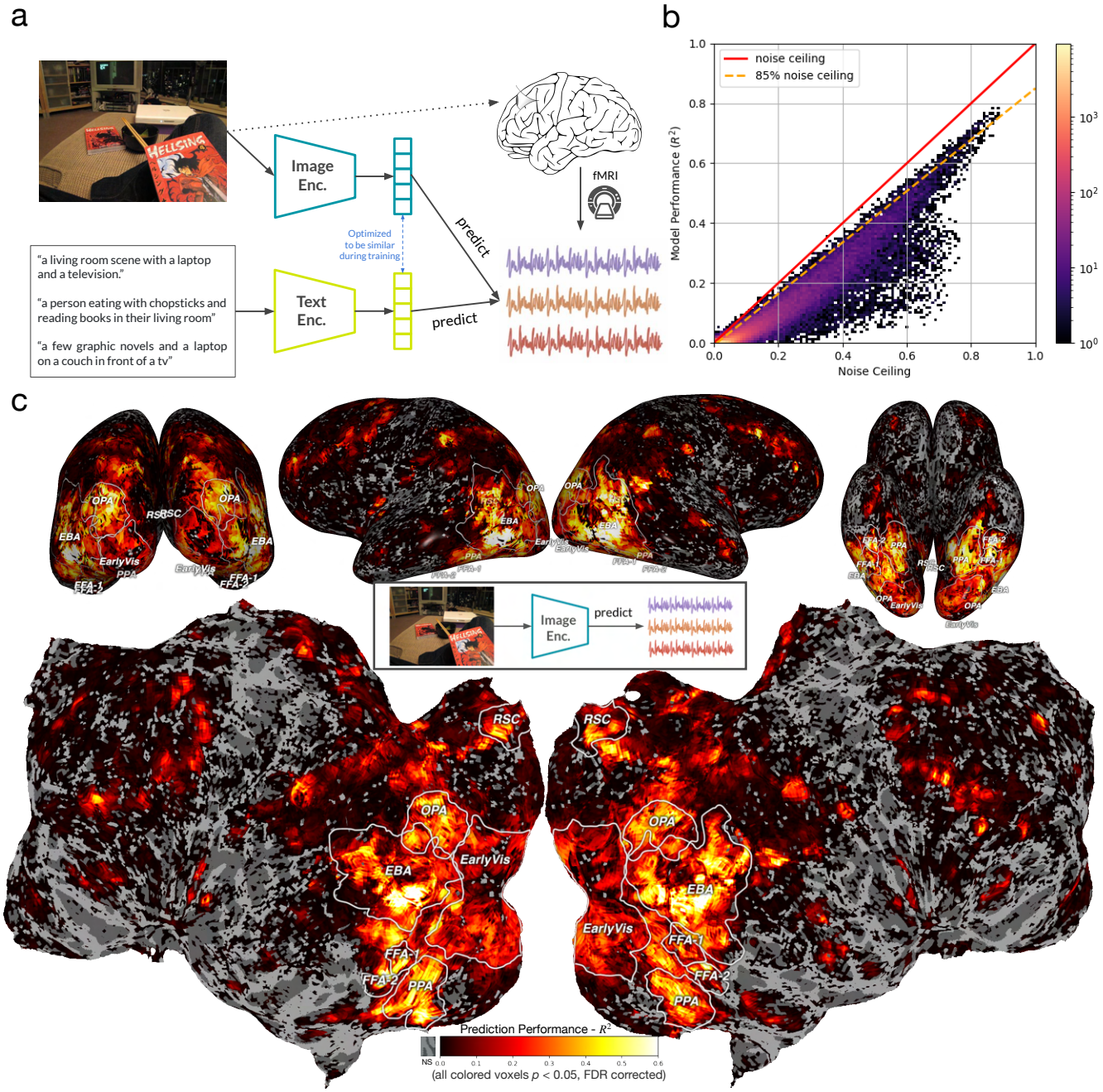


Figure 1. Model pipeline and prediction performance for the CLIP visual encoder. **(a)** Representations from the CLIP image and text encoders are extracted from images and captions, respectively. These representations are used in voxelwise encoding models to predict brain responses to each image. **(b)** A 2D histogram of model performance in R^2 against noise ceiling across all voxels in the whole brain. Density of voxels are shown in a log scale. Most voxels are predicted close to its noise ceiling. **(c)** Voxelwise prediction performance (measured in R^2) on a held-out test set is shown for Subject S5 in lateral (top-middle), posterior (top-left); bottom (top-right) views. (Bottom) The same prediction performance for S5 is shown in a flattened view of the brain. Performance improvements based on network architecture were ruled out by extracting features from two backbones: ViT-32 and ResNet50. Performance differences among architectures were small, indicating that the improvements afforded by CLIP are not due to any particular network architecture.

fact that both the image and text encoders have similar patterns of high predictive performance indicates that the information encoded in these high-level visual areas is highly anchored in semantics.

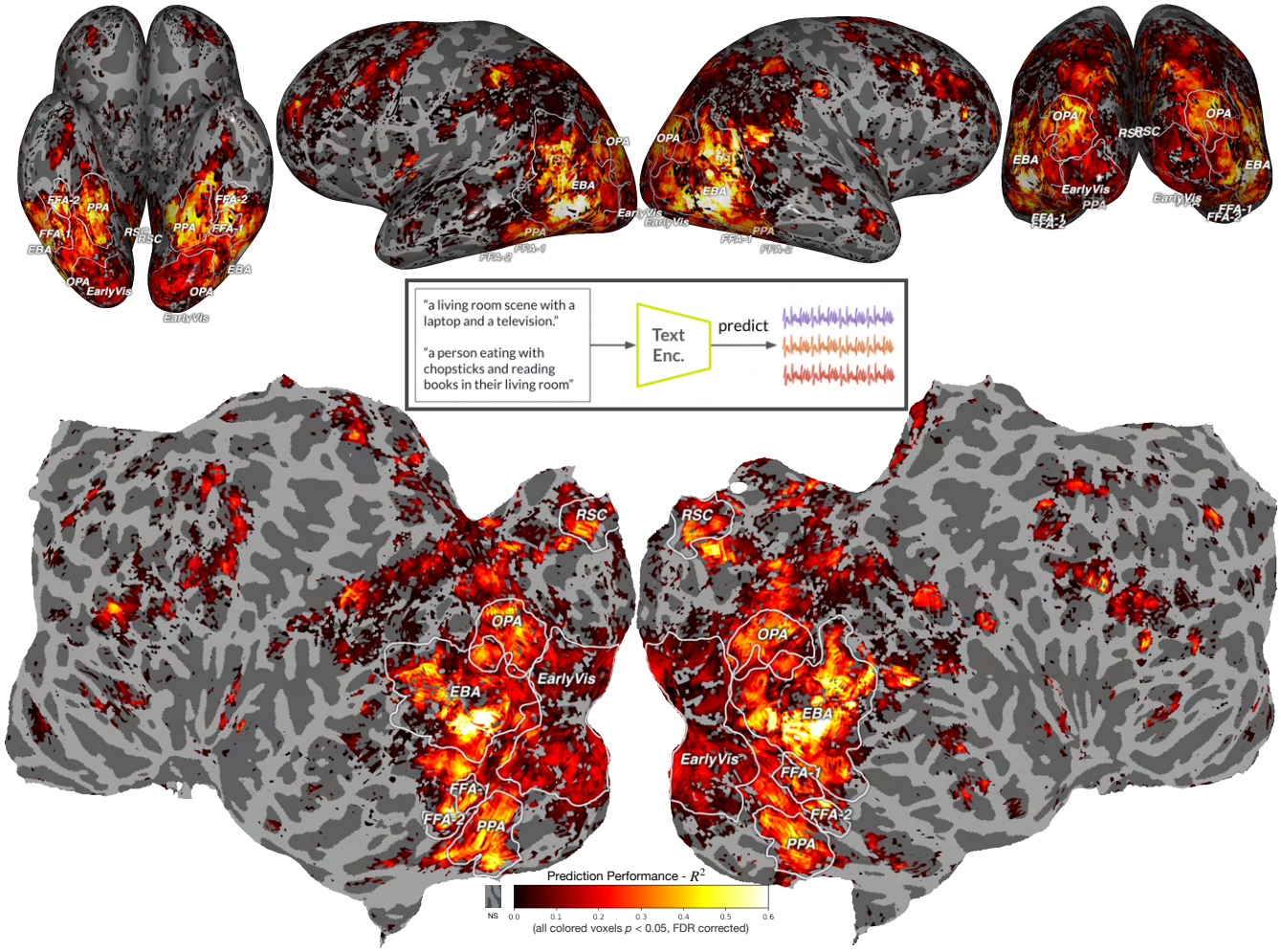


Figure 2. Prediction performance for the CLIP text encoder. Prediction performance for voxelwise responses – R^2 – in held out data for the CLIP text encoding model for S5 with overlays for functionally-defined, category-selective ROIs. Although only have access to the captions of the images that the subjects viewed, the CLIP text encoder is still able to predict fMRI data in many functionally-defined ROIs (e.g., EBA, PPA, RSC, FFA).

CLIP embeddings explain more unique variance than unimodal embeddings

As compared to the ImageNet trained ResNet50 (ResNet_I), CLIP explains more variances in individual voxels across the whole brain, as shown in 3a and Supplementary Fig. S5. In order to measure the unique variance accounted for by CLIP as compared to unimodal models, we performed a variance partitioning analysis[28,29] (Fig. 3). Only voxels with significantly higher than chance unique variance are plotted for both models ($p < 0.05$, FDR-corrected). We compared the unique variance accounted for by the last layer of the CLIP image encoder with a ResNet50 backbone to that accounted for by ResNet_I (which also has a ResNet50 architecture), ruling out potential performance differences arising from model architecture.

