# Visio-Linguistic Brain Encoding

**Subba Reddy Oota**[1,2], **Jashn Arora**[2], **Vijay Rowtula**[2], **Manish Gupta**[2,3], **Raju S. Bapi**[2]
[1]Inria Bordeaux, France, [2]IIIT Hyderabad, India, [3]Microsoft, Hyderabad, India
subba-reddy.oota@inria.fr, {jashn.arora, vijay.rowtula}@research.iiit.ac.in,
gmanish@microsoft.com, raju.bapi@iiit.ac.in

## Abstract

Brain encoding aims at reconstructing fMRI brain activity given a stimulus. Earlier neural encoding models focused on brain encoding for single-mode stimuli: visual (pretrained CNNs) or text (pretrained language models). Few recent papers have also obtained separate visual and text representation models and performed late-fusion using simple heuristics. However, the human brain perceives the environment using information from multiple modalities, and previous works have not explored the co-attentive multi-modal encoding for visual and text reasoning. This paper systematically explores image and multi-modal Transformers' efficacy for brain encoding. Extensive experiments on two popular datasets, BOLD5000 and Pereira, provide the following insights. (1) We find that VisualBERT, a multi-modal Transformer, significantly outperforms previously proposed single-mode CNNs, image Transformers, and other previously proposed multi-modal models, thereby establishing new state-of-the-art. (2) The regions such as LPTG, LMTG, LIFG, and STS, which have dual functionalities for language and vision, have a higher correlation with multi-modal models, which reinforces the fact that these models are good at mimicking the human brain behavior. (3) The supremacy of visio-linguistic models raises the question of whether the responses elicited in the visual regions are affected implicitly by linguistic processing even when passively viewing images. Future fMRI tasks can verify this computational insight in an appropriate experimental setting. We make our code publicly available[1].

## 1 Introduction

Brain encoding aims at constructing neural brain activity recordings given an input stimulus. The two most studied forms of stimuli include vision and language. Since discovering of the relationship between language/visual stimuli and functions of brain networks (Constable et al., 2004; Thirion et al., 2006), researchers have been interested in understanding how the neural encoding models predict the fMRI (functional magnetic resonance imaging) brain activity. Recently, several brain encoding models were developed to (i) understand the ventral stream in biological vision (Yamins et al., 2014; Kietzmann et al., 2019; Bao et al., 2020) and (ii) study higher-level cognition like language processing (Gauthier and Levy, 2019; Schrimpf et al., 2020a; Schwartz et al., 2019). Previous work has mainly focused on independently understanding vision and text stimuli. However, the biological systems perceive the world by simultaneously processing high-dimensional inputs from diverse modalities such as vision, auditory, touch, and proprioception (Jaegle et al., 2021). In particular, how the brain effectively processes and provides its visual understanding through natural language and vice versa is still an open question in neuroscience.

Earlier studies mainly were related to neural encoding models that predict brain activity using representations of single-mode stimuli: visual or text. Convolutional neural networks (CNNs) were known to encode semantics from visual stimuli effectively. Interestingly, intermediate layers in deep CNNs trained on the ImageNet (Deng et al., 2009) categorization task can partially account for how neurons in intermediate layers of the visual system respond to any given image (Yamins et al., 2013, 2014; Güçlü and van Gerven, 2015; Yamins and DiCarlo, 2016; Wang et al., 2019). However, more recent and deeper CNNs did not further improve on measures of brain-likeness, even though their ImageNet performance has vastly increased (Russakovsky et al., 2015). Recently, Kubilius et al. (2019) proposed a shallow recurrent anatomical network, CORnet, which provided state-of-the-art results on the Brain-score (Schrimpf et al., 2020b) benchmark. Similar to CNN based visual encoding models, various studies leveraged

[1]https://tinyurl.com/VLBEncoding



*Proceedings of the 29th International Conference on Computational Linguistics*, pages 116–133
October 12–17, 2022.

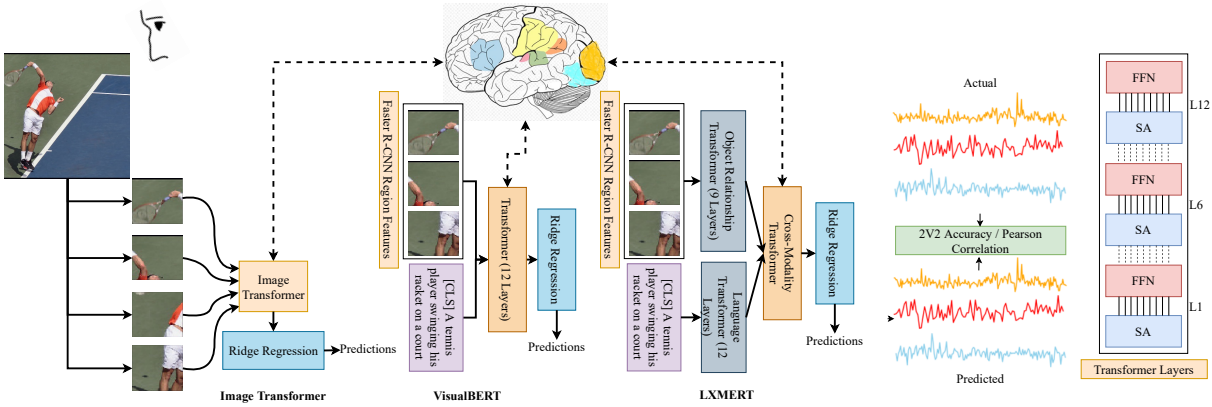


Fig. 1: Logical architecture of the proposed approach: We use features from image/multi-modal Transformers (like ViT, VisualBERT, and LXMERT) as input to the regression model to predict the fMRI activations for different brain regions. We evaluate the brain encoding results by computing 2V2 accuracy and Pearson correlation between actual and predicted activations. We also perform layer-wise correlation analysis between transformer layers and brain regions.

neural models like deep recurrent neural networks (RNNs), Transformer (Vaswani et al., 2017) based language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT-2 (Radford et al., 2019) to predict the brain activity corresponding to semantic vectors of linguistic items, including words, phrases, sentences, and paragraphs (Gauthier and Levy, 2019; Schrimpf et al., 2020a).

