Transformer brain encoders explain human high-level visual responses

Hossein Adeli[1*], **Minni Sun**[1], **Nikolaus Kriegeskorte**[1]

[1]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA

[*] corresponding author: `ha2366@columbia.edu`

Abstract

A major goal of neuroscience is to understand brain computations during visual processing in naturalistic settings. A dominant approach is to use image-computable deep neural networks trained with different task objectives as a basis for linear encoding models. However, in addition to requiring tuning a large number of parameters, the linear encoding approach ignores the structure of the feature maps both in the brain and the models. Recently proposed alternatives have focused on decomposing the linear mapping to spatial and feature components but focus on finding static receptive fields for units that are applicable only in early visual areas. In this work, we employ the attention mechanism used in the transformer architecture to study how retinotopic visual features can be dynamically routed to category-selective areas in high-level visual processing. We show that this computational motif is significantly more powerful than alternative methods in predicting brain activity during natural scene viewing, across different feature basis models and modalities. We also show that this approach is inherently more interpretable, without the need to create importance maps, by interpreting the attention routing signal for different high-level categorical areas. Our approach proposes a mechanistic model of how visual information from retinotopic maps can be routed based on the relevance of the input content to different category-selective regions. Code available at github.com/hosseinadeli/transformer brain encoder.

1 Introduction

An influential approach to study plausible neural computations in the brain is to train Deep Neural Network (DNN) models on different tasks [34, 21] and compare their learned representation to brain activity [42, 17]. There has been a great deal of discussion and research on best ways to compare the learned representations to the ones recorded from the brain (across models and across models and brains). One main approach is to build encoding models— learn a mapping function from one feature domain to another and measure the accuracy of the prediction in held-out sets [10, 28]. An alternative approach is to characterize the geometry or topology of the representation in each model or in the brain and then compare them (e.g. RSA; [22]). In this work, we focus on the learned encoding functions, as we believe that it can give us further insight into the computations in the brain.

The visual system uses structured retinotopic maps as it processes visual information in the cortex. Not surprisingly, models, such as Convolutional and transformer neural networks, that also maintain retinotopic maps of the space perform best on different visual tasks (e.g. recognition and segmentation) and consistently outperform other models in different brain activation prediction benchmarks [36, 13]. However the retinotopic feature maps from deep networks presents typically have a very large number of units posing us with a challenge when mapped unto the responses in the brain. Linear encoding models, although theoretically the simplest choice, can become very high-dimensional in that case

(the number of parameters equals the product of the number of model units and the number of units/voxels to be predicted) and require strong regularization (L2 penalty) given the size of typical fMRI datasets [28]. To address these limitations, approaches have been proposed that learn spatial receptive fields (RF) for different units or voxels in the brain data, using which the representation is first aggregated across space and then the lower dimensional representation is linearly mapped to the brain responses [19, 39, 27]. These models have been shown to perform on par with linear regression models despite having a fraction of the number of parameters and are also more plausible mechanisms of how information can route to different units. However, they can only capture fixed routing where input to a unit comes from a specific area in space regardless of the input content.

Transformer architectures has been extremely successful in many domains, including vision [9] and language [41]. Their success can be attributed to a general and simple (therefore scalable) computational motif where information is routed based on the content. In these models, each token (be a representation of a word in a sentence or a patch in an image) queries other tokens to find how relevant they are to updating its representation. The selective nature of this mixing has motivated naming this process "attention" in Transformers [41]. Then the new representation of this token becomes the average of the representation of all tokens, weighted by their degree of relevance (i.e. attention scores). We hypothesize that the optimal way for the routing of information from the retinotopic visual maps to category selective areas is to use the same computational motif where brain areas only attend to parts of the visual maps with the content relevant to what the area is selective for (Fig. 1). For example if there is face in the image, it could appear anywhere, but the FFA (fuisform face area) can learn to route only the information from the patches where the face-like stimuli are and then expand this lower dimensional representation in the area. Note that this approach is in a way a generalization of the aforementioned RF based methods going from fixed receptive fields to a dynamic content-based receptive fields.

2   Related works

Brain encoding models: Predicting brain activity is an important objective, both as an engineering challenge and also as a means of studying brain computations, reflected in the number of community benchmarks such as Algonauts [13], Brain-score [36], and Sensorium [40]. The availability of large-scale neural datasets has necessitated innovation in new encoding models [16]. Spatial-feature decomposition models have shown that considering the retinotopic maps and the receptive field organization can lead to more efficient encoding models [19, 39, 27, 35]. Generalizing these approaches to high-level visual areas would require considering more dynamic routing motifs.

Self-supervised Vision Transformers: Transformers have been shown to outperform convolutional and recurrent neural networks (CNNs) on a variety of visual tasks including object recognition [9]. More recent studies have explored training these models on self-supervised objectives, yielding some intriguing object-centric properties [1] that are not as prominent in the models trained for classification. When trained with self-distillation loss (DINO, [4] and DINOv2 [30]), the attention values contain explicit information about the semantic segmentation of the foreground objects and their parts, reflecting that these models can capture object-centric representations without labels. These findings show that features from these models can be a good basis for predicting neural activity in the brain. Recent work has also shown that networks trained using self-supervised contrastive losses (such as SimCLR; [7]) match the predictive power of supervised models for high-level ventral-stream visual representations in the brain [20, 6]. These works argue for self-supervised learning methods as a more plausible objective function for learning brain like visual representations.

Encoder-decoder Vision Transformers: Transformer-based encoder-decoder models provide a general framework that has achieved great performance in many domains [41] including domains where one modality (e.g. image) is mapped onto another one (e.g. language) [32]. A related pioneering work to our approach is the DETR model [3] applied to the problem of object detection and grouping in images. The encoder in this model converts the image to rich object-centric features. The decoder uses learnable embeddings, called queries, corresponding to different potential objects, that gather information from the encoder features using cross-attention over several layers. After the decoding process, each object query can then be linearly mapped into to the category and bounding box for an object. The model is trained end-to-end and can detect many objects in one feedforward pass. We also employ this general framework here.
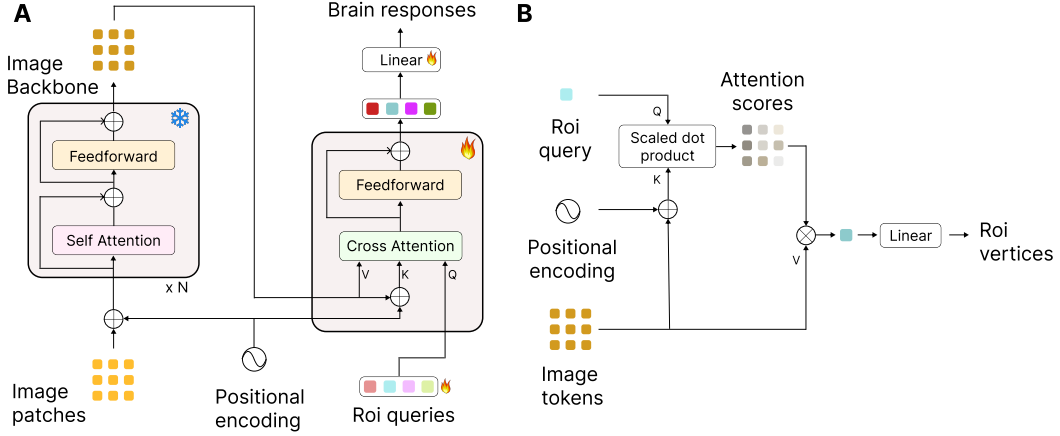
Figure 1: **A.** Brain encoder architecture. The input patches are first encoded using a frozen backbone model. The features are then mapped using a transformer decoder to brain responses. **B.** The cross attention mechanism showing how learned queries for each ROI can route only the relevant tokens to predict the vertices in the corresponding ROI.