Consistent with our results for prediction performance, CLIP accounts for the majority of the unique variance in areas anterior to primary visual cortex, particularly in OPA, PPA and EBA – all functional ROIs implicated in scene and person perception. To evaluate ROI-level improvement we also present a series of voxel scatter plots for a range of functional ROIs in Figure 3b. With the exception of early visual areas (e.g., V1v, h4v), CLIP accounts for a much

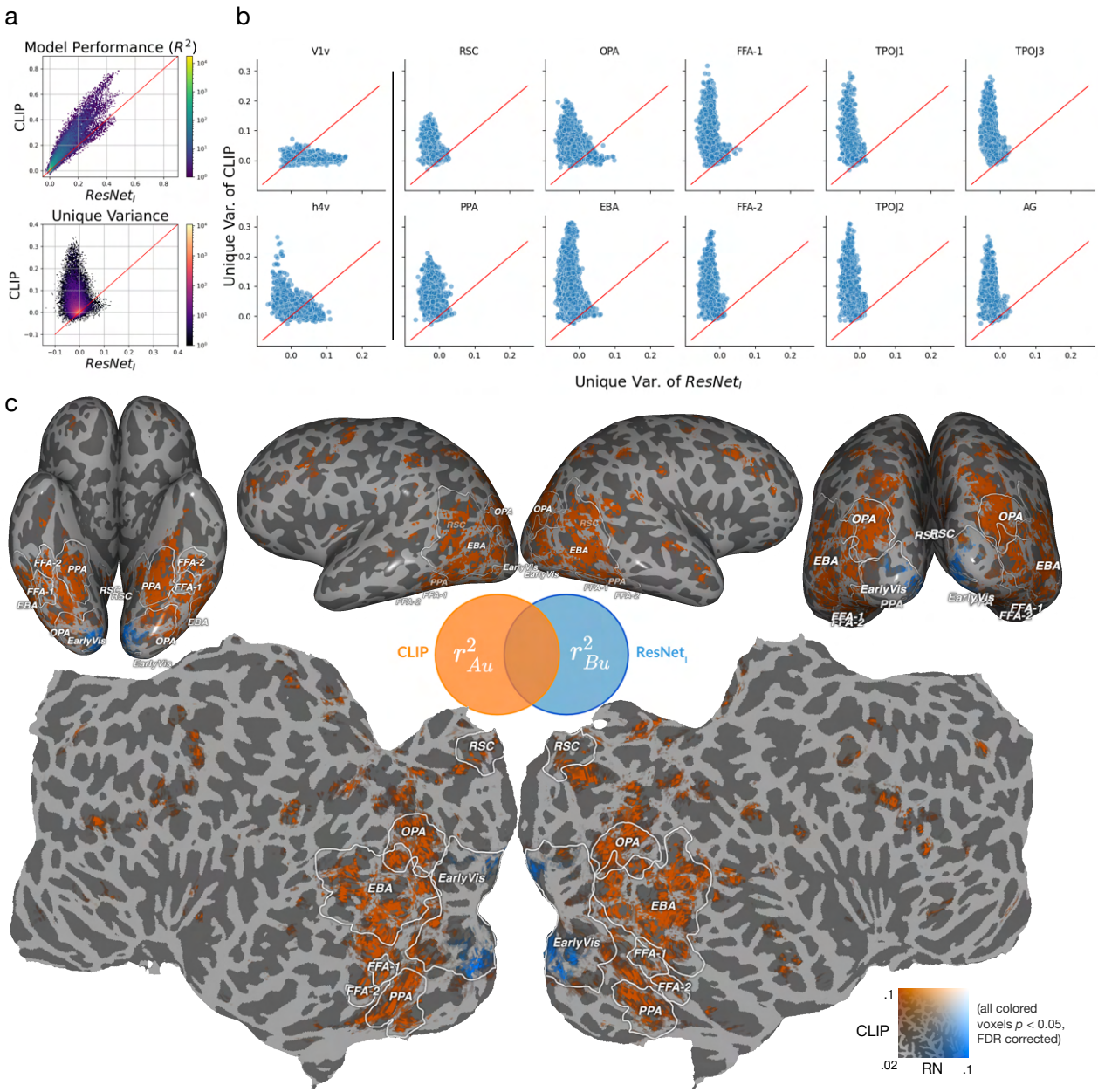


Figure 3. Performance for the CLIP visual encoder using a ResNet backbone as compared to ResNet$_I$ (a) 2D distribution plots of voxels from the whole brain in S5 in model performance (in R^2) and unique variance comparing between CLIP and ResNet$_I$. The red lines indicates equal performance for the two models. CLIP predicts much better in terms of total variance and unique variance. **(b)** Unique variance accounted for by CLIP as compared to ResNet$_I$ for 12 different ROIs for all eight subjects. Individual voxels are plotted as blue points. The red lines indicate iso-variance, that is, $(y = x)$. CLIP accounts for overwhelmingly more variance than ResNet$_I$ in higher-level visual cortex. In contrast, ResNet$_I$ only accounts for more variance in ventral V1 and a reasonable proportion of the variance in ventral V4. **(c)** Unique variance accounted for by CLIP as compared to ResNet$_I$ for S5 – obtained by subtracting R^2 for each model from that of the joint model (with concatenated feature spaces). Voxels where CLIP accounts for greater unique variance are orange and voxels where ResNet$_I$ accounts for greater unique variance are blue.

larger portion of the unique variance for the majority of voxels in these high-level ROIs. Beyond category-selective ROIs that respond to faces, places, and bodies, we also identified ROIs such as TPOJ and Angular Gyrus (AG) that were much better explained by CLIP. Interestingly, these two areas are held to be related to theory of mind and language[30].

Note that the last layer of CLIP explained less of the variance in early visual cortex as compared to ResNet$_I$; however, this does not imply that CLIP fails to capture information represented in these regions. The last layer of CLIP is the bottleneck layer that captures the image embeddings optimized to match in similarity with the text embeddings. As shown in Supplementary Figure S8, the entire visual pathway is best predicted by a progression of CLIP layers (including ones below the bottleneck layer). More generally, CLIP is the best predictive model for the whole of visual cortex.

A variance partitioning analysis comparing CLIP embeddings constructed from image captions to BERT embeddings of those same captions likewise found that CLIP accounts for almost all unique variance (Supplementary Fig. S9). Thus, the advances we observed in brain prediction using CLIP do not appear to arise from incorporating complex semantics alone, but rather, can be attributed to a meaningful mapping between visual and semantic representations.

CLIP embeddings capture contextual and semantic similarities in the absence of visual similarity

To further explore why CLIP outperforms unimodal models, we compared the similarities of the CLIP and ResNet$_I$ representations for 1000 randomly selected stimulus images. After obtaining the distances between each pair of images in both representations, we ranked each pair according to the differences between the similarities (measured in correlation). Namely, $S_{i,j} = Sim_{i,j}^{CLIP} - Sim_{i,j}^{ResNet_I}, \forall i,j \in \{1,...,1000\}$, where $Sim_{i,j}^{CLIP}$ and $Sim_{i,j}^{ResNet_I}$ are correlations of representations between Image i and Image j in CLIP and ResNet$_I$, respectively. Figure 4 plots the images that are most similar in CLIP and dissimilar in ResNet$_I$ (ranked by S_{ij}) and vice versa. These visualizations illustrate that with natural language as training feedback, representations within CLIP capture contextual similarities that are not present in ResNet$_I$ (which seems much more anchored in visual similarity). To the extent that higher-level visual representations in the brain reflect both semantics and visual appearance, this visualization of representation space helps to explain why CLIP is better than earlier models at predicting neural responses to complex, real-world scenes.

The principal semantic dimensions of the CLIP encoding model capture core axes of brain organization

To better understand the semantic dimensions learned in the encoding model, we performed principal component analysis (PCA) on the learned weight matrix concatenated across the 20000 top predicted voxels of all eight subjects. We projected the voxels onto the principal component (PC) dimensions to understand the tuning of the entire voxel space, following previous work[23,32]. By visualizing each PC of the learned model and its corresponding voxel projection, we were able to uncover some of the semantic bases that underlie semantic organization in the brain. To capture the information captured by different PCs, we visualized the images that correspond to the highest magnitude along a given PC (Fig. 5). These images were identified by computing the dot product of CLIP image embeddings with the vector corresponding to the PC direction. As illustrated in the middle row of Figure 5 and in Figure 6d, we observed that animate and inanimate images are separated by PC1; its brain projections correspond to functionally-defined body and face regions (e.g., FFA and EBA). As illustrated in the bottom row of Figure 5, we observed that scenes and food images are separated by PC2 when we split the functional areas identified from PC1 with PC2; its brain projections corresponded to functionally-defined place regions (e.g., PPA, RSC, OPA) and the food region[31,33,34]. Of note, we obtained interpretable PC dimensions up to PC10 (despite the relatively low explained variance from PC6 onwards), allowing us to identify more fined-grained semantic distinctions within high-level visual cortex. Images visualization of the rest of the PCs are shown in supplementary Fig. S11.