Brain encoding for more brain regions is vital since input stimuli elicit diverse and distributed representations in the brain. These activation responses could be internally repurposed for several other brain tasks. Although previous neural encoding models have demonstrated promising results for processing one of the two brain regions (visual cortex V4 and prefrontal cortex IT), more efforts are needed to improve brain encoding for other parts of the brain. Further, previous works manually choose[2] particular CNN layers, whose activations were used for predicting brain activity specific to the datasets they work with (Kubilius et al., 2019). Applying such methods to other datasets will need dataset-specific, time-consuming manual identification of the best layer. We observe in our experiments that using *last layer* activations from VisualBERT leads to the best accuracy.

Unlike previous studies, which focus on single-modality (visual or language stimuli), some authors demonstrated that multi-modal models formed by combining text-based distributional information with visual representations provide a better proxy for human-like intelligence (Anderson et al., 2015; Oota et al., 2019). However, these methods extract representations from each mode separately (image features from CNNs and text features from pretrained embeddings) and then perform a simple late-fusion. Thus, they cannot effectively exploit semantic correspondence across the two modes at different levels. Such late-fusion-based multi-modal models are the closest to our work, and our experiments show that our models outperform them.

Recently, Transformer-based models were found to be very effective than CNNs, in all language and image-related tasks (Devlin et al., 2019). Image-based transformer models like ViT (Dosovitskiy et al., 2020), DEiT (Touvron et al., 2021), and BEiT (Bao et al., 2021) have been shown to provide excellent results compared to traditional CNNs on image classification tasks. Also, multi-modal Transformers like VisualBERT (Li et al., 2019), LXMERT (Tan and Bansal, 2019), and CLIP (Radford et al., 2021) have shown excellent results on visio-linguistic tasks like visual question answering, visual common-sense reasoning. Inspired by the success of language, image, and multi-modal Transformers, we build multi-modal transformer models to learn the joint representations of image content and natural language and use them for brain encoding. Overall, in this work, we investigate whether *image-based and multi-modal Transformers* can accurately perform fMRI encoding on the *whole brain*. Fig. 1 illustrates our method for brain encoding.

Specifically, we make the following contributions to this paper. (1) We present state-of-the-art

[2]Quoting from (Kubilius et al., 2019): "After testing every layer on both V4 and IT, we report the model's score as the score of the best layer per region."

brain encoding results using multi-modal Transformers. We also study the effectiveness of our models in a cross-data setting. (2) Our approach generalizes the use of Transformer-based architectures, removing the need to manually select specific layers as in existing CNN-based fMRI encoding architectures. (3) We uncover several cognitive insights about the association between fMRI voxels and representations of multi-modal/image Transformers and CNNs.

## 2 Brain Imaging Datasets

The following datasets are popularly used in the literature for studying brain encoding: Vim-1 (Kay et al., 2008), Harry Potter (Wehbe et al., 2014; Pereira et al., 2018), BOLD5000 (Chang et al., 2019), Algonauts (Cichy et al., 2019), and SS-fMRI (Beliy et al., 2019). Vim-1 has only black and white images, is only related to object recognition, and is subsumed by BOLD5000. SS-fMRI is smaller and very similar to BOLD5000. The Harry Potter dataset does not have images. Lastly, fMRIs have not been made publicly available for the Algonauts dataset. Hence, we experiment with BOLD5000 and Pereira Pereira et al. (2018) datasets in this work.

**BOLD5000**: BOLD5000 dataset was collected from four subjects where three subjects viewed 5254 natural images (ImageNet: 2051, COCO: 2135, Scenes: 1068) while fMRIs were acquired. The fourth subject was shown 3108 images only. Details of the visual stimuli and fMRI protocols of the dataset have been discussed in (Chang et al., 2019). We briefly summarize the details of the dataset in Table 1. The data covers five visual areas in the human visual cortex, i.e., the early visual area (EarlyVis); object-related areas such as the lateral occipital complex (LOC); and scene related areas such as the occipital place area (OPA), the parahippocampal place area (PPA), and the retrosplenial complex (RSC). Each image also has corresponding text labels: ImageNet has a few out of 1000 possible tags per image, COCO has five captions per image, and Scenes have one out of 250 possible categories per image.

**Pereira**: For the Pereira dataset Pereira et al. (2018), participants were shown a concept word and a picture to observe brain activation when participants retrieved relevant meaning using visual information. Sixteen subjects were presented images (six per concept) corresponding to 180 concepts (abstract + concrete), while fMRIs were acquired. Out of 180 concepts, 116 are concrete, and others are abstract. Here, we augmented the image captions using the concept word associated with each image in the picture view. As in (Pereira et al., 2018), we focused on nine brain regions corresponding to four brain networks: Default Mode Network (DMN) (linked to the functionality of semantic processing), Language Network (related to language processing, understanding, word meaning, and sentence comprehension, Task Positive Network (related to attention, salience information), and Visual Network (related to the processing of visual objects, object recognition).

We show number of instances and voxel distribution across various brain regions for the BOLD5000 and Pereira datasets in Tables 1 and 2 respectively.

| | | Number of Voxels in Each ROI | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ROIs→ | | PPA | | LOC | | EarlyVis | | OPA | | RSC | |
| ↓Subjects | #Instances | LH | RH | LH | RH | LH | RH | LH | RH | LH | RH |
| Subject-1 | 5254 | 131 | 200 | 152 | 190 | 210 | 285 | 101 | 187 | 86 | 143 |
| Subject-2 | 5254 | 172 | 198 | 327 | 561 | 254 | 241 | 85 | 95 | 59 | 278 |
| Subject-3 | 5254 | 112 | 161 | 430 | 597 | 522 | 696 | 187 | 205 | 78 | 116 |
| Subject-4 | 3108 | 157 | 187 | 455 | 417 | 408 | 356 | 279 | 335 | 51 | 142 |

Table 1: BOLD5000 Dataset Statistics. LH=Left Hemisphere. RH=Right Hemisphere.