3 Methods

3.1 Dataset

We run our experiments on the Natural Scene Dataset (NSD; [2]) where the fMRI responses were collected from 8 subjects, each seeing up to 10,000 images. The reported results are from subjects 1, 2, 5, and 7 who completed all recording sessions. The surface-based fMRI responses across the three repetitions of each image were averaged for model training and testing. We use the train/test split that was introduced in the Algonauts benchmark [13] where the last three sessions for each subject were held out to ensure that no test data were accessed during the model development and to make the prediction task as natural as possible (predicting the future responses). Our analyzes also focused on the most visually responsive part of the brain, approximately 15k vertices for each left and right hemispheres (LH and RH) in the visual cortex, shown in Figure 2A on a surface map. ROI level labels were provided for all the selected vertices based on visual and categorical properties (using auxiliary experiment; refer to [2] for details). The labels are for early visual areas ('V1v', 'V1d', 'V2v', 'V2d', 'V3v', 'V3d', and 'hV4'), body selective areas ('EBA', 'FBA-1', 'FBA-2', and 'mTL-bodies'), face selective areas ('OFA', 'FFA-1', 'FFA-2', 'mTL-faces', and 'aTL-faces'), place selective areas ('OPA', 'PPA', 'RSC'), and word selective areas ('OWFA', 'VWFA-1', 'VWFA-2', 'mfs-words', and'mTL-words').

3.2 Transformer brain encoder

We apply the the general transformer encoder-decoder framework to map images to fMRI responses. Figure 1A shows the architecture of our model. The input image is first divided into patches (31×31 in our dataset) of size 14×14 pixels. These image patches are input to the backbone model which is a 12-layer vision transformer and frozen to be used as a feature backbone.

The decoder uses input queries corresponding to different brain ROIs in different hemispheres to gather relevant information from the backbone outputs for predicting neural activity in each ROI. Note that these queries are learnable embeddings for each ROI trained as part of the model training. We use a single-layer transformer for the decoder with one cross-attention and a feedforward projection. Figure 1B shows the cross-attention process. The positional encoding is added to the image token representation to create the keys. This allows the ROI query to attend either to the location or the content of the input tokens through scaled dot-product attention. The attention scores are then used to aggregate all the image tokens that are relevant to predict the brain activity in that ROI. The output decoder tokens are then mapped using a single linear layer to fMRI responses of the corresponding ROI. In our implementation, decoder output for each ROI is linearly mapped to a vector with the size
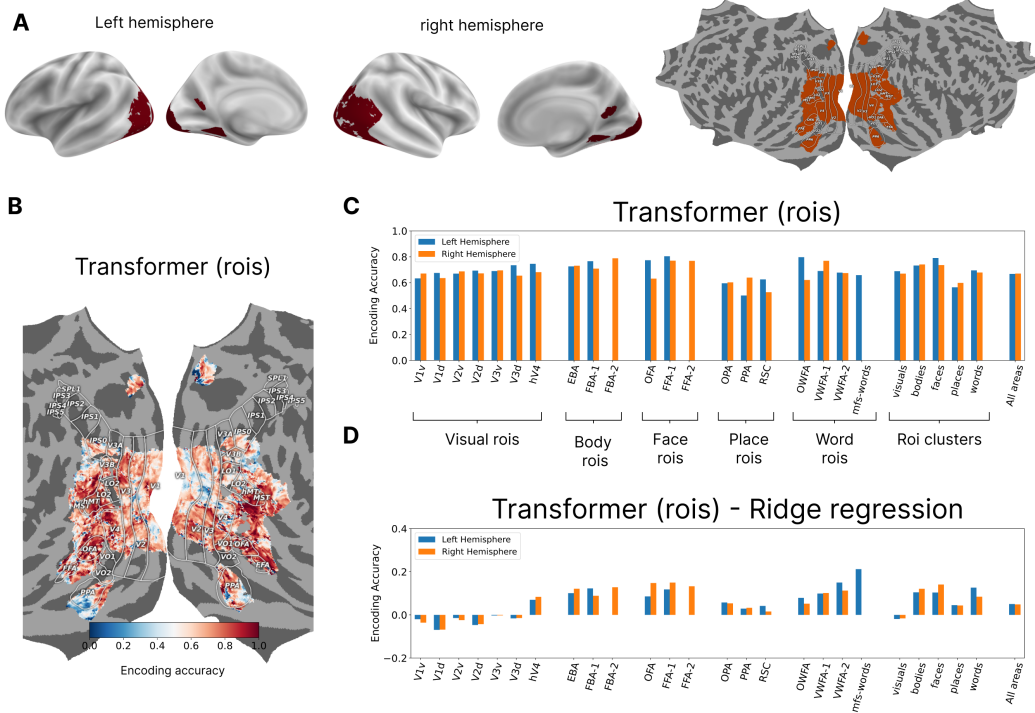
Figure 2: **A.** The general region of interest for highly visually responsive vertices in the back of the brain shown on different surface maps. **B.** Encoding accuracy (fraction of explained variance) shown for Subject 1 for all the vertices for the transformer model using ROIs for decoder queries. **C.** Encoding accuracy for individual ROIs and for ROI clusters based on category selectivity for the two hemispheres. **D.** The differences in encoding accuracy between the transformer and the ridge regression models showing that improvement in the former is driven by better prediction of higher visual areas.

equal to the number of vertices in that hemisphere. The response is then multiplied by a mask that is zero everywhere except for the vertices belonging to that ROI. This masking operation ensures that the gradient signal feeding back from the loss will only train linear mappings to the vertices of the queried ROI. The responses from different ROI readouts will then be combined using the same masks to generate the prediction for each hemisphere. The ROI queries, transformer decoder layer and the linear mappings are trained with the Adam optimizer [18] using mean-squared-error loss between the prediction and the ground truth fMRI activity for each image. We train and test the models separately for each subject.

4 Experiments

We did 10-fold cross validation using the training set for each subject and chose the model with the best validation performance in each fold and then averaged their predictions on the test split. The model predictions were evaluated first using Pearson correlation between the predictions and the ground truth data. The squared correlation coefficient were then divided by the noise ceiling (see [2] Methods, Noise ceiling estimation) to calculate the encoding accuracy as the fraction of the explained variance.

We present results using multiple different feature backbones namely, DINOv2 base model [30], ResNet50 [15], and CLIP large model [32]. For the DINOv2 backbone, inspired by prior work on human attention prediction [1], we did some preliminary analyses and found the patch level query representations (instead of values) to have slightly more predictive power and chose to use them in all our experiment. For ResNet50, the feature maps from the last layer were extracted and reshaped to create the visual tokens comparable to transformers. For CLIP, we chose the large model to have

the same image patch size (14) and transformer token dimension (768) to the DINOv2 base model. Unless otherwise stated, the features from the last layer of the backbone models are used as the input representation to the decoder.

We consider multiple different mapping functions to compare to our proposed method. The Ridge regression model flattens the feature representation across space and feature dimensions and learns one linear mapping to the fMRI responses. We used a grid search to select the best ridge penalty to maximize performance on the validation data. For the spatial-feature factorized method, the model learns a (H * W) spatial map and applies that to the input feature similar to the attention map in Figure 1B. The scores however are only learned for a given ROI or a vertex and are not dependent on the content of the image. The spatial map then aggregates the features to be linearly mapped to the brain responses. For the transformer brain encoder, we used 24 queries per hemisphere corresponding to the 24 ROIs. Note that not all ROIs were present in all the subjects, therefore we present results and figures for subjects individually. If an ROI is not mapped in a subject the decoder output is not mapped to any vertices. The figures in the main text are generated using the results from subject 1, but the figures for the remaining three subjects are presented in the supplementary section A.1.

Table 1: Encoding accuracy using DINOv2 backbone

Encoder	Subjects				Model size (M)
	S1	S2	S5	S7	
Ridge regression	0.56	0.52	0.50	0.37	∼1200
Spatial-feature factorized (rois)	0.49	0.46	0.48	0.37	∼31
Transformer (rois)	**0.60**	**0.56**	**0.56**	**0.42**	∼37

Table 1 shows the encoding accuracy of the three encoding models using the DINOv2 backbone. Ridge regression requires tuning a larger number of parameters compared to the other two approaches (all model sizes reported as multiples of millions of parameters). Our model consistently outperforms the spatial-feature factorized model and the ridge regression model across all subjects.