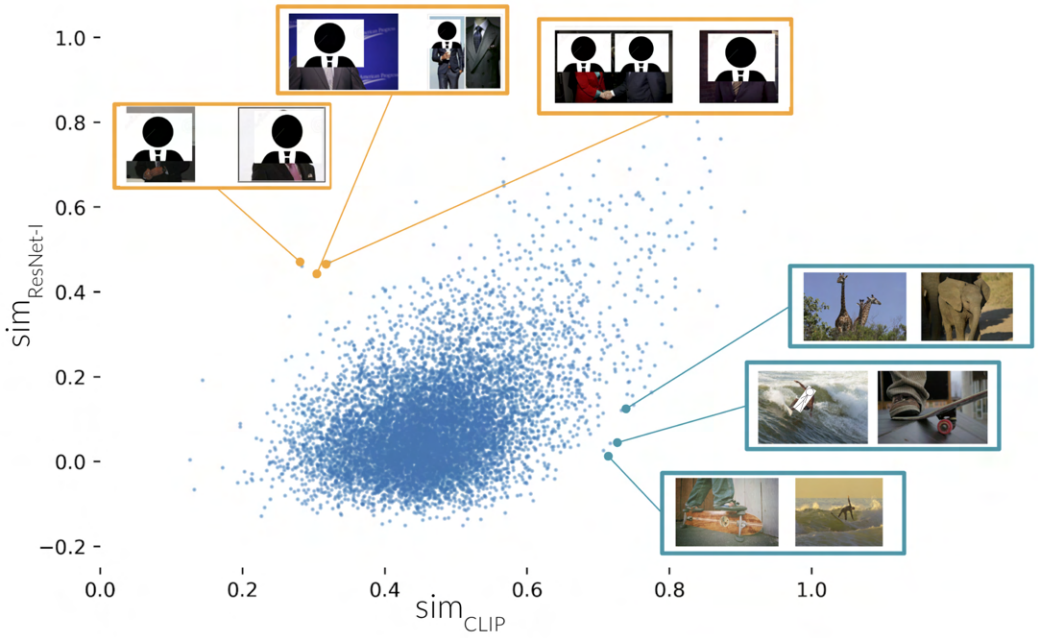


Figure 4. Pairwise representation similarity emphasizes semantically similar images using CLIP and visually similar images using ResNet$_I$. Images represented similarly in CLIP, but not ResNet$_I$, are semantically related. For example, within CLIP, images of people surfing and skateboarding are grouped together and images of giraffes and an elephant are grouped together. In contrast, within ResNet$_I$, images with different contexts are grouped according to visual similarity. For example, people wearing dark suits with a white shirt and contrasting tie.

Regions that benefit most from CLIP embeddings encode scenes of humans interacting with their environment

We directly compared the brain projection for PC1 and the unique variance map for CLIP. We found that voxels that have large negative values on the PC1 overlay the majority of the time with voxels where CLIP has the largest unique variance (Figs. 6a and 6b). These voxels clustered in ventral EBA, FFA-1, FFA-2, as well as ventral RSC. Figure 6c further validates this finding by showing a strong negative correlation for the voxels with a negative projection between the magnitude of this projection and the unique variance explained by CLIP. Note that the sign of the PC is arbitrary and can be flipped; we use "negative" here to refer to one of the sides of PC1.) Thus, PC1 appears to separate the regions of high-level visual cortex that benefit the most when CLIP is used to predict performance.

Figure 6d shows the top 10 images for both ends of PC1. Top negative images are people participating in sports, whereas the top positive images are indoor scenes. This separation is consistent with the location of the best predicted voxels from CLIP being centered on the EBA. Category distributions of images that are on the two sides of the PC1 further validate this finding. We leveraged the known category and super-category labels of images in COCO and found that images that lie on the negative end of the PC1 are more likely to contain people, animals, and sports items. These observations suggest that the representation of people in CLIP is the domain for which the model provided the most leverage in terms of predicting brain responses (i.e., as compared to ResNet$_I$). From an ecological standpoint this finding appears to capture high-level semantic statistics regarding the world around us: scenes of people and human interactions are heavily present in our daily life. Returning to our original hypothesis, by including natural language as input (image captions) along with complex scenes, CLIP is more effective at capturing

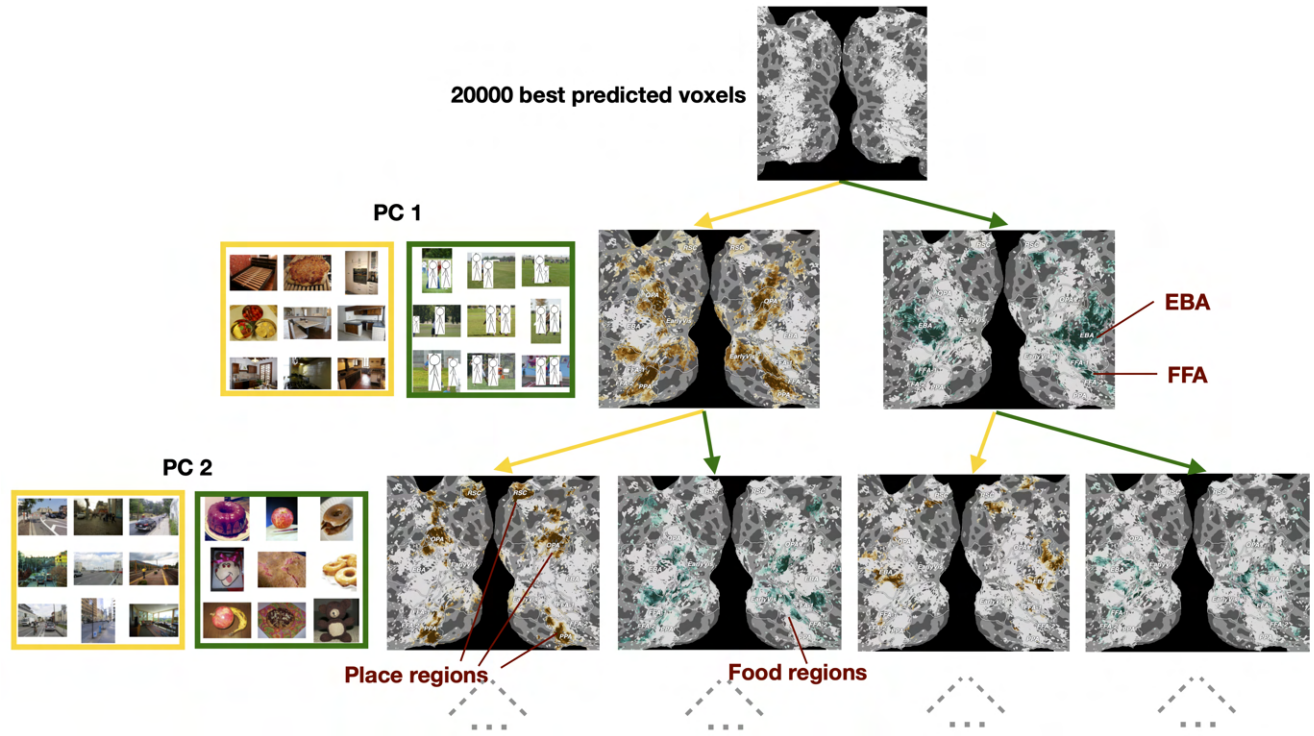


Figure 5. Cortical semantic organization as revealed by the principal components of the CLIP encoding model. Brain regions well predicted by CLIP can be hierarchically decomposed using the model PCs. PC1 separates animacy regions (EBA and FFA) from other regions, which are themselves separated by PC2 into place and food regions[31]. The rest of the tree is not shown due to space constraints.

the rich semantics of scenes as compared to models trained with image/label pairs pre-training (e.g., ImageNet). As such, we hold that CLIP is a better candidate model for understanding representation in high-level visual cortex, especially for regions such as EBA, FFA and the cingulate cortex.

Discussion

We evaluated and quantified the contribution of multimodal pre-training in generating more human-like, semantically-grounded representations. We found that the multimodal CLIP model is extraordinarily good at predicting voxelwise neural responses to viewing scenes in the Natural Scenes Dataset[20]. A related studies confirm our findings, reporting that CLIP excels at predicting responses in NSD as compared to 85 other deep network models[35]. However, our work provides a range of analyses that are an advance over that study. First, we applied variance partitioning to localize where in the brain we observed the most benefit from CLIP predictions. We found that neural responses in high-level visual cortex are exceptionally well predicted by CLIP. Second, we visualized the representation from CLIP and the unimodal network. We showed that CLIP captures more semantic based representations, which corroborates our hypothesis about complex semantic representations in high-level visual cortex. Third, we used PCA analyses to reveal that, the more fine-grained representation of scenes depicting human interaction in CLIP gain the most leverage in brain prediction, which in turn elucidated some of the underlying reasons why CLIP yields such excellent performance. This analysis suggests that CLIP captures information about humans interacting with the world, and that this information is predictive of these regions.

It is hard to definitively state that natural language feedback is what makes CLIP excel at few-shot tasks or its superior brain prediction. In fact this question is still being debated within the field of computer vision. Alternatively,