| | Number of Voxels in Each ROI | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ROIs→ | Language | | Vision | | | | | DMN | Task Positive |
| ↓Subj | LH | RH | Body | Face | Object | Scene | Vision | RH | LH |
| P01 | 5265 | 6172 | 3774 | 4963 | 8085 | 4141 | 12829 | 17190 | 35120 |
| M01 | 5716 | 5561 | 3934 | 4246 | 7357 | 3606 | 12075 | 17000 | 34582 |
| M02 | 4930 | 5861 | 3873 | 4782 | 7552 | 3173 | 11729 | 15070 | 30594 |
| M03 | 3616 | 4247 | 2838 | 3459 | 5956 | 2822 | 9074 | 12555 | 24486 |
| M04 | 5906 | 5401 | 3867 | 4803 | 7812 | 3602 | 12278 | 18011 | 34024 |
| M05 | 4607 | 4837 | 2961 | 4023 | 6609 | 3135 | 10417 | 14096 | 28642 |
| M06 | 4993 | 5099 | 3424 | 4374 | 7300 | 4058 | 11986 | 16289 | 30109 |
| M07 | 5629 | 5001 | 4190 | 4993 | 8617 | 3721 | 12454 | 17020 | 30408 |
| M08 | 5083 | 5062 | 2624 | 4082 | 6463 | 3503 | 10439 | 14950 | 29972 |
| M09 | 3513 | 3650 | 2876 | 3343 | 5992 | 2815 | 9003 | 12469 | 25167 |
| M10 | 5458 | 5581 | 3232 | 4844 | 7445 | 3474 | 11530 | 16424 | 29400 |
| M13 | 4963 | 4811 | 2675 | 4008 | 5809 | 3323 | 9848 | 14489 | 30608 |
| M15 | 5315 | 6141 | 4112 | 4941 | 8323 | 3496 | 12383 | 15995 | 31610 |
| M16 | 4726 | 5534 | 4141 | 4669 | 8060 | 4142 | 12503 | 15104 | 31758 |
| M17 | 5854 | 5698 | 4416 | 4801 | 8831 | 4521 | 13829 | 16764 | 37463 |

Table 2: Pereira Dataset Statistics. LH=Left Hemisphere. RH=Right Hemisphere.

## 3 Task Descriptions

We train fMRI encoding models using Ridge regression on stimuli representations obtained using various models for both datasets, as shown in Fig. 1. The main goal of each fMRI encoder model is to predict fMRI voxel values for each brain region given stimuli. In all cases, we train a model per subject separately. Different brain regions are involved in processing stimuli involving objects and scenes. Similarly, some regions specialize in understand-

ing vision inputs while others interpret linguistic stimuli better.

To evaluate the generalizability of our models across objects vs. scenes understanding, we also perform cross-data experiments where the train images belong to one sub-dataset, and the test images belong to the other sub-dataset. Thus, for each subject, we perform (1) three same-sub-dataset train-test experiments and (2) six cross-sub-dataset train-test experiments.

**Full dataset fMRI Encoding**: Whenever we train and test on the same dataset, we follow K-fold (K=10) cross-validation. All the data samples from K-1 folds were used for training, and the model was tested on samples of the left-out fold.

**Cross-data fMRI Encoding**: In the BOLD5000 dataset, we have three sub-datasets: COCO, ImageNet, and Scenes. ImageNet images mainly contain objects. Scenes images are about natural scenes, while COCO images relate to both objects and scenes. For each of the three sub-datasets, we perform K-fold (K=10) cross-validation within the sub-dataset.

## 4 Methodology

We trained a ridge regression-based encoding model to predict the fMRI brain activity associated with the stimuli representation for each brain region. Each voxel value is predicted using a separate ridge regression model. Formally, we encode the stimuli as $X \in \mathbb{R}^{N \times D}$ and brain region voxels $Y \in \mathbb{R}^{N \times V}$, where $N$ denotes the number of training examples, $D$ denotes the dimension of input stimuli representation, and $V$ denotes the number of voxels in a particular region. Although ridge regression is a very naïve way of modeling, it has been the most popular brain encoding technique in this line of work. We plan to experiment with other forms of regression methods in the future.

The input stimuli representation can be obtained using any of the following models: (i) pretrained CNNs, (ii) pretrained text Transformers (iii) image Transformers, (iv) late-fusion models, or (v) multi-modal Transformers. The ridge regression objective function for the $i^{th}$ example is given as follows.

$$f(X_i) = \min_W \|Y_i - X_i W\|_F^2 + \lambda \|W\|_F^2$$

Here, $W$ are the learnable weight parameters, $\|.\|_F$ denotes the Frobenius norm, and $\lambda > 0$ is a tunable hyper-parameter representing the regularization weight. $\lambda$ was tuned on a small disjoint validation set obtained from the training.

Next, we discuss different input stimuli representation methods. Pretrained CNNs and Image Transformers encode image stimuli only, while Pretrained text Transformers encode text stimuli only. Late fusion models and Multi-modal Transformers encode both text and image stimuli.

**Pretrained CNNs**: Inspired by the Algonauts challenge (Cichy et al., 2019), we extract the layer-wise features from different pretrained CNN models such as VGGNet19 (Simonyan and Zisserman, 2014) (MaxPool1, MaxPool2, MaxPool3, MaxPool4, MaxPool5, FC6, FC7, FC8), ResNet50 (He et al., 2016) (Block1, Block2, Block3, Block4, FC), InceptionV2ResNet (Szegedy et al., 2017) (Conv2D5, Conv2D50, Conv2D100, Conv2D150, Conv2D200, Conv2D_7b), and EfficientNetB5 (Tan and Le, 2019) (Conv2D2, Conv2D8, Conv2D16, Conv2D24, FC), and use them for predicting fMRI brain activity. We use adaptive average pooling on each layer to get features for each image.

**Pretrained text Transformers**: RoBERTa (Liu et al., 2019) builds on BERT's language masking strategy and has been shown to outperform several other text models on the popular GLUE NLP benchmark. We use the average-pooled representation[3] from RoBERTa to encode text stimuli.

**Image Transformers**: We used three image Transformers: Vision Transformer (ViT), Data Efficient Image Transformer (DEiT), and Bidirectional Encoder representation from Image Transformer (BEiT). Given an image, image Transformers output two representations: pooled and patches. We experiment with both representations.

**Late-fusion models**: In these models, the stimuli representation is obtained as a concatenation of image stimuli encoding obtained from pretrained CNNs and text stimuli encoding obtained from pretrained text Transformers. Thus, we experiment with these late-fusion models: VGGNet19+RoBERTa, ResNet50+RoBERTa, InceptionV2ResNet+RoBERTa and EfficientNetB5+RoBERTa. These models do not incorporate real information fusion but do concatenation across modalities.

**Multi-modal Transformers**: We experiment with these multi-modal Transformer models: Contrastive Language-Image Pre-training (CLIP),

[3] Average-pooled representation gave us better results compared to using the CLS representation.