Figure 2B shows the encoding accuracy of our model for subject 1 for the areas of interests projected onto the cortical surface using Pycortex [11]. Figure 2C shows the encoding accuracy divided over all the individual ROIs and also clusters of ROIs. When we compare the transformer encoder to the ridge regression model (Fig. 2D), we see that our model achieves higher encoding accuracies through better performance for categorical areas. This suggests that content based routing can be part of the brain computation for higher level visual areas.

To examine whether our results depend on the specific choice of the transformer backbone architecture, we tested all the encoding models on the ResNet50 backbone features (a fully convolutional network). Table 2 shows that we replicate the exact same pattern of accuracy as the DINOv2 backbone, where the transformer encoder outperforms the other two alternatives across all subjects. This shows that the transformer encoder can map differently learned features (transformer vs convolution) well to the brain data.

Table 2: Encoding accuracy using ResNet50 backbone

Encoder	Subjects				Model size (M)
	S1	S2	S5	S7	
Ridge regression	0.49	0.48	0.47	0.37	∼1200
Spatial-feature factorized (rois)	0.42	0.42	0.43	0.33	∼80
Transformer (rois)	**0.52**	**0.50**	**0.50**	**0.38**	∼37

4.1 Vertex-based routing

So far the presented transformer encoding models used ROIs as units of routing. But the routing could be made more granular by learning a decoder query for each vertex where the gathered features from the decoder would be mapped linearly to the corresponding vertex value. This approach can also be applied in the spatial-feature encoding models where a spatial map is learned per vertex.
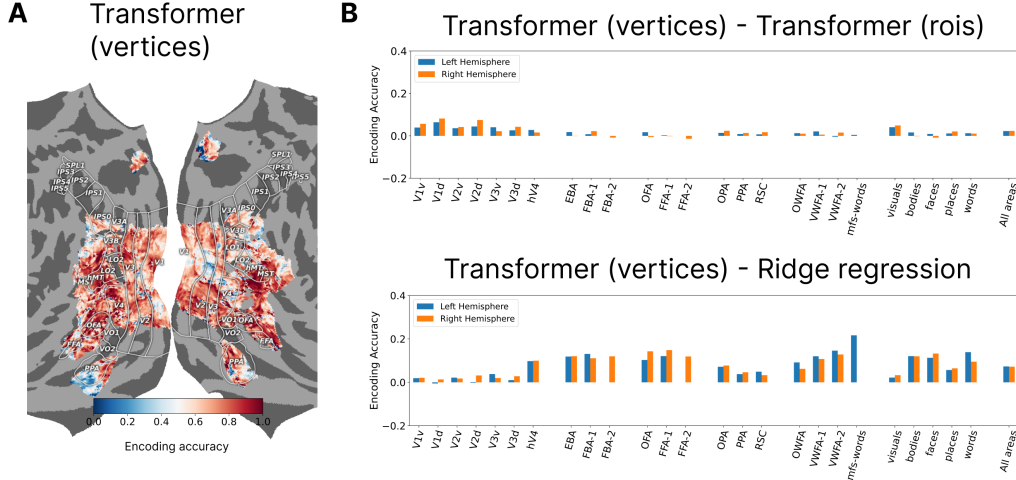
Figure 3: **A.** The encoding accuracy for subject 1 shown on the brain surface for the transformer model with vertices as decoder queries. **B.** The difference in encoding accuracies going from ROIs to vertices as the decoder queries shows the improvement is almost entirely from the early visual areas. **C.** The vertex-based transformer model outperforms the ridge regression model for almost all the ROIs.

Table 3 shows model accuracies for these two approaches using the vertex-based routing, indicating improvements for both models across all the subjects. Examining the encoding accuracy for individual ROIs (Fig. 3), we can see that the performance boost came almost entirely from early visuals areas for the transformer based model. The fact that shifting from ROI-based to vertex-bases routing does not improve encoding accuracy for higher visual ares indicates that ROIs may be the right level of routing for those regions, however the early visual areas requires more granular routing because the receptive fields of the vertices are smaller and less content dependent. Comparing the vertex-based transformer model to the ridge regression model (Fig. 3B) shows that the former now outperforms the latter in almost all the ROIs, as a result of more granular routing in early visual areas.

Table 3: Encoding accuracy for different decoder queries

Encoder	Subjects				Model size (M)
	S1	S2	S5	S7	
Spatial-feature factorized (rois)	0.49	0.46	0.48	0.37	~31
Spatial-feature factorized (vertices)	0.52	0.48	0.48	0.37	~68
Transformer (rois)	0.60	0.56	0.56	0.42	~37
Transformer (vertices)	**0.63**	**0.59**	**0.57**	**0.44**	~67
Transformer (vertices) backbone layers ensemble	**0.65**	**0.62**	**0.59**	**0.45**	~400

Motivated by previous encoding models of the brain having used CLIP embeddings [32] to represent images [24], we tested the different mapping functions using this feature backbone. Table 4 shows while the performance is generally not as good as the DINOv2 backbone, it yields the same exact pattern of results. The Transformer-based models outperform other alternatives with the vertex-based routing reaching higher performance overall. Taken together with also the lower performance we saw with ResNet50 backbone, the DINOv2 features, a self-supervised trained vision transformer, deserve consideration as models of human visual brain representations.

4.2 Ensemble

A concern with using complex encoding models for neural system identification is that the non-linear mapping may obscure the differences in the underlying representations [16]. However, our

Table 4: Encoding accuracy using CLIP vision backbone

Encoder	Subjects				Model size (M)
	S1	S2	S5	S7	
Ridge regression	0.51	0.48	0.47	0.38	∼650
Spatial-feature factorized (rois)	0.38	0.35	0.40	0.31	∼30
Spatial-feature factorized (vertices)	0.44	0.40	0.42	0.32	∼40
Transformer (rois)	0.53	0.49	0.50	0.38	∼37
Transformer (vertices)	**0.55**	**0.52**	**0.52**	**0.40**	∼67

Figure 4: **A.** Encoding accuracy of the transformer encoding model with vertex-based queries ensembled across backbone layers. **B.** Showing the backbone layer from which each vertex was best predicted. **C.** The improved performance of ensembling is almost entirely from better prediction of early visual areas.

results with different feature backbones show that the ones that perform better using the linear model consistently perform better using our transformer encoding model as well, just with the latter achieving higher accuracies.

To address this concern further, we consider a robust phenomenon shown consistently using linear encoding with convolutional neural network backbones, where the earlier layers of the network are better features for predicting the earlier visual areas [42, 14, 29, 17, 43]. We trained different transformer decoders with image tokens coming from different layers of the DINOv2 backbone. We then use a softmax operation across the ensemble of models to get the final prediction for each voxel. The softmax weights are based on goodness of the prediction for each model for that vertex in the validation set. Figure. 4A shows the accuracy of the overall model on the brain surface for subject 1. The layers that had the highest weights in the ensemble for predicting for each given voxel is shown in 4B; higher visual areas were better predicted by later encoder layers, indicating that encoder layers capture similar feature abstractions as the brain.

Comparing the ensemble model to the model trained using only the final backbone layer features (Fig. 4C), we can see that the performance increase is entirely driven by better prediction of earlier visual areas. These results show that our encoding model does not obscure the differences in the underlying representation pointing further to its plausibility.

4.3 Attention maps

Different methods have been developed to interpret linear encoding models to make claims about the the selectivity learned for each ROI. Some methods tend to retrieve or generate images that highly activate the ROI vertices [25, 26, 5], and others focus on creating importance maps to show which parts of the input images are important for predicting the activity of an ROI [33].

The difference in our approach is that the cross-attention scores (Fig. 1B) can be examined to reveal the selectivity for each ROI making our model inherently more interpretable. We visualize the attention maps for 3 different ROIs in Figure 5 for the transformer encoder trained with ROI decoder queries with DINOv2 backbone. First is an early visual area, V2d (dorsal) in the left hemisphere.

7

Figure 5: **Attention maps.** Transformer decoder cross attention scores for three ROIs overlaid on the images. The selected ROIs show different ways in which the learned ROI queries can route information— based on location (V2d), content (FBA), or a combination of the two (OFA) depending on the location of the ROI in the brain processing hierarchy.