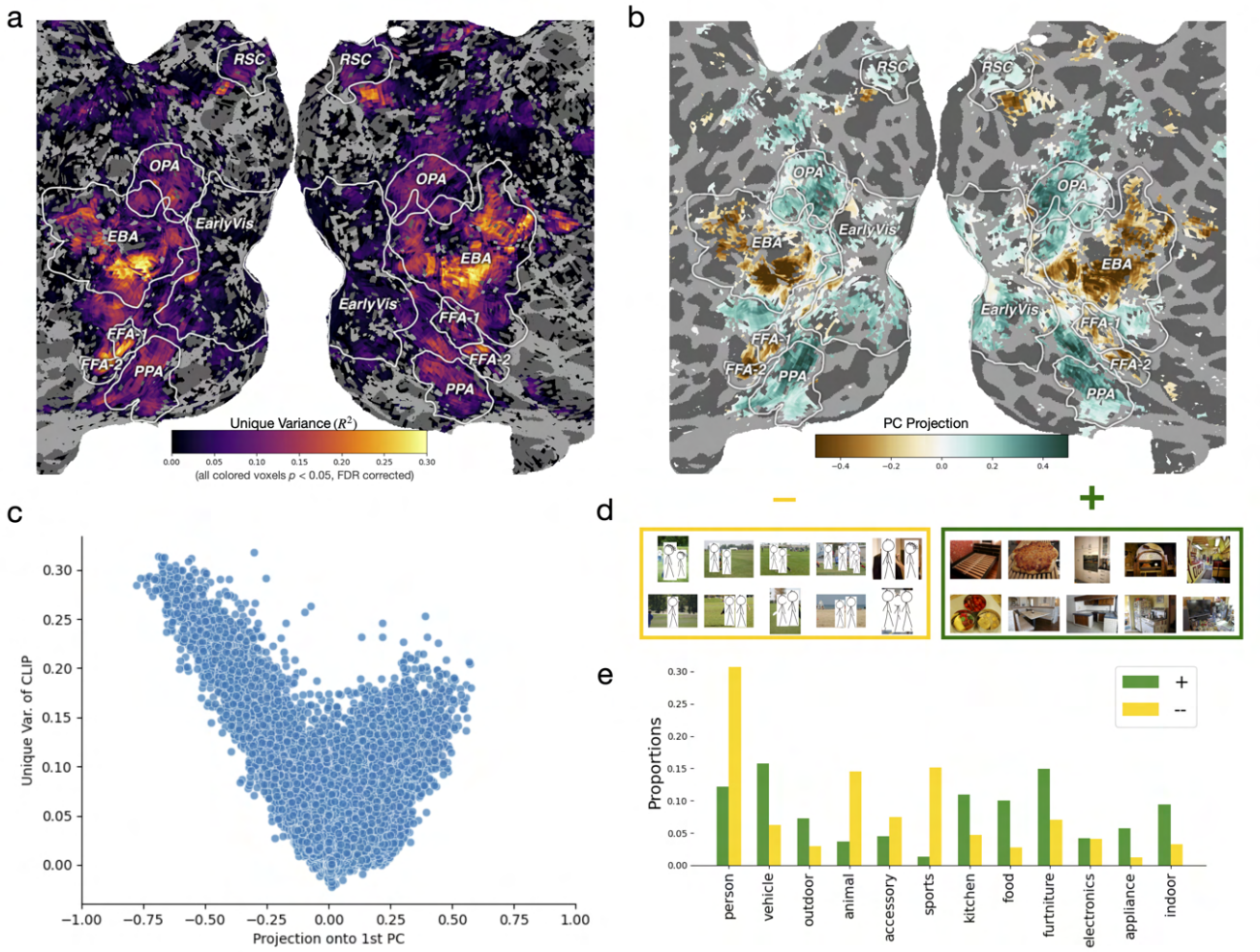


Figure 6. Better representations of scenes with people in CLIP can account for gains in unique variance. **(a)** Unique variance explained by CLIP plot on flatmap from S5. **(b)** Projection of voxels onto PC1 of the learned CLIP model for S5. Voxels that are best explained by CLIP overlap largely with the voxels that lie on positive side when projected onto the 1st PC. **(c)** Voxelwise scatter plot illustrating that for voxels lying on the negative side of 1st PC projection, the further down the voxel lies on the projection, the better it is explained by CLIP. **(d)** Images are grouped in to "+" and "−" depending on which side the image lies on when projected onto the PC1. The top 10 images that best align with either end of the PC1 are shown in the yellow and green boxes respectively. For the positive projection we observe images of indoor scenes, whereas for the negative projection we observe images of people participating in outdoor sports. **(e)** Category distribution of two groups of images validates that images on the negative side consist more of people, animal, and sports, relative to images on the positive side.

one could attribute CLIP's superior performance to the extremely large size of the training dataset. However, we are skeptical that training set size is the main contributor to the high level of brain prediction we obtain with CLIP. In particular, we conjecture that even if we were able to re-train ResNet50 on a 400x bigger dataset, but still included only category labels, the resultant model would be unlikely to learn fine-grained representations of human centered scenes. Such information is simply not carried by category labels. Therefore, we hypothesize that the natural language feedback CLIP receives is crucial to its excellent performance, at least in the case of brain prediction as presented in our work here.

Building on the findings detailed above, PCA on the learned CLIP encoding model allowed us to tease apart important semantic dimensions and gain further insight into how fine-grained visual knowledge is represented within

visual cortex. Taken together, our results suggest that both better overall prediction performance and the ability to capture fine-grained dimensions in representation are rooted in the natural language feedback that CLIP receives as part of training. As such, earlier approaches to brain prediction that focused on object or scene representation in and of themselves, failed to capture a fundamental dimension of visual representation – that of complex scenes in which humans and other actors are interacting with one another and the world around them.

CLIP's ability to predict neural responses opens up new possibilities for developing a better understanding of cortical functional architecture. Our results provide evidence, consistent with the developmental literature discussed earlier, that the organization of knowledge within visual cortex may be best characterized as multimodal. Exploring this idea further will require new ways of thinking about visual cortex. As such, we suggest that any future large-scale studies of visual representation should incorporate stimuli, representations, and models that reflect such complexity.

Methods

Datasets

fMRI data. Neural data were obtained from the the Natural Scenes Dataset (NSD)[20], an open dataset of 7T whole brain high-resolution fMRI responses from eight subjects (S1-S8) who each viewed ∼10,000 unique images of natural scenes, each image repeated 3 times. These scene images were a subset of the images in the annotated Microsoft Common Objects in Context (COCO) dataset[36]. Of the 70,566 total images presented across subjects, ∼1,000 images were viewed by all subjects. fMRI data were collected during 30-40 scan sessions. Stimulus images were square cropped, presented for 3 s at a size of $8.4^\circ \times 8.4^\circ$ with 1 s gaps in between image presentations. Subjects were instructed to fixate on a central point and to press a button after each image if they had seen that image previously.

The functional MRI data were acquired at 7T using whole-brain gradient-echo EPI at 1.8-mm resolution and 1.6-s repetition time. Preprocessing steps included a temporal interpolation (correcting for slice time differences) and a spatial interpolation (correcting for head motion). Single-trial beta weights were estimated with a general linear model. In this paper we used the $betas_fithrf_GLMdenoise_RR$ preparation of the betas. FreeSurfer[37,38] was used to generate cortical surface reconstructions to which the beta weights were mapped. The beta weights were z-scored across run and were averaged across repetitions of the image (up to 3 repetitions of each image), resulting in one averaged fMRI response to each image per voxel, in each subject. NSD also includes several visual ROIs that were identified using separate functional localization experiments. We drew the boundaries of those ROIs for each subject on their native surface for better visualization and interpretation of the results (e.g., Fig. 1). All brain visualizations were produced using Pycortex software[39].

Natural scene images. All stimulus images used in NSD and in our experiments were drawn from the COCO dataset[36]. COCO is unique among large-scale image datasets in that COCO images contain contextual relationships and non-iconic (or non-canonical) object views. In comparison to ImageNet[17], COCO contains fewer labeled categories (91), but includes more examples for each category ($> 5,000$ for 82 of the categories). Note, however, that many labeled categories in ImageNet are at the subordinate level – COCO likely contains at least as many *unlabeled* subordinate categories. The complete set of COCO images and additional details can be found on the COCO website: `https://cocodataset.org`.

Feature extraction for models

Each NSD stimuli images are input to the both the standard ImageNet pretrained ResNet 50 and CLIP model. For CLIP model we use pretrained CLIP model released by OpenAI for both ResNet50 and ViT-32 transformer backbone. Model activations across layers in both architectures are used in the voxelwise encoding models. For image captions, we use the human generated captions for each of the NSD images provided by the COCO dataset and input them into both BERT and CLIP text encoder for their layerwise activations. On average, COCO provides 5-6 captions for each image. Caption embeddings for a image are extracted individually and the average is used in the encoding models.

Voxelwise encoding models

We build ridge regression model (implemented in PyTorch; see koushik2017torchgel) to predict one averaged fMRI response to each image per voxel, in each subject. We chose to use a ridge regression model instead of more complicated models in order to retain the interpretability of model weights, which may provide insights into the underlying dimensions of the brain responses. We randomly split the total number of images a subject sees into training and test set with a 4-to-1 ratio. For each subject, each voxel's regularization parameter was chosen independently via 7-fold cross-validation across the training set. Model performance was evaluated on the test data using both Pearson's correlation and coefficient of determination (R^2). To determine the significance of the predictions, we perform a bootstrap test where we resample the test set with replacement for 2000 times and compute the FDR corrected p-values threshold for various performance statistics.

Variance Partitioning

To obtain unique variance by two model A and B, we first create joint model of A and B by concatenate features from these two models. We then fit voxelwise ridge regression model to the joint model and obtain R^2_{AB}. The variance explained by individual model A and B are R^2_A and R^2_B, respectively. We calculate the unique variance for model A and B, where $R^2_A = R^2_{AB} - R^2_B$, $R^2_B = R^2_{AB} - R^2_A$.

PCA analysis

We performed principal component analysis (PCA) on the learned matrix to recover the semantic basis of the learned model. We select the 20000 best predicted voxels after noise correction and concatenate weight matrices corresponding to these voxels from all eight subjects along the voxel dimension. We then apply PCA on this weight matrix and obtain the first 20 PCs. Explained variance by these PCs are plotted in supplementary figure S10.

References

1. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci U S A* **111**, 8619–8624 (2014).
2. Toneva, M., Mitchell, T. M. & Wehbe, L. Combining computational controls with natural text reveals new aspects of meaning composition. *bioRxiv* (2020).
3. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* **19**, 356–365 (2016).
4. Wang, A., Tarr, M. & Wehbe, L. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 32 (Curran Associates, Inc., 2019).
5. Gauthier, I., James, T., Curby, K. & Tarr, M. The influence of conceptual knowledge on visual discrimination. *Cogn. Neuropsychol.* **20** (2003).
6. Maier, M. & Abdel Rahman, R. No matter how: Top-down effects of verbal and semantic category knowledge on early visual perception. *Cogn. Affect. & Behav. Neurosci.* **19**, 859–876 (2019).
7. Charest, I., Allen, E., Wu, Y., Naselaris, T. & Kay, K. Precise identification of semantic representations in the human brain. *J. Vis.* **20**, 539–539 (2020).
8. Lupyan, G., Rakison, D. H. & McClelland, J. L. Language is not just for talking: redundant labels facilitate learning of novel categories. *Psychol. Sci.* **18**, 1077–1083 (2007).
9. Waxman, S. R. & Markow, D. B. Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cogn. Psychol.* **29**, 257–302 (1995).
10. Shusterman, A. & Spelke, E. Language and the development of spatial reasoning. *The innate mind: Struct. contents* 89–106 (2005).