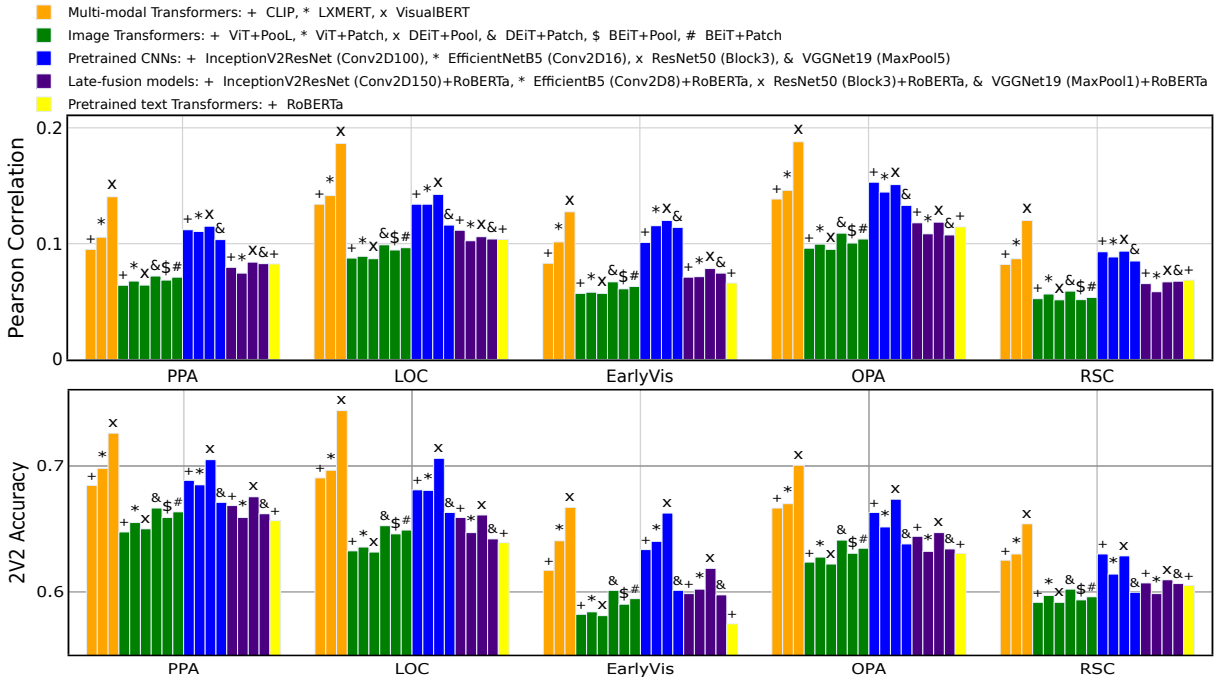


Fig. 2: BOLD5000 Results: Pearson correlation coefficient (top figure) and 2V2 (bottom figure) between predicted and actual responses across different brain regions using various models. Results are averaged across all participants. VisualBERT performs the best.

Learning Cross-Modality Encoder Representations from Transformers (LXMERT), and VisualBERT. These Transformers take both image and text stimuli as input and output a joint visio-linguistic representation. Specifically, the image input for these models comprises region proposals as well as bounding box regression features extracted from Faster R-CNN (Ren et al., 2015) as input features. These models incorporate information fusion across modalities at different levels of processing using co-attention and hence are expected to result in high-quality visio-linguistic representations.

**Hyper-parameter Settings**: We used sklearn's ridge-regression with default parameters, 10-fold cross-validation, Stochastic-Average-Gradient Descent Optimizer, Huggingface for Transformer models, MSE loss function, and L2-decay ($\lambda$) as 1.0. We used Word-Piece tokenizer for the linguistic Transformer input and Faster-RCNN (Ren et al., 2015) for extracting region proposals. All experiments were conducted on a machine with 1 NVIDIA GEFORCE-GTX GPU with 16GB GPU RAM. We make our code publicly available[1].

## 5 Experiments

### 5.1 Evaluation Metrics

We evaluate our models using popular brain encoding evaluation metrics described in the following. Given a subject and a brain region, let $N$ be the number of samples. Let $\{Y_i\}_{i=1}^{N}$ and $\{\hat{Y}_i\}_{i=1}^{N}$ denote the actual and predicted voxel value vectors for the $i^{th}$ sample. Thus, $Y \in R^{N \times V}$ and $\hat{Y} \in R^{N \times V}$ where $V$ is the number of voxels in that region.

$$\text{2V2 Accuracy} = \frac{1}{N_{C_2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} I[\{cosD(Y_i, \hat{Y}_i) + cosD(Y_j, \hat{Y}_j)\} < \{cosD(Y_i, \hat{Y}_j) + cosD(Y_j, \hat{Y}_i)\}]$$

where $cosD$ is the cosine distance function. $I[c]$ is an indicator function such that $I[c] = 1$ if $c$ is true, else it is 0. The higher the 2V2 accuracy, the better. Pearson Correlation (PC) is computed as PC=$\frac{1}{N}\sum_{i=1}^{n} corr[Y_i, \hat{Y}_i]$ where corr is the correlation function.

### 5.2 Do multi-modal Transformers outperform other models?

Unfortunately, no previous work uses image Transformers or multi-modal Transformers for brain encoding. StepEnCog (Oota et al., 2019) is a late-fusion method, but it has a different setting where the model expects voxel values per brain slice rather than per brain region. Besides performing extensive evaluation using a large variety of models, we also compare our results with those obtained by two previously proposed baselines that lever-