Since the visual field is flipped around both horizontal and vertical meridians in the cortex (starting from the retina), we expect the brain activity in this area to represent visual information from the bottom-right of the input (given that the subjects were instructed to hold fixation at the center of the screen for the presentation duration). We see this exact pattern emerge in the attention maps. Recall that the decoder queries can learn to attend to both patch locations or their content (since the key value is the sum of backbone image patches and positional encoding). In this case, the attention seems to completely be driven by the location, similarly for all the images, ignoring the content. This is exactly what we would expect from an early visual area. The fact that all the vertices in this ROI have to share the same attention map hurts accuracy as we saw in Figure 2D since the vertices do have smaller RFs in this area than a quadrant, however this can be addressed by vertex level routing.

The second ROI is OFA in the right hemisphere, a mid-level face selective area [12]. The attention maps is this ares consistently focus on faces. Since this area is in the right hemisphere it also has a preference for visual input in the left visual field. We can see this for cases with multiple faces where the second face in the right visual field in not strongly attended. The decoder query therefore makes use of both the positional encoding and he content component of the key to attend to the most relevant part of the image to predict vertices in this ROI. The attention could also be spread across multiple faces in different locations. This is the important dynamic aspect of the receptive field in higher visual areas that can be captured using the transformer attention mechanism. The third area is FBA in the right hemisphere, a high level body selective area [31]. The attention maps are more spread across bodies for this ROI and not just faces. In the Supplementary section A.2, we provide an analyses of the similarity between the learned queries for different ROIs (capturing visual and semantic similarity between them) and also show how our model can be used in an interpretability pipeline using diffusion models [24] to generate stimuli that maximally activate different ROIs (section A.3).

Table 5: Encoding accuracy using BERT backbone

Encoder	Subjects				Model size (M)
	S1	S2	S5	S7	
Ridge regression	0.19	0.21	0.25	0.19	$\sim$1200
Transformer (rois)	**0.27**	**0.27**	**0.33**	**0.27**	$\sim$37

4.4 Text modality

We have tested the transformer encoding model on a few vision backbones but it remains to test whether this approach is generalizable to other modalities. TO test this, we first used the BLIP
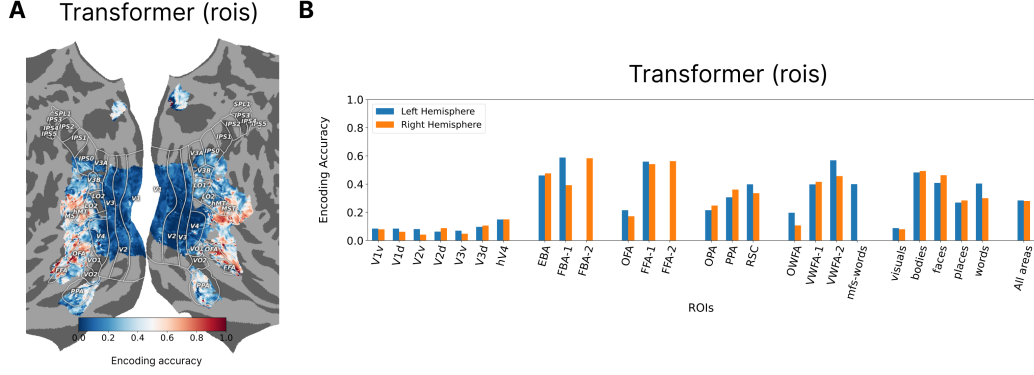
Figure 6: **A.** Transformer encoder accuracy using image caption as input **B.** Only high-level visual areas are predicted by semantic information in a caption.

model [23] to generate short captions for all the images in the dataset. Using BERT [8] as the feature backbone, the decoder work exactly as before, using ROI queries to map backbone features to fMRI responses. Table 5 shows how the transformer model outperforms the regression model across all subjects (with a fraction of the parameters). Given only semantic information available in the captions, the model can only predict the high level visual areas as shown in Figures 6A and 6B.

5 Discussion

Linear encoding models have been the dominant method used for learning the mapping from model features to brain activity [10]. The reasons for this (see [16] for a review of these points) include theoretical simplicity, allowing comparison among backbone features, biological plausibility, and the ability to interpret the learned weights. However, this approach is parameters inefficient for a typical number of voxels and image features, ignores the organization of the features, and does not capture nonlinear computations between brain areas such as ubiquitous normalizations [38]. Our proposed routing based method not only reaches state of the art accuracy, it also achieves the aforementioned desiderata for encoding models, as we have shown in our results.

Foundation models (e.g. DONOv2 or CLIP for vision) trained with self-supervised objectives can serve as general visual representation backbones. However these task agnostic models do not capture all the computations in the brain and between brain areas, which needs be addressed by learning better encoding models. Our work suggests a mechanism for how different brain areas dynamically gate their input based on the input content and the area selectivity. Our results showing that the encoding accuracy for high-level areas cannot be improved beyond ROI-based routing also agrees with prior work on between area interactions using communication subspaces [37]. The routed information that is relevant to an area can then get expanded more in-depth. This process allows for cutting down on wiring cost in the brain by not connecting all the units in one area to another area but rather only a subset of relevant information getting routed with more local connections expanding the representation.

Limitations: We performed our experiments on NSD [2], the largest image viewing fMRI dataset to date. It will be important to test the generality of our approach on other datasets using different recording techniques (Neurophysiology, EEG, etc) and on different input modalities (such as video and audio). We used vertex-wise routing to capture the responses in early visual areas but while the computations for smaller receptive fields can be learned by this approach, the way the RFs are implemented in the brain are through different anatomical and wiring constraints. Also we chose for the model to read out the brain responses from a backbone for both early and high-level visual areas. Future work will seek to explore the connectivity between early and high-level visual areas in a more integrated system and test whether making the model further aligned with known anatomy of the visual cortex will improve performance.

9

References

[1] Hossein Adeli, Seoyoung Ahn, Nikolaus Kriegeskorte, and Gregory Zelinsky. Affinity-based attention in self-supervised transformers predicts dynamics of object grouping in humans. *arXiv preprint arXiv:2306.00294*, 2023.

[2] 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, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[5] Diego García Cerdas, Christina Sartzetaki, Magnus Petersen, Gemma Roig, Pascal Mettes, and Iris Groen. Brainactiv: Identifying visuo-semantic properties driving cortical selectivity using diffusion-based image manipulation. *bioRxiv*, pages 2024–10, 2024.

[6] Honglin Chen, Rahul Venkatesh, Yoni Friedman, Jiajun Wu, Joshua B Tenenbaum, Daniel LK Yamins, and Daniel M Bear. Unsupervised segmentation in real-world images via spelke object inference. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 719–735. Springer, 2022.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Jack L Gallant, Shinji Nishimoto, Thomas Naselaris, and MC Wu. System identification, encoding models, and decoding models: a powerful new approach to fmri research. *Visual population codes: Toward a common multivariate framework for cell recording and functional imaging*, pages 163–188, 2012.

[11] James S. Gao, Alexander G. Huth, Mark D. Lescroart, and Jack L. Gallant. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9, September 2015. ISSN 1662-5196. doi: 10.3389/fninf.2015.00023. URL http://journal.frontiersin.org/Article/10.3389/fninf.2015.00023/abstract.

[12] Isabel Gauthier, Michael J Tarr, Jill Moylan, Pawel Skudlarski, John C Gore, and Adam W Anderson. The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of cognitive neuroscience*, 12(3):495–504, 2000.

[13] Alessandro T Gifford, Benjamin Lahner, Sari Saba-Sadiya, Martina G Vilas, Alex Lascelles, Aude Oliva, Kendrick Kay, Gemma Roig, and Radoslaw M Cichy. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv preprint arXiv:2301.03198*, 2023.

[14] Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Anna A Ivanova, Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. Beyond linear regression: mapping models in cognitive neuroscience should align with research goals. *Neurons, Behavior, Data analysis, and Theory*, 1, 2022.

[17] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] David Klindt, Alexander S Ecker, Thomas Euler, and Matthias Bethge. Neural system identification for large populations separating "what" and "where". *Advances in neural information processing systems*, 30, 2017.

[20] Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.

[21] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446, 2015. ISSN 2374-4642.

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

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[24] Andrew Luo, Maggie Henderson, Leila Wehbe, and Michael Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36:75740–75781, 2023.