11. Nappa, R., Wessel, A., McEldoon, K. L., Gleitman, L. R. & Trueswell, J. C. Use of Speaker's Gaze and Syntax in Verb Learning. *Lang. Learn. Dev.* **5**, 203–234 (2009).

12. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763 (PMLR, 2021).

13. Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J. & Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

14. Tan, H. & Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

15. Murray, S. O., Boyaci, H. & Kersten, D. The representation of perceived angular size in human primary visual cortex. *Nat Neurosci* **9**, 429–434 (2006).

16. Gilbert, C. D. & Li, W. Top-down influences on visual processing. *Nat. Rev. Neurosci.* **14**, 350–363 (2013).

17. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (IEEE, 2009).

18. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

19. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

20. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).

21. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

22. Güçlü, U. & van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).

23. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).

24. Epstein, R. A. & Baker, C. I. Scene perception in the human brain. *Annu. Rev. Vis. Sci.* **5**, 373–397 (2019).

25. Downing, P. E., Jiang, Y., Shuman, M. & Kanwisher, N. A cortical area selective for visual processing of the human body. *Science* **293**, 2470–2473 (2001).

26. Sergent, J., Ohta, S. & MacDonald, B. Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain* **115**, 15–36 (1992).

27. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci* **17**, 4302–4311 (1997).

28. Lescroart, M. D., Stansbury, D. E. & Gallant, J. L. Fourier power, subjective distance, and object categories all provide plausible models of bold responses in scene-selective visual areas. *Front. computational neuroscience* **9**, 135 (2015).

29. de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L. & Theunissen, F. E. The hierarchical cortical organization of human speech processing. *J. Neurosci.* **37**, 6539–6557 (2017).

30. Saxe, R. & Kanwisher, N. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". In *Social neuroscience*, 171–182 (Psychology Press, 2013).

31. Jain, N. *et al.* Food for thought: selectivity for food in human ventral visual cortex. *bioRxiv* (2022).

32. Çukur, T., Nishimoto, S., Huth, A. G. & Gallant, J. L. Attention during natural vision warps semantic representation across the human brain. *Nat. neuroscience* **16**, 763–770 (2013).

33. Pennock, I. M. L. *et al.* Color-biased regions in the ventral visual pathway are food-selective. *bioRxiv* (2022).

34. Khosla, M., Apurva Ratan Murty, N. & Kanwisher, N. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Curr. Biol.* **32**, 1–13 (2022).

35. Conwell, C., Prince, J. S., Alvarez, G. A. & Konkle, T. Large-scale benchmarking of diverse artificial vision models in prediction of 7t human neuroimaging data. *bioRxiv* (2022).

36. Lin, T. Y. *et al.* Microsoft COCO: Common objects in context. *Lect. Notes Comput. Sci. (including subseries Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **8693 LNCS**, 740–755 (2014).

37. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage* **9**, 179–194 (1999).

38. Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. *NeuroImage* **9**, 195–207 (1999).

39. Gao, J. S., Huth, A. G., Lescroart, M. D. & Gallant, J. L. Pycortex: an interactive surface visualizer for fmri. *Front. Neuroinformatics* **9** (2015).

Acknowledgements

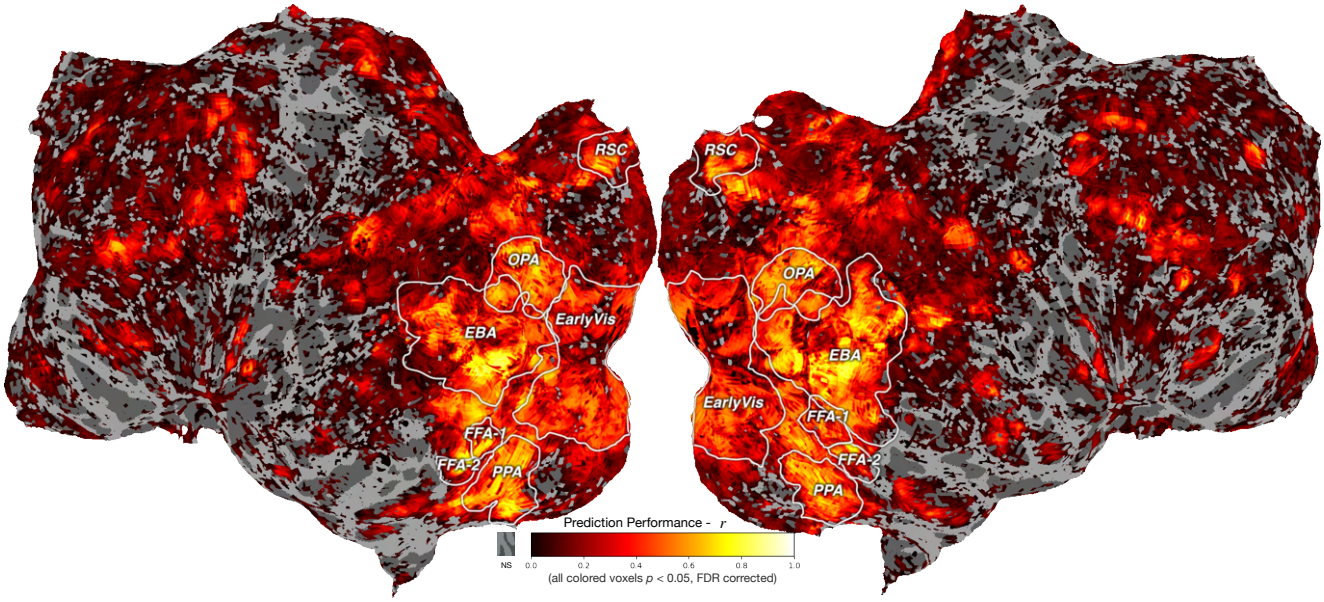
AYW and MJT were supported by AFRL/AFOSR award FA9550-18-1-0251. The NSD was supported by NSF IIS-1822683 and NSF IIS-1822929. We thank the following people for contributing technical assistance, ideas and commentary to this project: Jayanth Koushik, Nadine Chang, and Maggie Henderson.

Author contributions statement

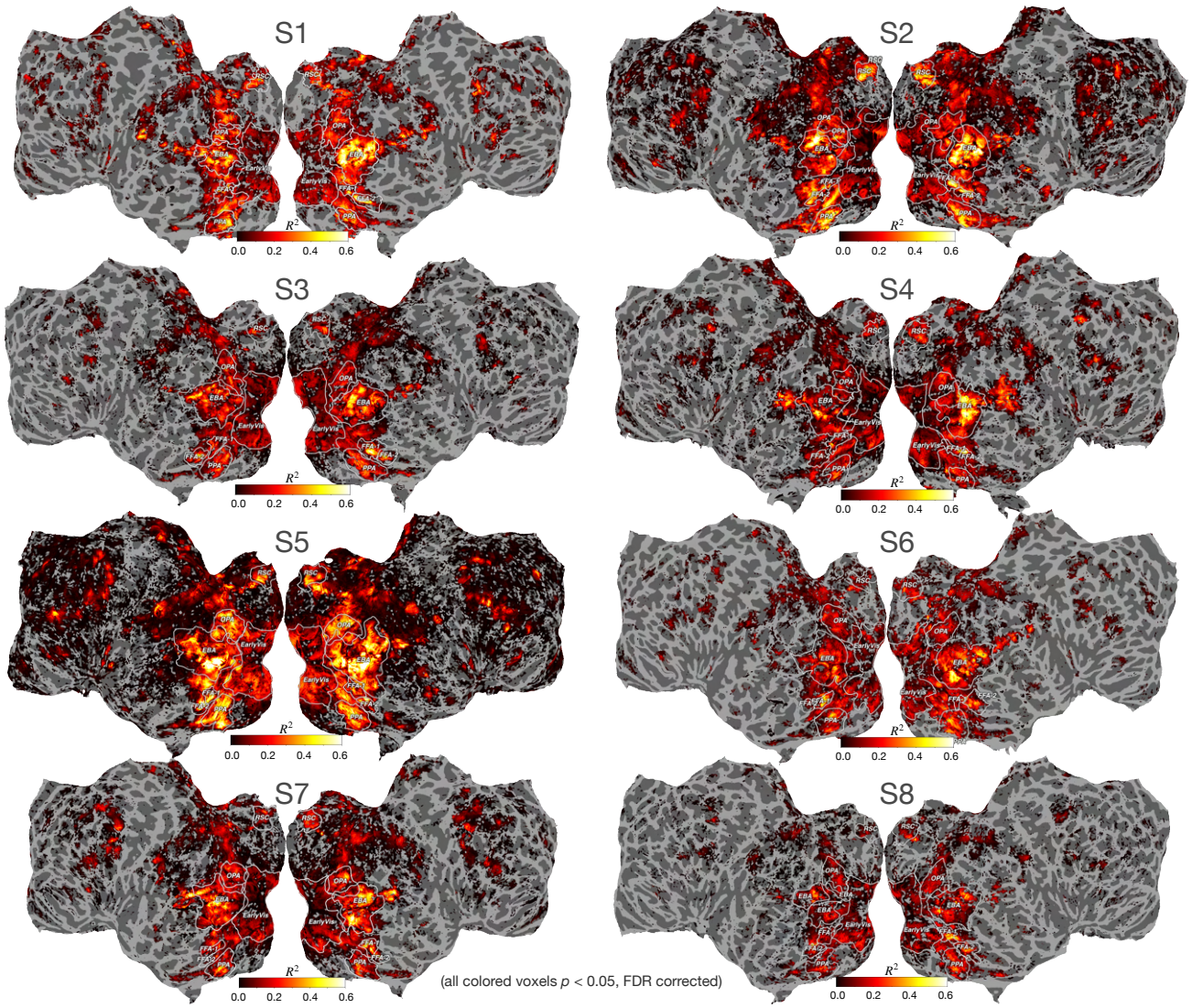
AYW, MJT, and LW conceived the experiments, KK and TN collected the neuroimaging data, AYW conducted the experiments, AYW analysed the results. All authors reviewed the manuscript.