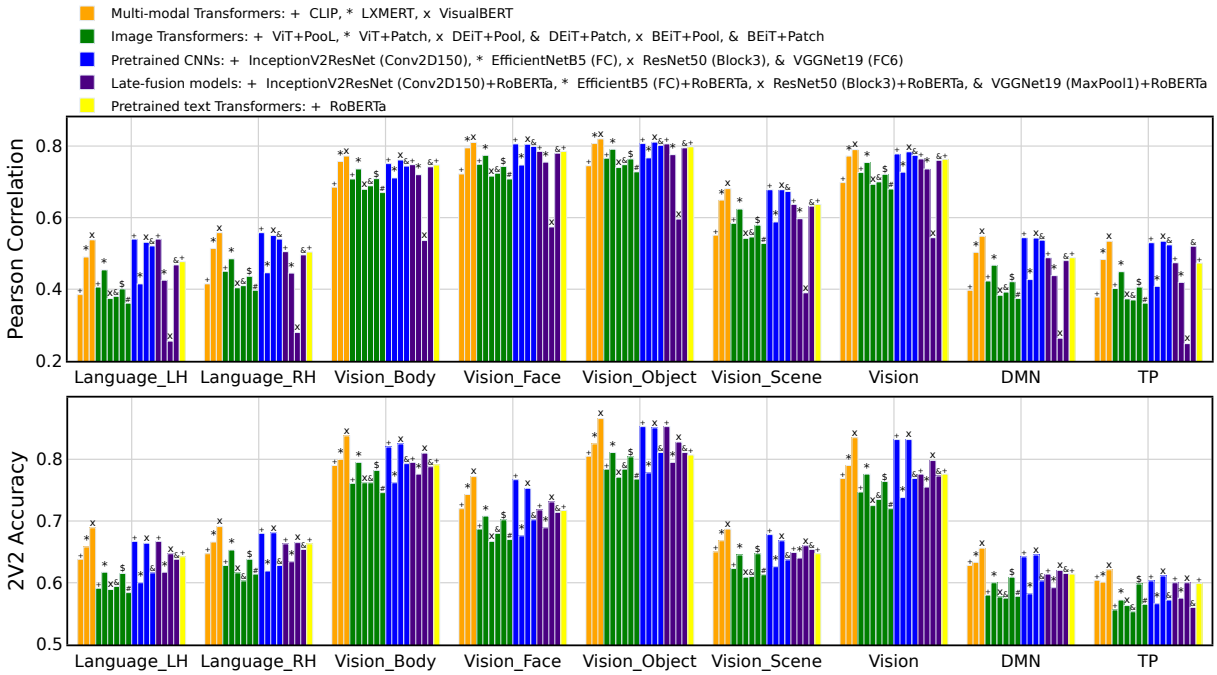


Fig. 3: Pereira Results: Pearson correlation coefficient (top figure) and 2V2 (bottom figure) between predicted and actual responses across different brain regions using a variety of models. Results are averaged across all participants. VisualBERT performs the best.

age pretrained CNN models: (Blauch et al., 2019) and (Wang et al., 2019) which use VGGNet.

We present the 2V2 accuracy and Pearson correlation results for models trained with different input representations (extracted from the best-performing layer of every pretrained CNN model and the last output layer of the Transformer model) on the two datasets: BOLD5000 and Pereira in Figs. 2 and 3, respectively. We also compare the results using many intermediate layer activations (not just the best) for CNN models and the last layer of Transformer models in Figs. 9 and 10 in the Appendix. Further, we also compare the results using all intermediate layer activations for Transformer models in Figs. 11 and 12 in the Appendix.

**BOLD5000:** We make the following observations from Fig. 2: (1) On both 2V2 accuracy and Pearson correlation, VisualBERT is better across all the models. (2) Other multi-modal Transformers such as LXMERT and CLIP perform as good as pretrained CNNs. We observed that image Transformers perform worse than pretrained CNNs. Late fusion models and RoBERTa has the least performance. (3) Late visual areas such as OPA (scene-related) and LOC (object-related) display a higher Pearson correlation with multi-modal Transformers, which is in line with the visual processing hierarchy. A higher correlation with all the visual brain ROIs with multi-modal Transformers demonstrates the power of jointly encoding visual and language information. (4) The patch representation of image Transformers shows an improved 2V2 accuracy and Pearson correlation compared to the Pooled representation. (5) Both InceptionV2ResNet and ResNet-50 have better performance among uni-modality models.

In order to estimate the statistical significance of the performance differences, we performed *post hoc* pairwise comparisons for all the subjects across the five brain ROIs. We found that VisualBERT is significantly better than LXMERT (second-best multi-modal Transformer) and InceptionV2ResNet (best pretrained CNN) for all ROIs except EarlyVis. Lastly, InceptionV2ResNet is significantly better than BEiT (best image Transformer) for all ROIs. Detailed p-values are mentioned in Table 3.

**Pereira:** We make the following observations from Fig. 3: (1) Similar to BOLD5000, multi-modal Transformers such as VisualBERT and LXMERT perform better. (2) Lateral visual areas such as

| Models compared | PPA | LOC | EarlyVis | OPA | RSC |
|---|---|---|---|---|---|
| VisualBERT vs. LXMERT | 0.044* | 0.004* | 0.076 | 0.049* | 0.029* |
| VisualBERT vs. InceptionV2ResNet | 0.049* | 0.032* | 0.521 | 0.041* | 0.0354* |
| InceptionV2ResNet vs. BEiT | 0.041* | 0.003* | 0.014* | 0.188 | 0.203 |

Table 3: p-values for *post hoc* pairwise comparisons for BOLD5000 dataset

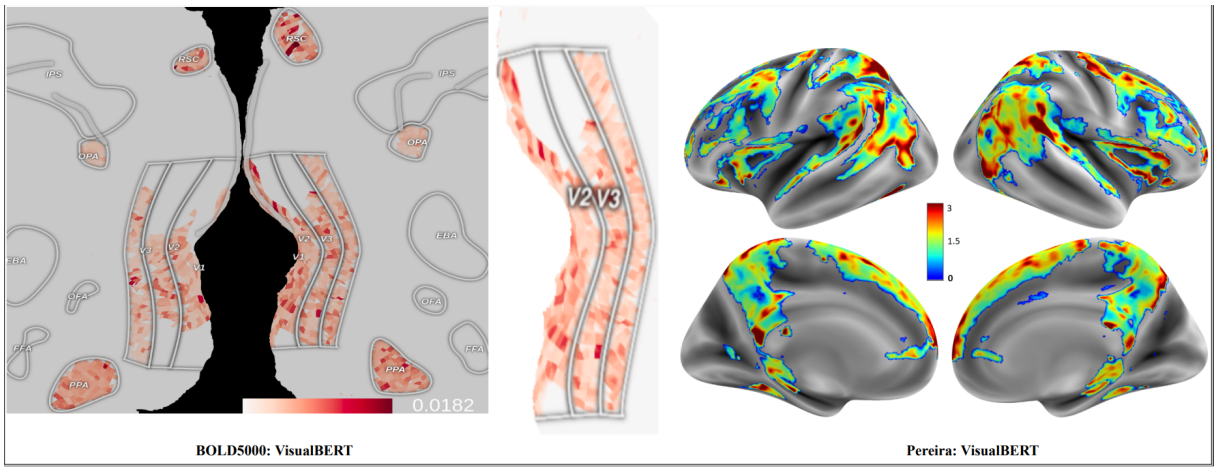


Fig. 4: MAE between actual and predicted voxels: (a) left figure is zoomed on V2 and V3 brain areas for VisualBERT on BOLD5000 subject 1. Note that V1 and V2 are also called EarlyVis areas, while V3 is also called LOC area. (b) the right figure is for VisualBERT on the Pereira dataset subject 2.