[25] Andrew F Luo, Margaret M Henderson, Michael J Tarr, and Leila Wehbe. Brainscuba: Fine-grained natural language captions of visual cortex selectivity. *arXiv preprint arXiv:2310.04420*, 2023.

[26] Andrew F Luo, Jacob Yeung, Rushikesh Zawar, Shaurya Dewan, Margaret M Henderson, Leila Wehbe, and Michael J Tarr. Brain mapping with dense features: Grounding cortical semantic selectivity in natural images with vision transformers. *arXiv preprint arXiv:2410.05266*, 2024.

[27] Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay K Jagadish, Eric Wang, Edgar Y Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S Tolias, et al. Generalization in data-driven models of primary visual cortex. *BioRxiv*, pages 2020–10, 2020.

[28] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[29] Aran Nayebi, Javier Sagastuy-Brena, Daniel M Bear, Kohitij Kar, Jonas Kubilius, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel LK Yamins. Goal-driven recurrent neural network models of the ventral visual stream. *bioRxiv*, 2021.

[30] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[31] Marius V Peelen and Paul E Downing. Selectivity for the human body in the fusiform gyrus. *Journal of neurophysiology*, 93(1):603–608, 2005.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[33] N Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):5540, 2021.

[34] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[35] Shreya Saha, Ishaan Chadha, et al. Modeling the human visual system: Comparative insights from response-optimized and task-optimized vision models, language models, and different readout mechanisms. *arXiv preprint arXiv:2410.14031*, 2024.

[36] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[37] João D Semedo, Amin Zandvakili, Christian K Machens, Byron M Yu, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.

[38] H Sebastian Seung. Interneuron diversity and normalization specificity in a visual system. *bioRxiv*, pages 2024–04, 2024.

[39] Ghislain St-Yves and Thomas Naselaris. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, 180:188–202, 2018.

[40] Polina Turishcheva, Paul G Fahey, Michaela Vystrčilová, Laura Hansel, Rachel Froebe, Kayla Ponder, Yongrong Qiu, Konstantin F Willeke, Mohammad Bashiri, Eric Wang, et al. The dynamic sensorium competition for predicting large-scale mouse visual cortex activity from videos. *ArXiv*, pages arXiv–2305, 2024.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[42] Daniel LK 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, 2014.

[43] Huzheng Yang, James Gee, and Jianbo Shi. Brain decodes deep nets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23030–23040, 2024.

A   Supplementary Material

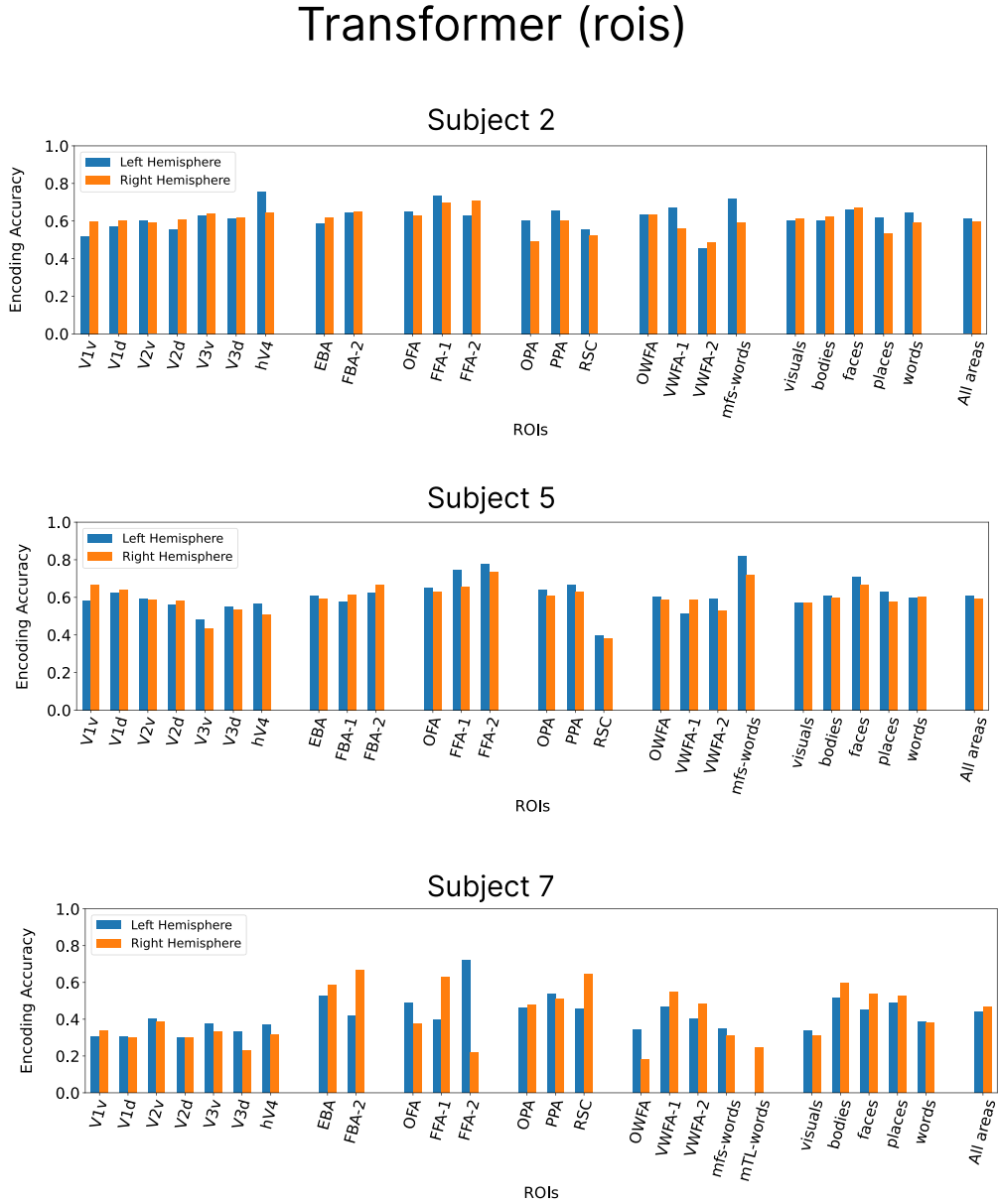## A.1   Encoding accuracies for Subjects 2, 3 and 7