Author Declaration

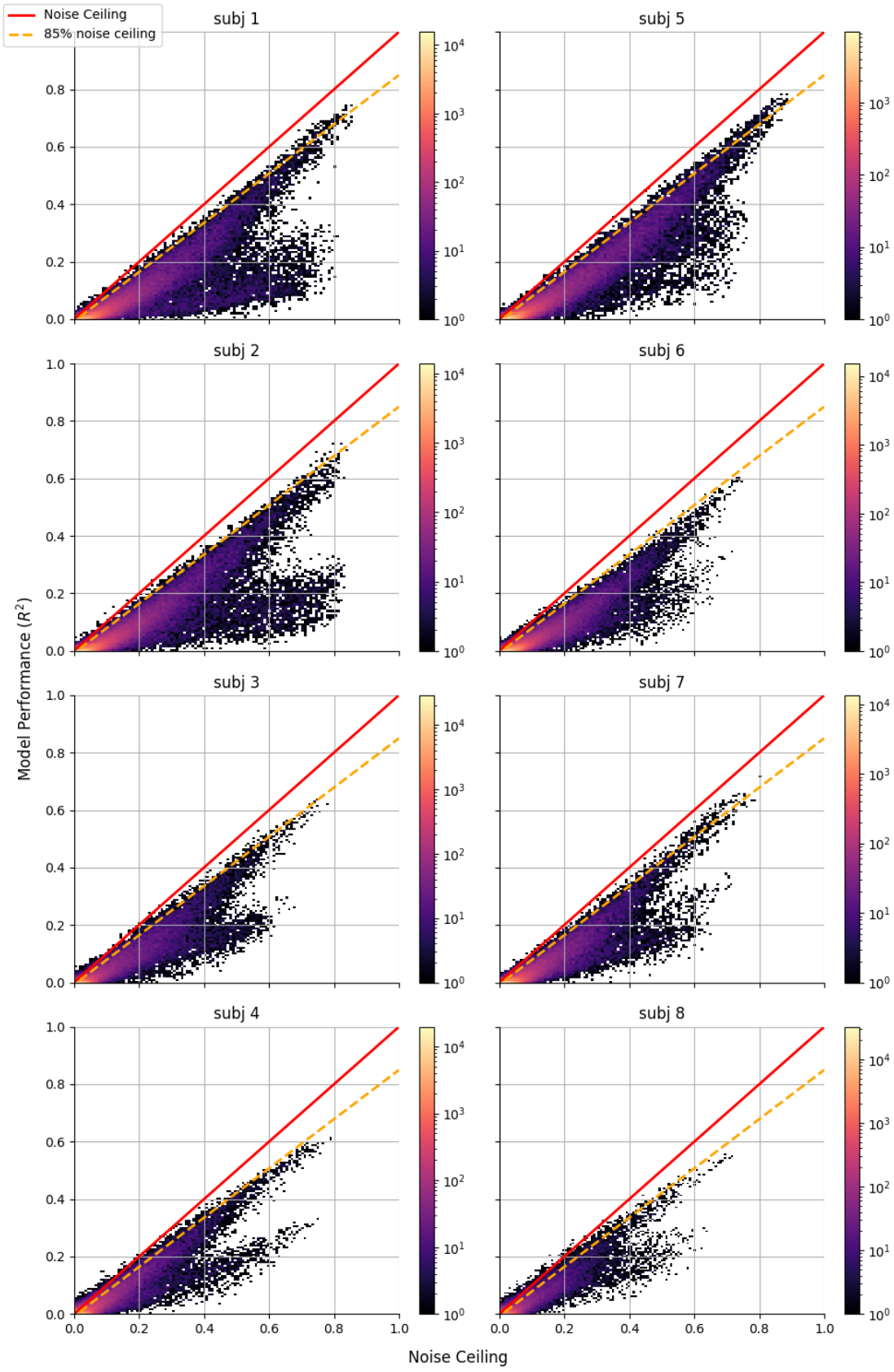
The authors declare no competing interests.



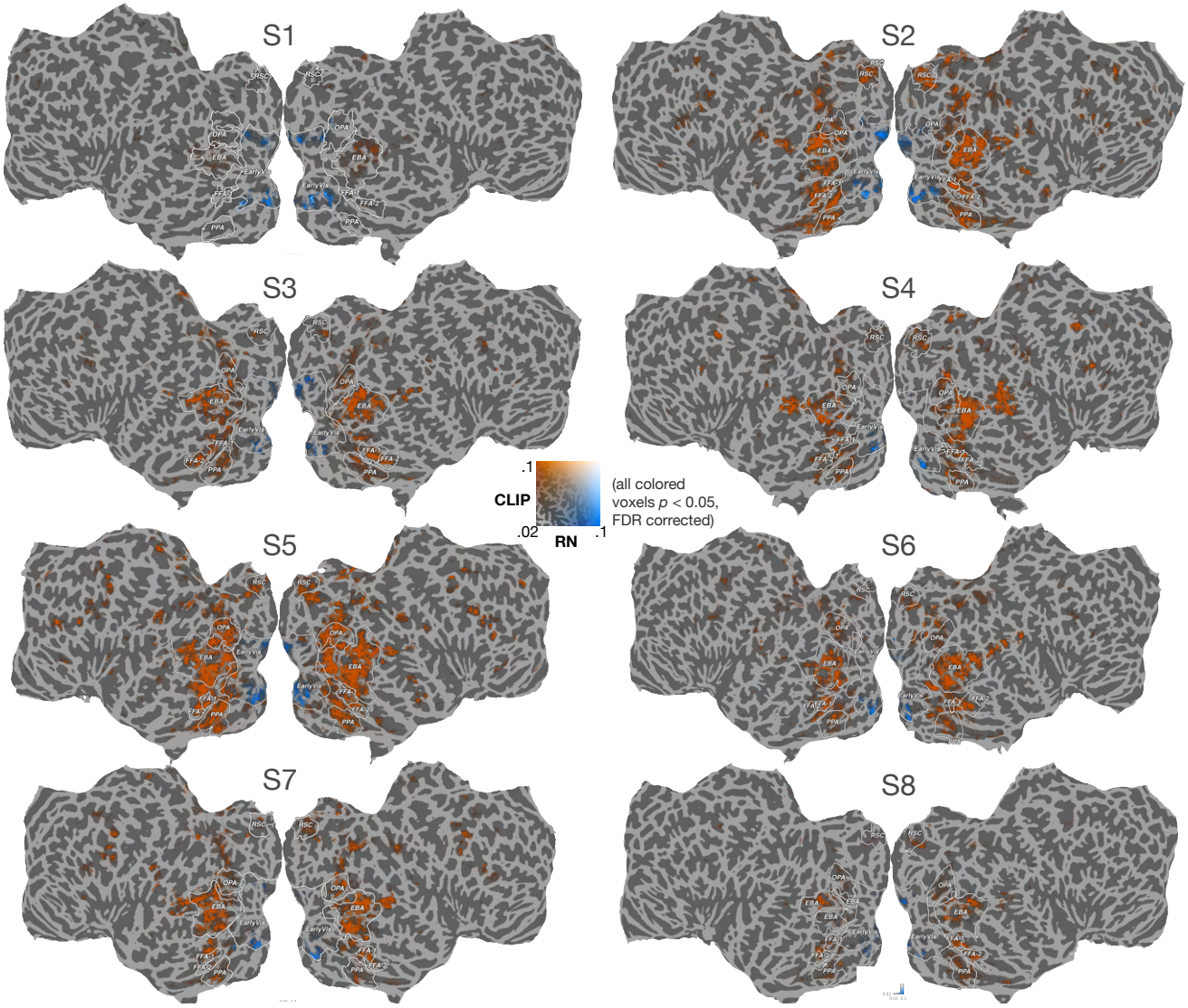
Supplementary Figure S1. Prediction performance meansured in corrlation using the CLIP visual encoder. Voxelwise prediction performance (measured in *r*) on a held-out test set is shown for S5 in a flattened view of the brain.



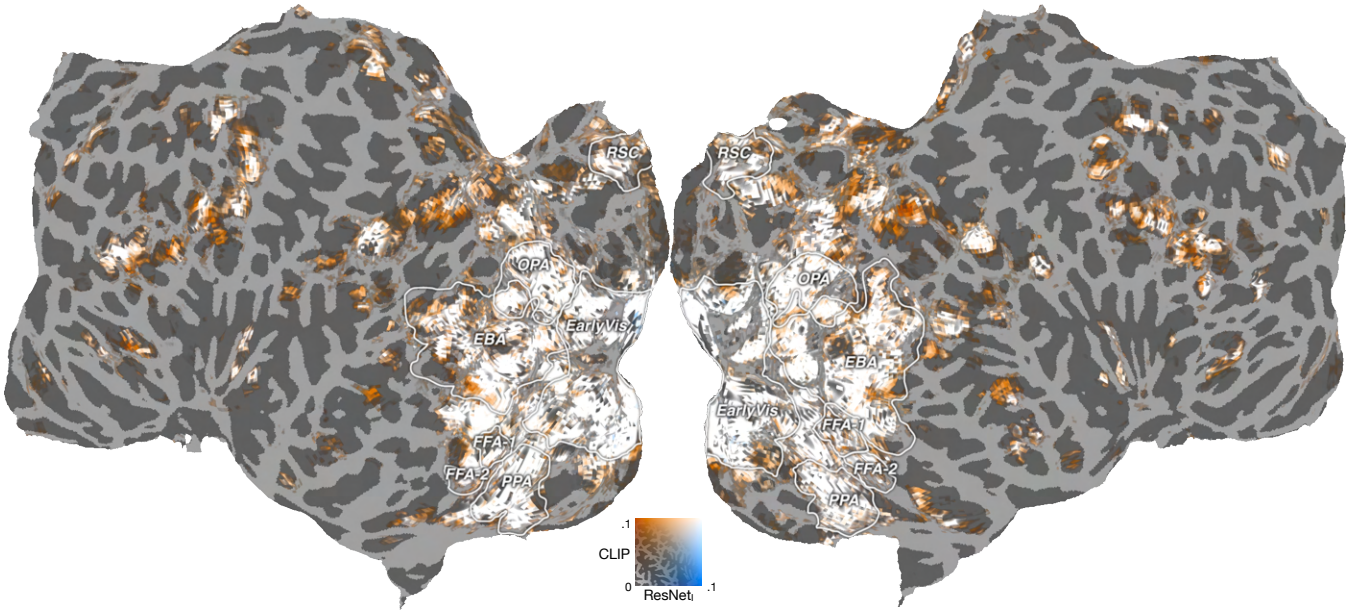
Supplementary Figure S2. Prediction performance with CLIP visual encoder for all eight subjects. Voxelwise prediction performance (measured in R^2) on a held-out test set is shown for S1-S8 in a flattened view of the brain.