| Models compared | Language_LH | Language_RH | Vision_Body | Vision_Face | Vision_Object | Vision_Scene | Vision | DMN | TP |
|---|---|---|---|---|---|---|---|---|---|
| VisualBERT vs. LXMERT | 0.046* | 0.039* | 0.052 | 0.048* | 0.046* | 0.045* | 0.047* | 0.040* | 0.035* |
| VisualBERT vs. ResNet | 0.049* | 0.038* | 0.048* | 0.048* | 0.078 | 0.217 | 0.048* | 0.046* | 0.049* |
| ResNet vs. ViT | 0.009* | 0.043* | 0.041* | 0.047* | 0.046* | 0.042* | 0.038* | 0.022* | 0.023* |

Table 4: p-values for *post hoc* pairwise comparisons for Pereira dataset

Vision_Object, Vision_Body, Vision_Face, and Vision areas display higher correlation with multi-modal Transformers. A higher correlation with all the visual brain regions, language regions, DMN, and TP with multi-modal Transformers, demonstrates that the alignment of visual-language understanding helps.

In order to estimate the statistical significance of the performance differences, we performed *post hoc* pairwise comparisons for all the subjects across the nine brain ROIs. We found that VisualBERT is statistically significantly better than LXMERT (second-best multi-modal Transformer) for all ROIs except Vision_Body. Further, VisualBERT is statistically significantly better than ResNet (best pretrained CNN) for all ROIs except Vision_Object and Vision_Scene. Lastly, ResNet is statistically significantly better than ViT (best image Transformer) for all ROIs. Detailed p-values are mentioned in Table 4.

As further analysis, in Fig. 4, we show the mean absolute error (MAE) between the actual and predicted voxels across brain regions using VisualBERT. Comparing with similar brain charts for other models (shown in Figs. 13 and 14 in the Appendix), we notice that the error magnitudes are very small for the majority of the voxels. We observe that MAE values are relatively higher for EarlyVis areas and lowest for OPA for BOLD5000.

### 5.3 Model size vs. Efficacy Comparison

We plot a comparison of model size with Pearson Correlation averaged across all subjects for BOLD5000 in Fig. 5. Compared to LXMERT, VisualBERT is not just more accurate but also much smaller. VisualBERT is much more accurate than image Transformers while being almost the same size. Lastly, pretrained CNNs are smaller than VisualBERT but are less accurate even when the particular layer activations are cherry-picked. We observe similar trends for the Pereira dataset, as shown in Fig. 6. We hope that smaller models can be helpful for faster fine-tuning of new datasets.

### 5.4 Single Stream vs. Dual Stream Models

Since single stream (VisualBERT) and dual-stream (CLIP, LXMERT, and ViLBERT) models fuse language and images at different times. We report

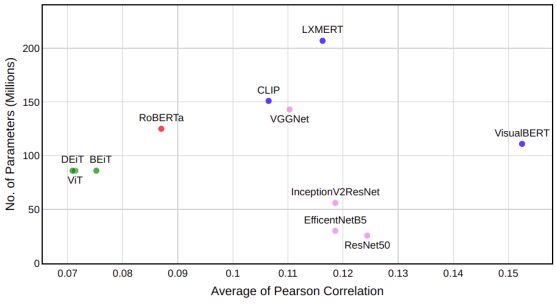


Fig. 5: BOLD5000: #Parameters vs. Avg Pearson Corr.

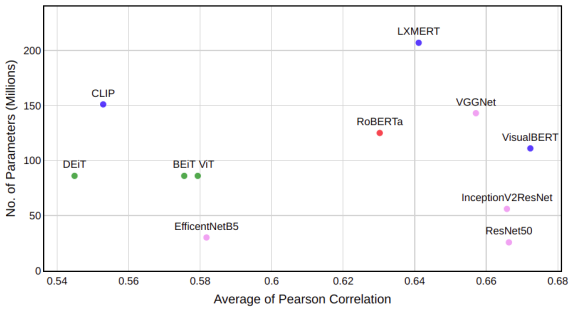


Fig. 6: Pereira: #Parameters vs. Avg Pearson Corr.

the comparison of single-stream vs. dual-stream with Pearson Correlation (PC) averaged across all subjects for BOLD5000 in Table 5 (top). Compared to dual-stream models (CLIP, LXMERT, and ViLBERT), VisualBERT showcases much better performance.

### 5.5 Is Linguistic Information Important in Multi-Modal Transformers?

Is the improvement in prediction performance of vision+language models over vision-only models due to the added linguistic information? For example, what happens if we randomize the language captions in BOLD5000, feed the model the correct image with a wrong caption, and train the encoding model to predict the correct image-elicited brain recording? We report the comparison of multi-modal transformers with correct caption vs. random caption using Pearson Correlation averaged across all subjects for BOLD5000 in Table 5. We observe that linguistic information is crucial to better performance with multi-modal Transformers.

### 5.6 Cross-Data fMRI Encoding

Fig. 7 illustrates PC for cross-data encoding on BOLD5000 using three multi-modal Transformers (VisualBERT, LXMERT, and CLIP). We also show results for a baseline method (Blauch et al., 2019). We observe that (1) multi-modal Transformers outperform the baseline results across all the five brain regions for cross-data tasks. (2) PC score is higher for the model trained on COCO and tested on ImageNet in the object-selective visual area LOC (lateral occipital cortex), which makes sense since COCO has many objects. (3) Similarly, the scene-selective brain areas such as RSC and OPA have a higher correlation for the COCO-Scenes, ImageNet-Scenes, and Scenes-Scenes tasks. (4) EarlyVisual areas have a lower correlation than other brain regions across the three tasks. (5) Overall, the models trained on COCO or ImageNet report a higher correlation than those trained on Scenes.