Figure S1: Encoding accuracy (fraction of explained variance) shown for Subjects 2, 5, and 7 for individual ROIs and for ROI clusters for the two hemispheres. The transformer model uses ROIs for decoder queries and features from the last layer of the DINOv2 backbone.
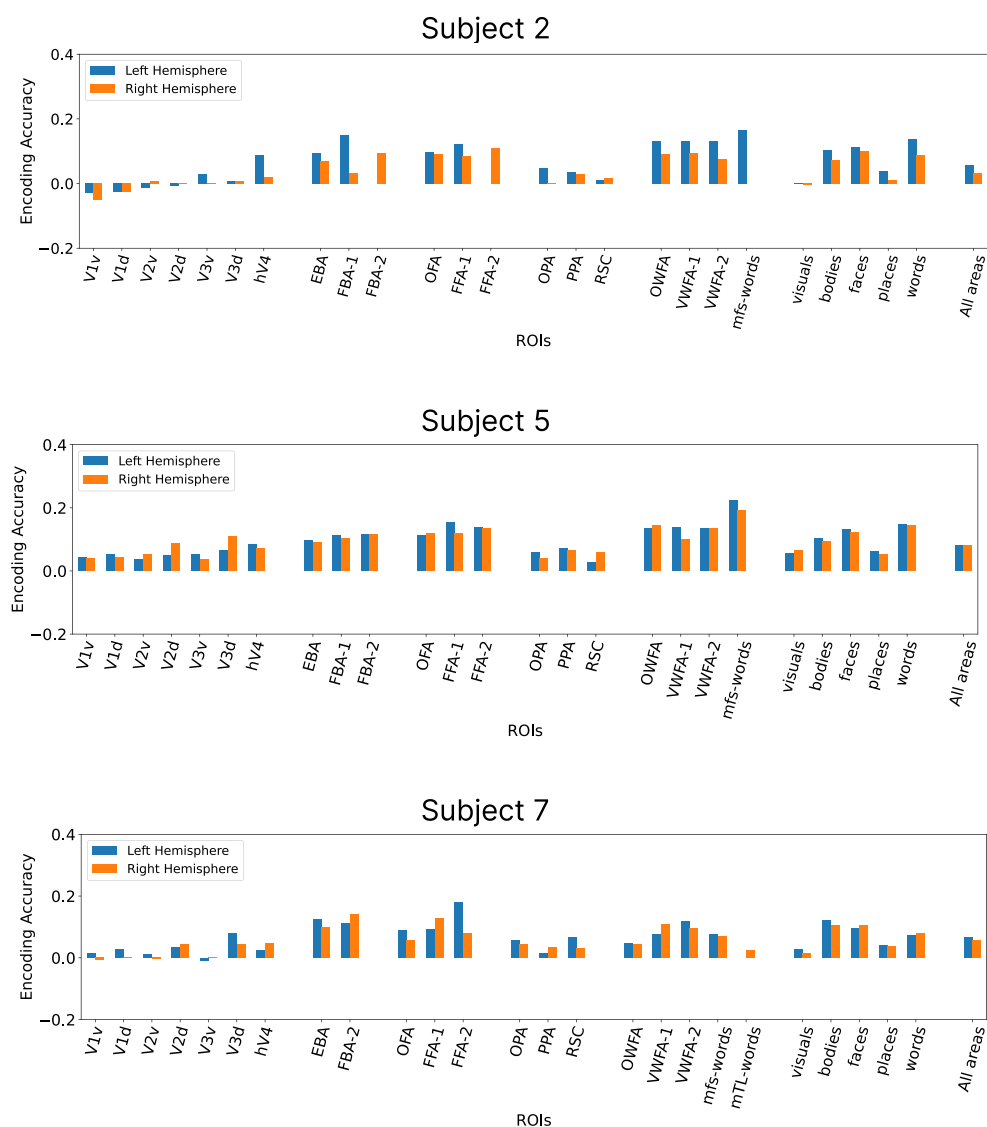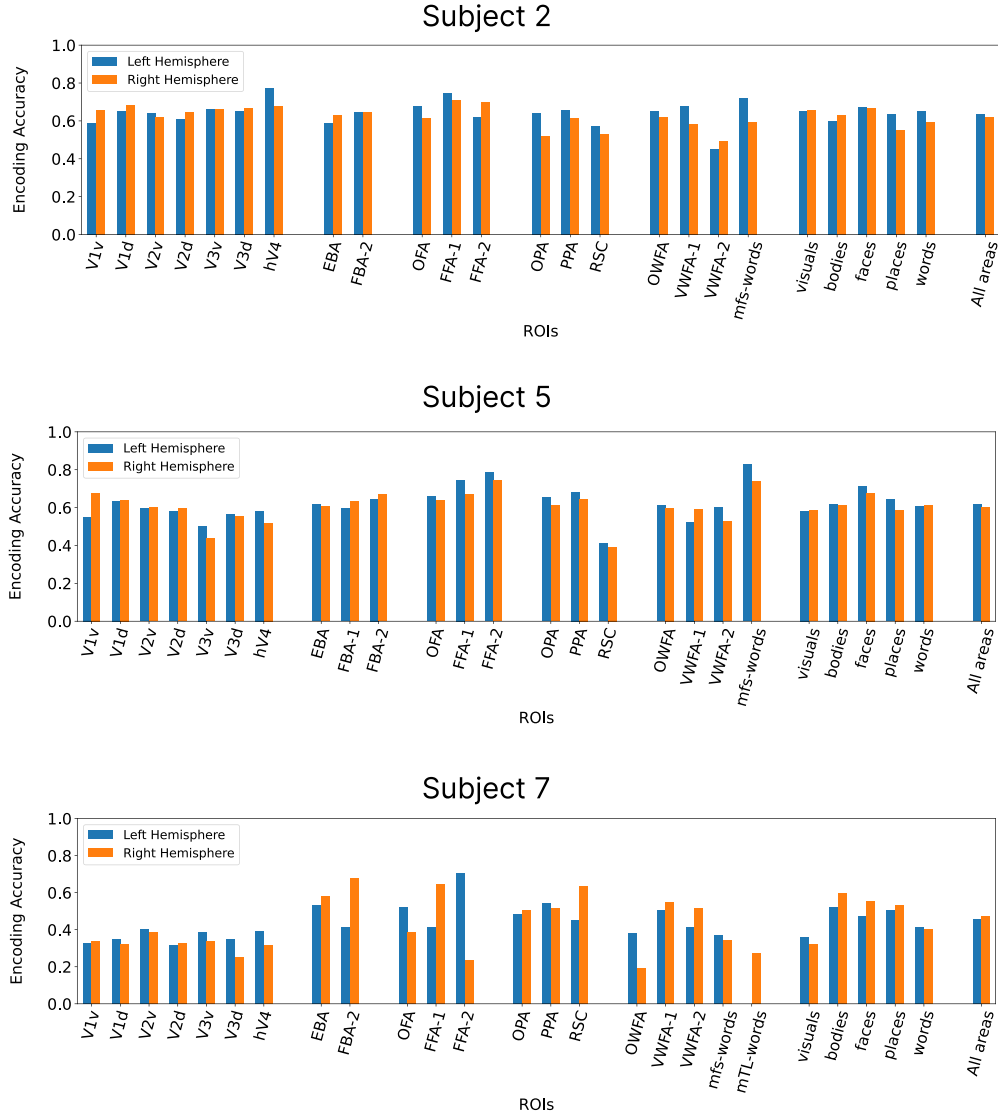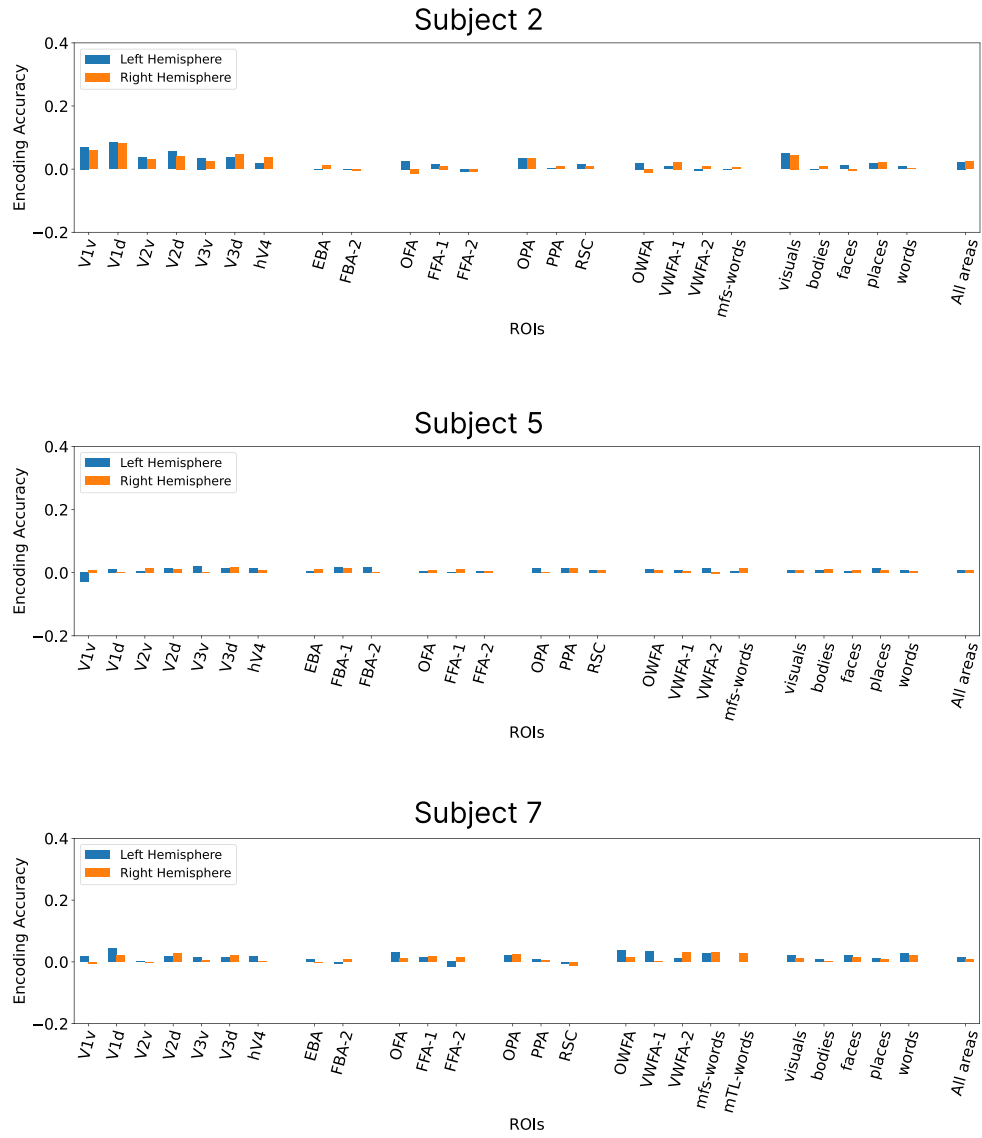
Transformer (rois) - Ridge Regression

Figure S2: The differences in encoding accuracy between the transformer and the ridge regression models showing that any improvement in the former is driven by better prediction of higher visual areas.