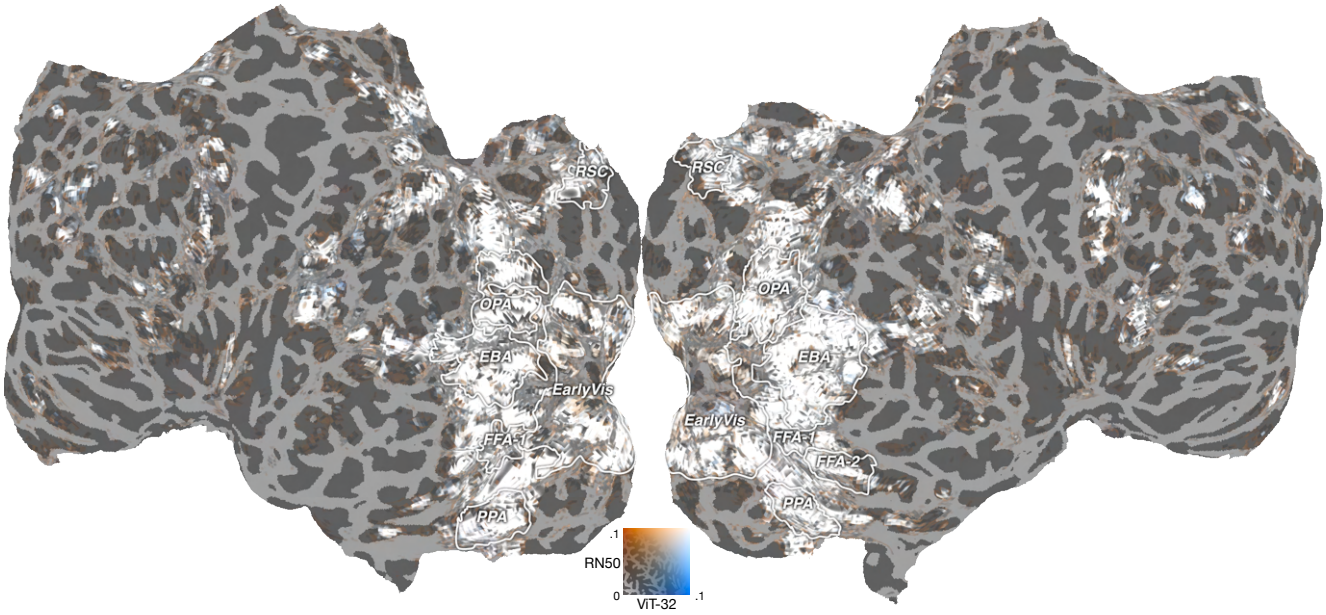
Supplementary Figure S3. Scatterplots of noise ceiling against model performance in R^2 for all subjects.



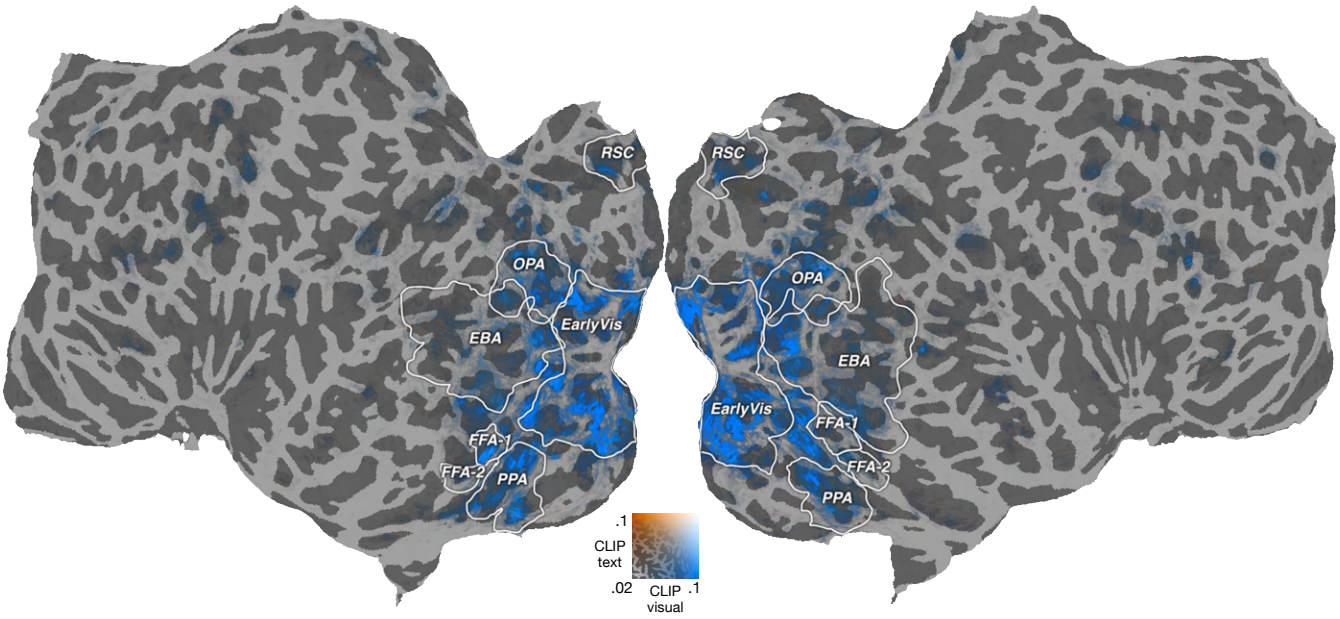
Supplementary Figure S4. Unique variance accounted for by CLIP as compared to ResNet$_I$ (noted as RN in the figure) for all eight subjects. Voxels where CLIP accounts for greater unique variance are orange and voxels where ResNet$_I$ accounts for greater unique variance are blue.



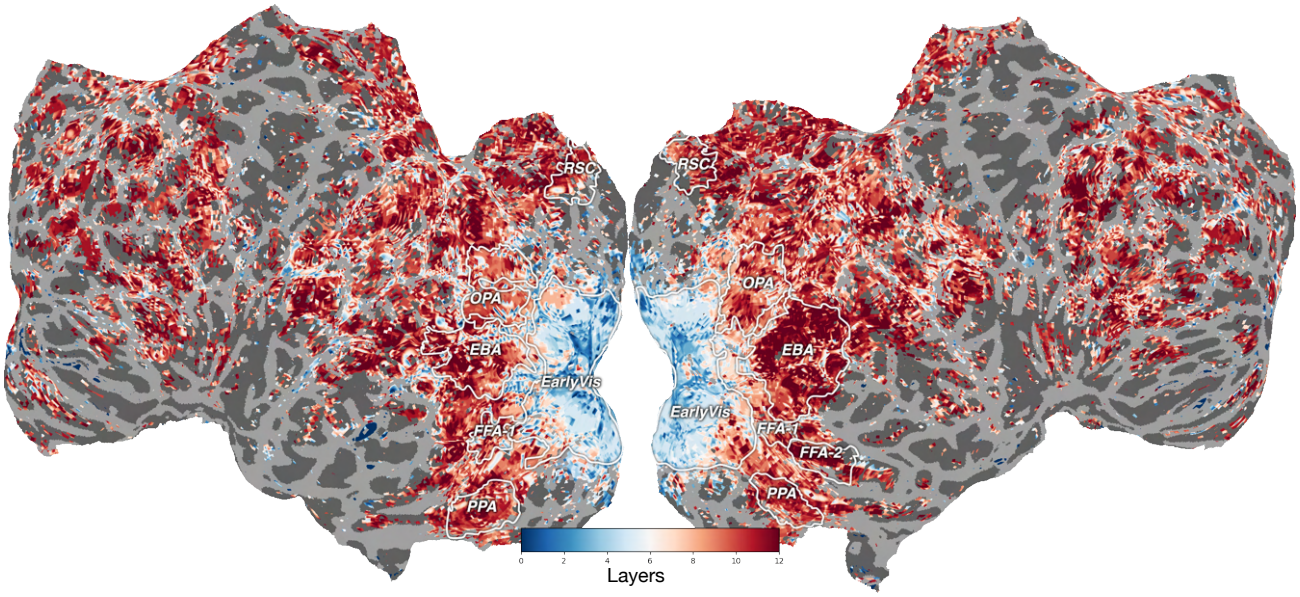
Supplementary Figure S5. Total variance accounted for by CLIP as compared to ResNet$_l$ for S5 Voxels where CLIP accounts for greater variance are orange and voxels where ResNet$_l$ accounts for greater variance are blue.