| Models compared | PPA | LOC | EarlyVis | OPA | RSC |
|---|---|---|---|---|---|
| CLIP | 0.095 | 0.134 | 0.083 | 0.139 | 0.082 |
| LXMERT | 0.106 | 0.142 | 0.102 | 0.146 | 0.087 |
| VisualBERT | **0.141** | **0.187** | **0.128** | **0.188** | **0.12** |
| ViLBERT | 0.057 | 0.078 | 0.052 | 0.087 | 0.045 |
| CLIP-Random | 0.020 | 0.024 | 0.033 | 0.031 | 0.002 |
| LXMERT-Random | 0.035 | 0.041 | 0.035 | 0.049 | 0.029 |
| VisualBERT-Random | 0.072 | 0.102 | 0.062 | 0.109 | 0.060 |
| ViLBERT-Random | 0.018 | 0.011 | 0.013 | 0.017 | 0.017 |

Table 5: Single stream (VisualBERT) vs. Dual stream (CLIP, LXMERT, and ViLBERT) models with BOLD5000: Pearson correlation computed between predicted and actual responses across different brain regions. Results are averaged across all participants. VisualBERT performs the best. The bottom four rows display the model performance when a random-caption is provided with the correct image as input.

## 6 Cognitive Insights: Does Language Influence Vision?

BOLD5000 dataset comprises brain responses from visual areas (early visual, scene-related, and object-related) when visual stimuli are presented to the subjects. Although only visual information is present in the stimuli, it is conceivable that participants implicitly invoke appropriate linguistic representations that, in turn, influence visual processing (Lupyan et al., 2020). Thus, it is not surprising that computational models such as multi-modal Transformers (VisualBERT, and LXMERT) that learn a joint representation of language and vision show superior performance on the 'purely' visual response data in BOLD5000 (see Figs. 2 and 4(a)).

Further, the performance of these models is naturally good in the case when text and image are shown to the participants, and whole-brain responses are captured as in the case of the Pereira dataset (see Figs. 3 and 4(b)). We further investigate the role of different sub-ROIs of the language and visual networks. For this, we compare the predicted responses of the best encoding model, i.e., VisualBERT, with the ground truth (observed) responses of various language and visual sub-regions (see Fig 8). We notice that the classical language areas in the temporal gyrus (LMTG and LPTG) and the inferior frontal gyrus (LIFG) are more accurately predicted than the other sub ROIs of the language network. These sub-ROIs (LMTG, LPTG, and LIFG) are highly involved in language comprehension and semantic processing. Interestingly, the second-best correlations are seen for multi-modal

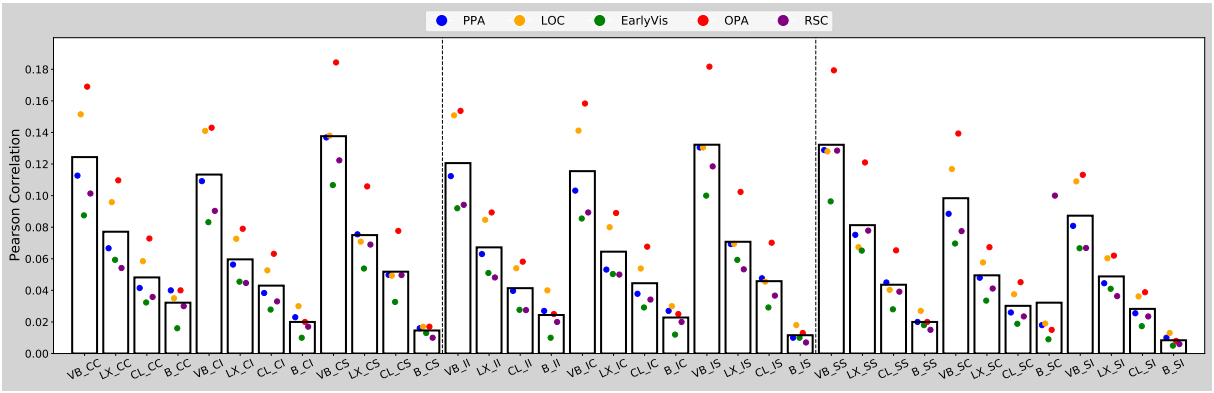


Fig. 7: Cross-Data Results for BOLD5000 dataset. VB=VisualBERT, LX=LXMERT, CL=CLIP, B=Baseline (Blauch et al., 2019), INC=InceptionV2ResNet. CC=Train and test on COCO, CI=Train on COCO and test on ImageNet, CS=Train on COCO and test on Scenes.)

integration areas in the temporo-parietal regions (LAngG, LFus, LPar) and higher-order processing and attention-related areas in the middle frontal region (MFG).

In the visual sub-ROIs (Fig. 8), we observe that the superior temporal sulcus (bilaterally but more in the left: LSTS) is more accurately predicted than other sub-ROIs. Surprisingly, LSTS is implicated in various social processes, ranging from language perception to simulating the mental processes of others. Also, the sub-ROIs such as LLOC, LFFA, LOFA, and LEBA have a higher correlation. These areas are involved in more visual-related functions such as object recognition, face perception, face recognition, and body recognition.

Based on the intuition from the computational experiments, we make the following testable prediction for future fMRI experiments. Instead of a passive viewing task, if participants were to perform a naming task/decision-making task on the objects/scenes, we expect to see more pronounced and focused activation in the visual areas during the language-based task compared to passive viewing.

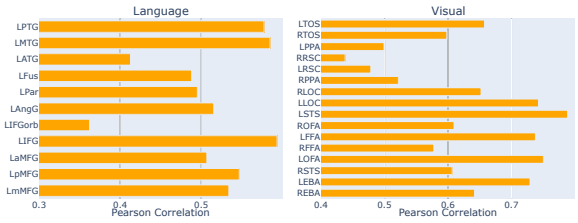


Fig. 8: Pearson correlation results across the Language Sub ROIs and Visual Sub ROIs for Pereira dataset, between predicted and true responses across different brain regions using a variety of models. Results are averaged across all participants and are obtained using VisualBERT.

## 7 Conclusion

We studied the effectiveness of multi-modal modeling for brain encoding. We found that VisualBERT, which jointly encodes text and visual input using cross-modal attention at multiple levels, performs the best. Our experiments on BOLD5000 and Pereira datasets lead to interesting cognitive insights. These insights indicate that fMRIs reveal reliable responses in scenes and object selection visual brain areas, which shows that cross-view decoding tasks like image captioning or image tagging are practically possible with reasonable accuracy. We plan to explore this as part of future work. We also plan to explore correlations between brain voxel space and representational feature space in the future. Finally, the combined strength of joint (audio, vision, and text) modalities remains to be investigated.