Transformer (vertices)

Figure S3: Encoding accuracy (fraction of explained variance) shown for Subjects 2, 5, and 7 for individual ROIs and for ROI clusters for the two hemispheres. The transformer model uses vertices for decoder queries and features from the last layer of the DINOv2 backbone.

Transformer (vertices) - Transformer (rois)

Subject 2

Subject 5

Subject 7

Figure S4: The differences in encoding accuracy between the transformer model using vertices and the model using ROIs as decoder queries. The figure shows that any potential improvement in the former is driven by better prediction of early visual areas.

A.2 Analyzing learned ROI queries

We analyzed the representational similarity of learned ROI queries, and report the average cosine similarity between each pair of ROIs across 20 models trained using five different random seeds and four different DINOV2 backbone layers in Figures S5, S6, S7, S8. These figures show the visual and semantic similarity between the ROIs as reflected in the learned queries for the subjects. We observed that ROIs with shared category selectivity form clusters (faces, places, bodies, or words) in the similarity matrix, exhibiting greater representational similarity within each category type.

We also see a clear divide between categorical and non-categorical areas. Additionally, ROIs within the ventral early visual areas (V1v, V2v, V3v) are more similar to one another than to their dorsal counterparts (V1d, V2d, V3d), and vice versa (the checkerboard patterns), reflecting the anatomical and functional organization of the visual cortex, and that the attention will be mostly driven by spatial information.

Figure S5: Cosine similarity between learned ROI queries for subject 1. Each entry in the matrix represents the average cosine similarity between the query for the ROI indicated by the row label and that indicated by the column label. ROIs from the left hemisphere are labeled with 'lh', and those from the right hemisphere with 'rh'. Results are averaged across 20 models, trained using five random seeds and four different backbone layers.
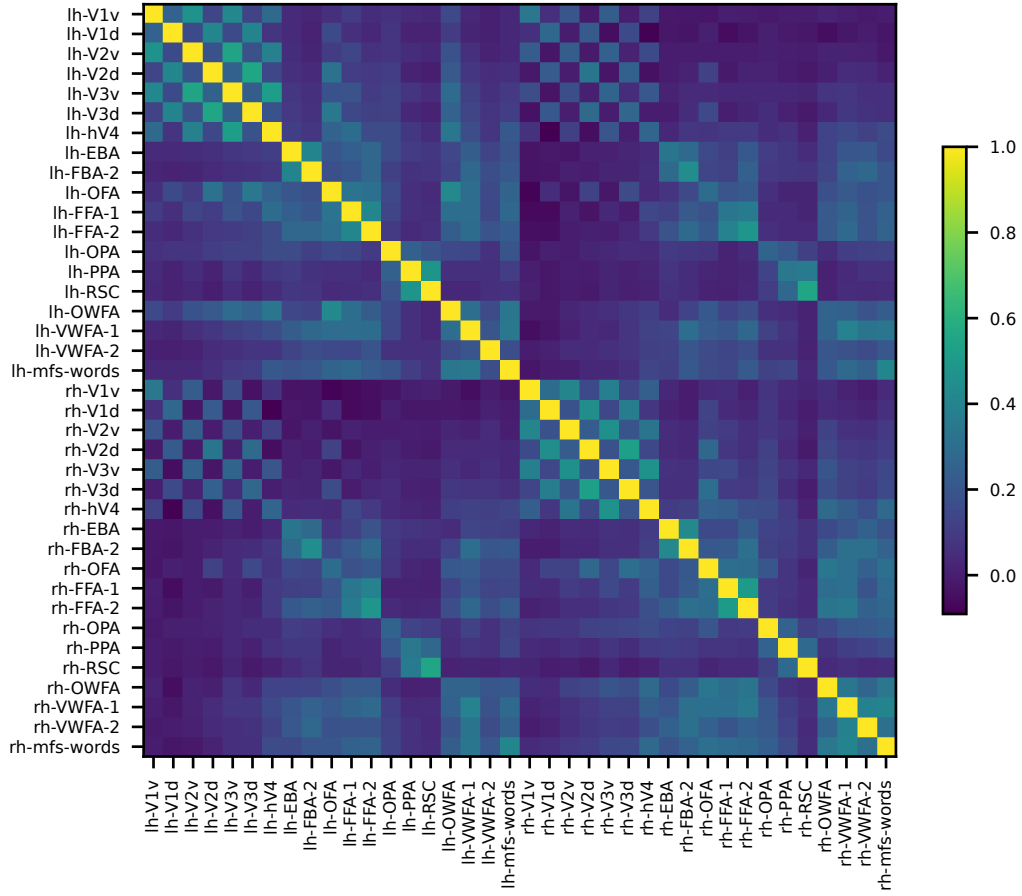
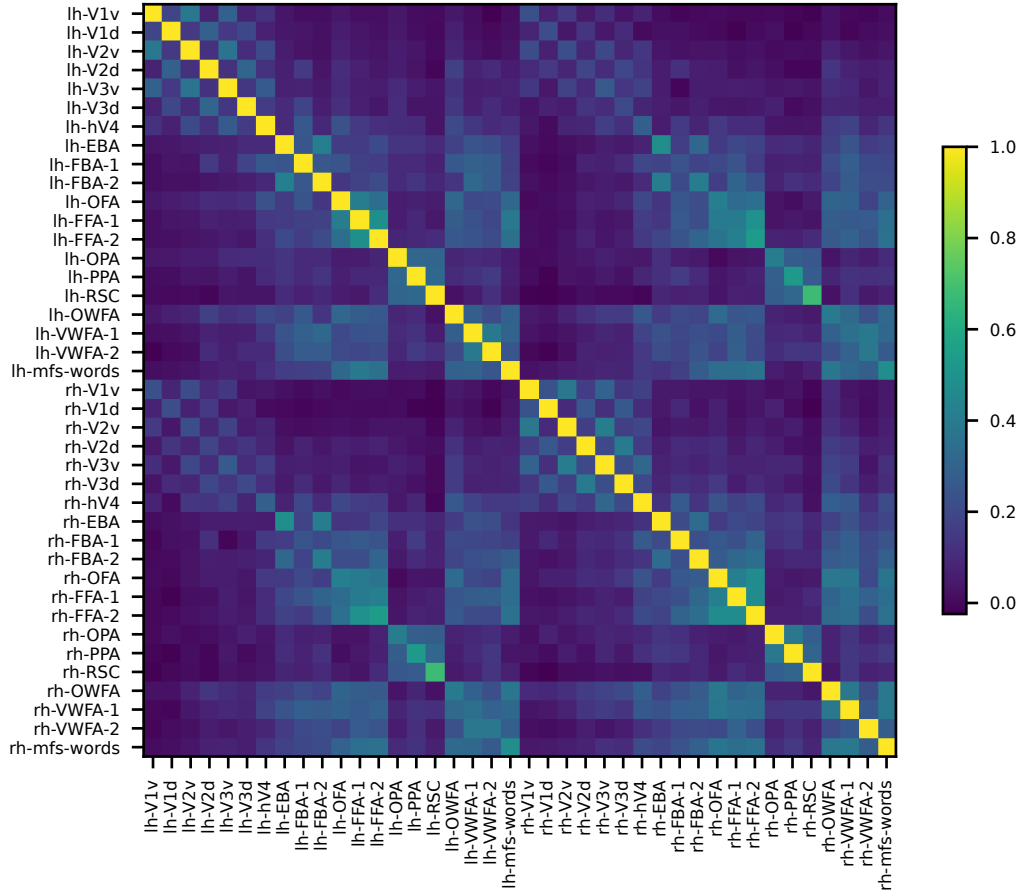Figure S6: Cosine similarity between learned ROI queries for subject 2.