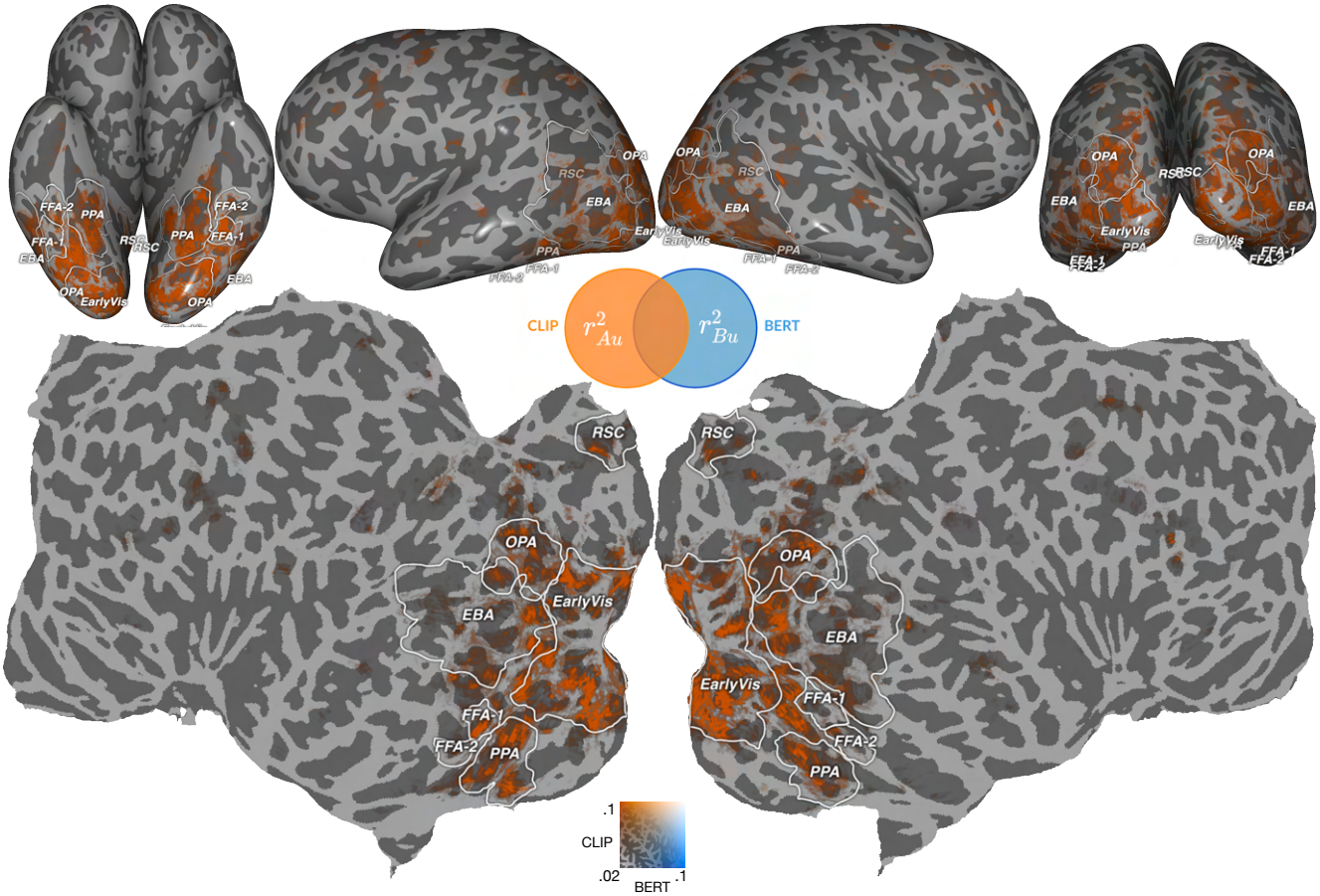
Supplementary Figure S6. Performance 2D map between CLIP (ViT-32) and CLIP (RN50).



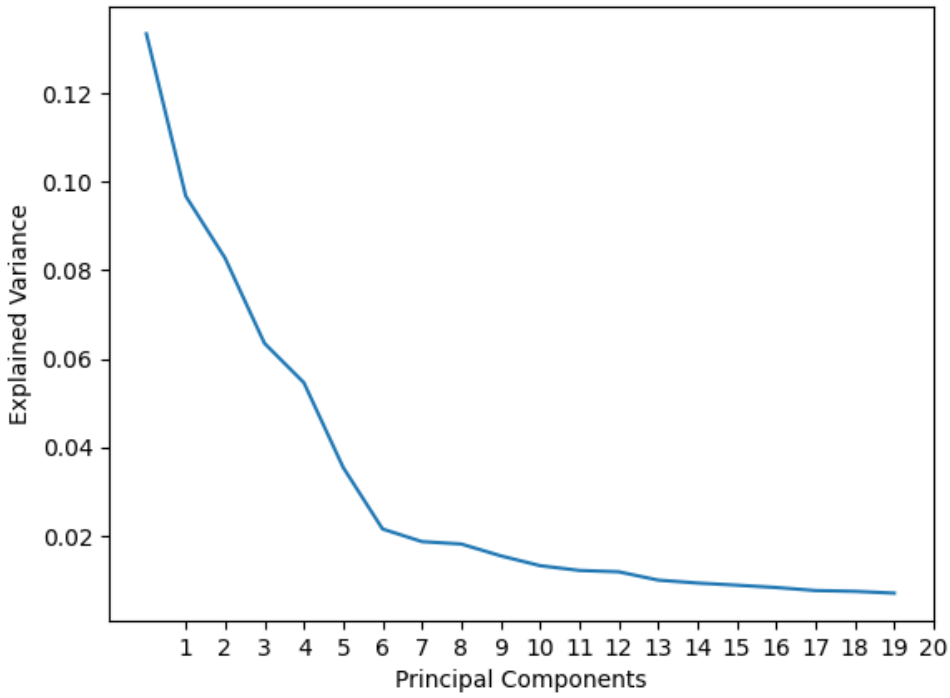
Supplementary Figure S7. Unique variance by CLIP visual encoder and CLIP text encoder.



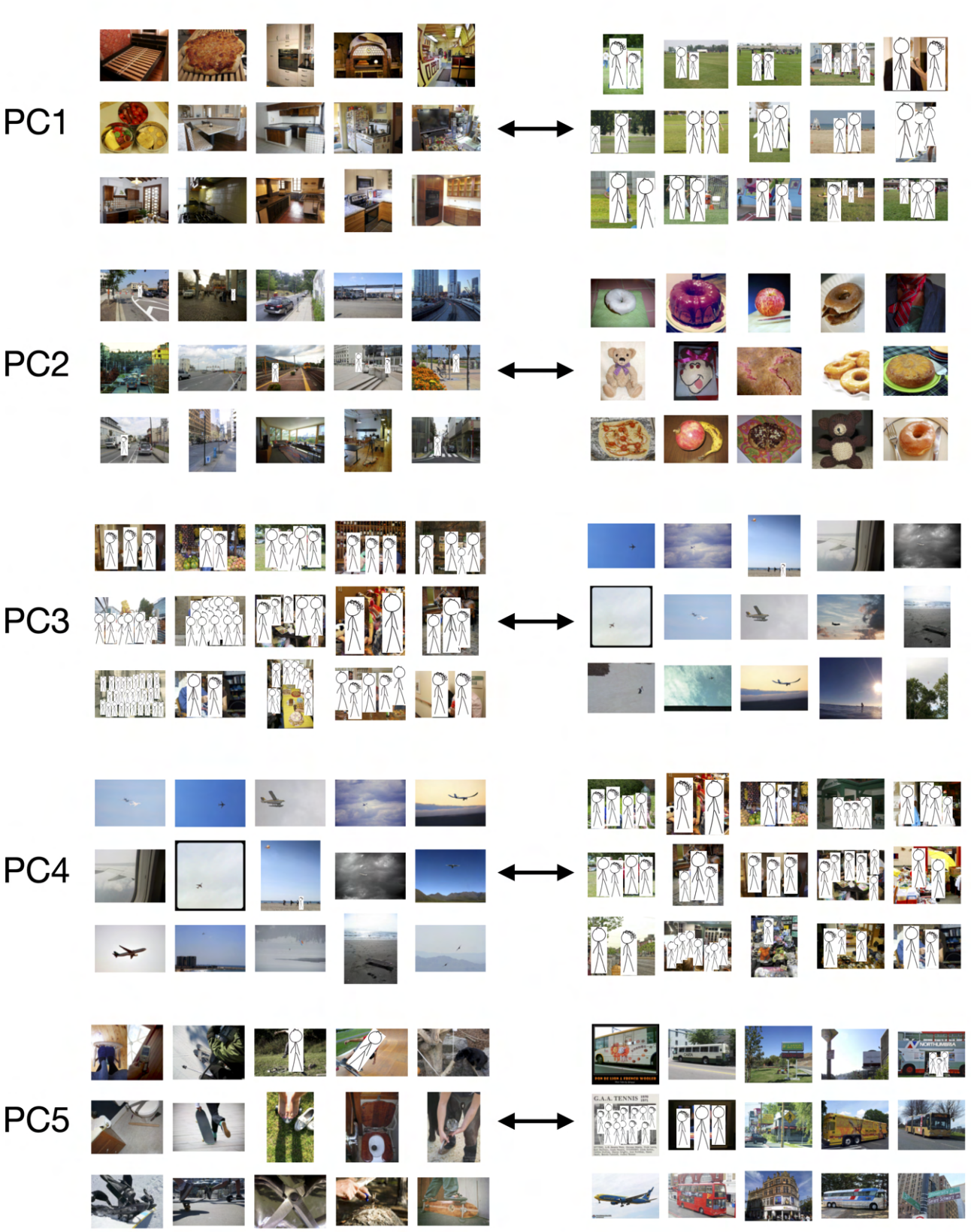
Supplementary Figure S8. Layer preference by voxels across the brain.



Supplementary Figure S9. Performance comparison between CLIP text encoder with BERT Unique variance accounted for by CLIP as compared to BERT for S5 – obtained by subtracting R^2 for each model from that of the the concatenated model. Voxels where CLIP accounts for greater unique variance are orange and voxels where BERT accounts for greater unique variance are blue.



Supplementary Figure S10. Explainable variances across 20 PCs



Supplementary Figure S11. Top 15 images for top 5 PCs