Figure S7: Cosine similarity between learned ROI queries for subject 5.
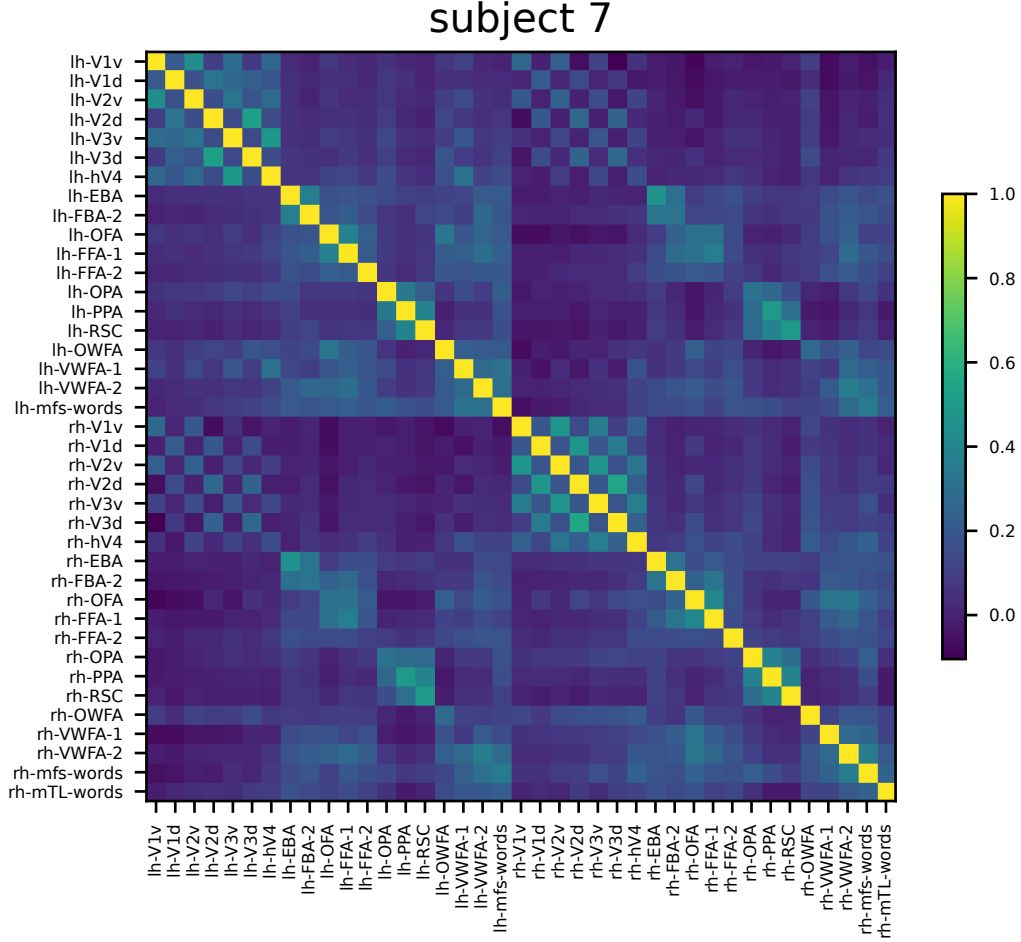
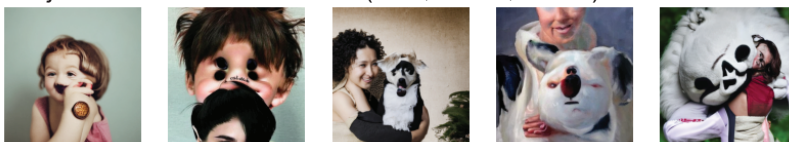Figure S8: Cosine similarity between learned ROI queries for subject 7.

A.3 Generating maximally activating images for ROIs

BrainDiVE [24] is a generative framework for synthesizing images predicted to activate specific regions of the human visual cortex. It guides the denoising steps of a diffusion model using gradients derived from a brain encoding model. Given the strong performance of our encoding model in predicting brain activity, we tested whether it could also effectively guide image generation within the BrainDiVE framework. We generated 200 images optimized to maximally activate the average predicted response of a specific ROI cluster, and display the top five in Figure S9, S10. The categories of the generated images are consistent with the reported category selectivity of each ROI cluster in the literature.

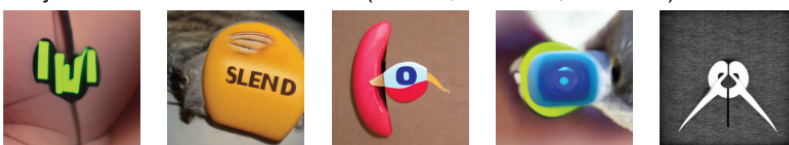subject 1: Body selective areas (EBA, FBA-1, FBA-2)

subject 1: Face selective areas (OFA, FFA-1, FFA-2)

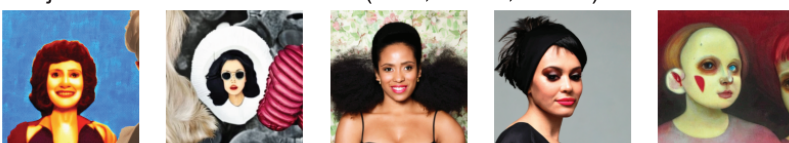subject 1: Place selective areas (OPA, PPA, RSC)

subject 1: Word selective areas (OWFA, VWFA-1, VWFA-2)

subject 2: Body selective areas (EBA, FBA-1, FBA-2)

subject 2: Face selective areas (OFA, FFA-1, FFA-2)

subject 2: Place selective areas (OPA, PPA, RSC)

subject 2: Word selective areas (OWFA, VWFA-1, VWFA-2)

Figure S9: Images generated to maximally activate different ROI clusters for subjects 1 and 2. Using our encoding model within the BrainDiVE framework, we generated 200 images predicted to maximally activate a specific ROI cluster for a given subject (indicated by the row titles). For each cluster, we display the top five images with the highest predicted activation, as determined by our encoding model.

subject 5: Body selective areas (EBA, FBA-1, FBA-2)

subject 5: Face selective areas (OFA, FFA-1, FFA-2)

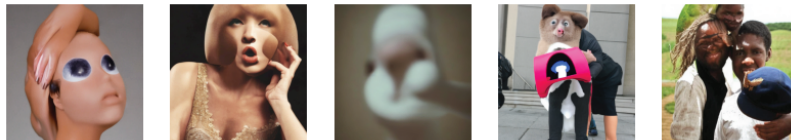subject 5: Place selective areas (OPA, PPA, RSC)

subject 5: Word selective areas (OWFA, VWFA-1, VWFA-2)

subject 7: Body selective areas (EBA, FBA-1, FBA-2)

subject 7: Face selective areas (OFA, FFA-1, FFA-2)

subject 7: Place selective areas (OPA, PPA, RSC)

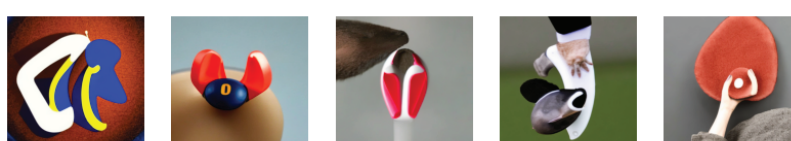
subject 7: Word selective areas (OWFA, VWFA-1, VWFA-2)

Figure S10: Images generated to maximally activate different ROI clusters for subjects 5 and 7