## 8 Ethical Statement

We reused publicly available datasets for this work: BOLD5000 and Pereira. We did not collect any new dataset. BOLD5000 dataset, except the stimulus images and their original annotations, is licensed under a Creative Commons 0 License. Please read their terms of use[4] for more details. Pereira dataset can be downloaded from https://osf.io/crwz7/. Please read their terms of use[5] for more details. We do not foresee any harmful uses of this technology.

[4] https://bold5000-dataset.github.io/website/terms.html

[5] https://github.com/CenterForOpenScience/cos.io/blob/master/TERMS_OF_USE.md

## References


Andrew James Anderson, Elia Bruni, Alessandro Lopopolo, Massimo Poesio, and Marco Baroni. 2015. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*, 120:309–322.

Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. 2020. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108.

Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. 2019. From voxels to pixels and back: self-supervision in natural-image reconstruction from fmri. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6517–6527.

Nicholas M Blauch, Filipe De Avila Belbute Peres, Juhi Farooqui, Alireza Chaman Zar, David Plaut, and Marlene Behrmann. 2019. Assessing the similarity of cortical object and scene representations through cross-validated voxel encoding models. *Journal of Vision*, 19(10):188d–188d.

Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. 2019. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):1–18.

Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. 2019. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv e-prints*, pages arXiv–1905.

R Todd Constable, Kenneth R Pugh, Ella Berroya, W Einar Mencl, Michael Westerveld, Weijia Ni, and Donald Shankweiler. 2004. Sentence complexity and input modality effects in sentence comprehension: an fmri study. *NeuroImage*, 22(1):11–21.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539.

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

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

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. 2008. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355.

Tim C Kietzmann, Courtney J Spoerer, Lynn KA Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. 2019. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. 2019. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in Neural Information Processing Systems*, 32:12805–12816.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Gary Lupyan, Rasha Abdel Rahman, Lera Boroditsky, and Andy Clark. 2020. Effects of language on visual perception. *Trends in cognitive sciences*.

Subba Reddy Oota, Vijay Rowtula, Manish Gupta, and Raju S Bapi. 2019. Stepencog: A convolutional lstm autoencoder for near-perfect fmri encoding. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *Image*, 2:T2.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. 2020a. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *BioRxiv*.

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. 2020b. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. Inducing brain-relevant bias in natural language processing models. *Advances in Neural Information Processing Systems*, 32:14123–14133.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. 2006. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Aria Wang, Michael Tarr, and Leila Wehbe. 2019. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *Advances in Neural Information Processing Systems*, 32:15501–15511.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *in press*.

Daniel Yamins, Ha Hong, Charles Cadieu, and James J DiCarlo. 2013. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream.

Daniel LK Yamins and James J DiCarlo. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.

## A Do multi-modal Transformers perform better encoding compared to intermediate layer representations from pretrained CNNs?

We present the 2V2 accuracy and Pearson correlation for models trained with representations extracted from the last layer of multi-modal Transformers and all the lower to higher-level representations from pretrained CNNs on the two datasets: BOLD5000 and Pereira in Figs. 9 and 10, respectively.

We make the following observations from Fig. 9: (1) With respect to 2V2 and Pearson correlation, the multi-modal Transformer, VisualBERT, performs better than all the internal representations of pretrained CNNs. (2) In the pretrained CNNs, intermediate blocks have better correlation scores as compared to lower or higher level layer representations. (3) Other multi-modal Transformers, CLIP, and LXMERT, have marginal improvements over all the models except intermediate blocks such as Conv2D150 in InceptionV2ResNet.

We make the following observations from Fig. 10: (1) With respect to 2V2 and Pearson correlation, the multi-modal Transformer, VisualBERT, performs better than all the internal representations of pretrained CNNs. (2) Similar to BOLD5000, the intermediate blocks have better correlation scores as compared to lower or higher level layer representations in the pretrained CNNs on Pereira Dataset. (3) Other multi-modal Transformer, LXMERT, have equal performance with intermediate blocks of each pretrained CNN model.

## B Do multi-modal Transformers perform better encoding in their layers?

Given the hierarchical processing of visual or visual-language information across the Transformer layers, we further examine how these Transformer layers encode fMRI brain activity using image and mulit-modal Transformers. We present the layer-wise encoding performance results on two datasets: BOLD5000 and Pereira in Figs. 11 and 12, respectively.

We make the following observations from Fig. 11: (i) The multi-modal Transformer, VisualBERT, have consistent performance across the layers from 1 to 12. (ii) The LXMERT model have marginal decreasing performance from intermediate layer (L7) to higher layers. (iii) The image Transformers have higher Pearson correlation for early visual areas in the lower layers whereas higher visual areas such as LOC, OPA, and PPA have an increasing correlation in higher layers. (iv) This clearly indicates that the hierarchy of processing of visual stimulus in the human brain is similar to image Transformer layers.

We make the following observations from Fig. 12: (i) The multi-modal Transformers, VisualBERT, have consistent performance across the layers from 1 to 12. (ii) The LXMERT model have marginal decreasing performance from lower to higher layers. (iii) The image Transformer, ViT, has higher Pearson correlation for early visual areas in the lower layers whereas higher visual areas such as Vision_Body, Vision_Face, and Vision_Obj have an increasing correlation in higher layers.

## C Brain Maps for various models for BOLD5000 Dataset

Fig. 13 shows mean absolute errors (MAE) between actual and predicted voxels for various models on the BOLD5000 dataset. Notice that the magnitude of errors is much higher for a majority of voxels, compared to that with the VisualBERT model as shown in Fig. 4(a). Also, the multi-modal Transformers, VisualBERT (MAE range: 0 to 0.0181) and LXMERT (MAE range: 0 to 0.0188), have lower MAE compared to both image Transformers (MAE range: 0 to 0.02) and pretrained CNNs (MAE range: 0 to 0.0236).

## D Brain Maps for various models for Pereira Dataset

Fig. 14 shows mean absolute errors (MAE) between actual and predicted voxels for various models on the Pereira dataset. Notice that the magnitude of errors is much higher for a majority of voxels, compared to that with the VisualBERT model as shown in Fig. 14(a). Also, the multi-modal Transformers, VisualBERT and LXMERT, and InceptionV2ResNet+Conv2D150 have lower MAE compared to both image Transformers and other pretrained CNNs.

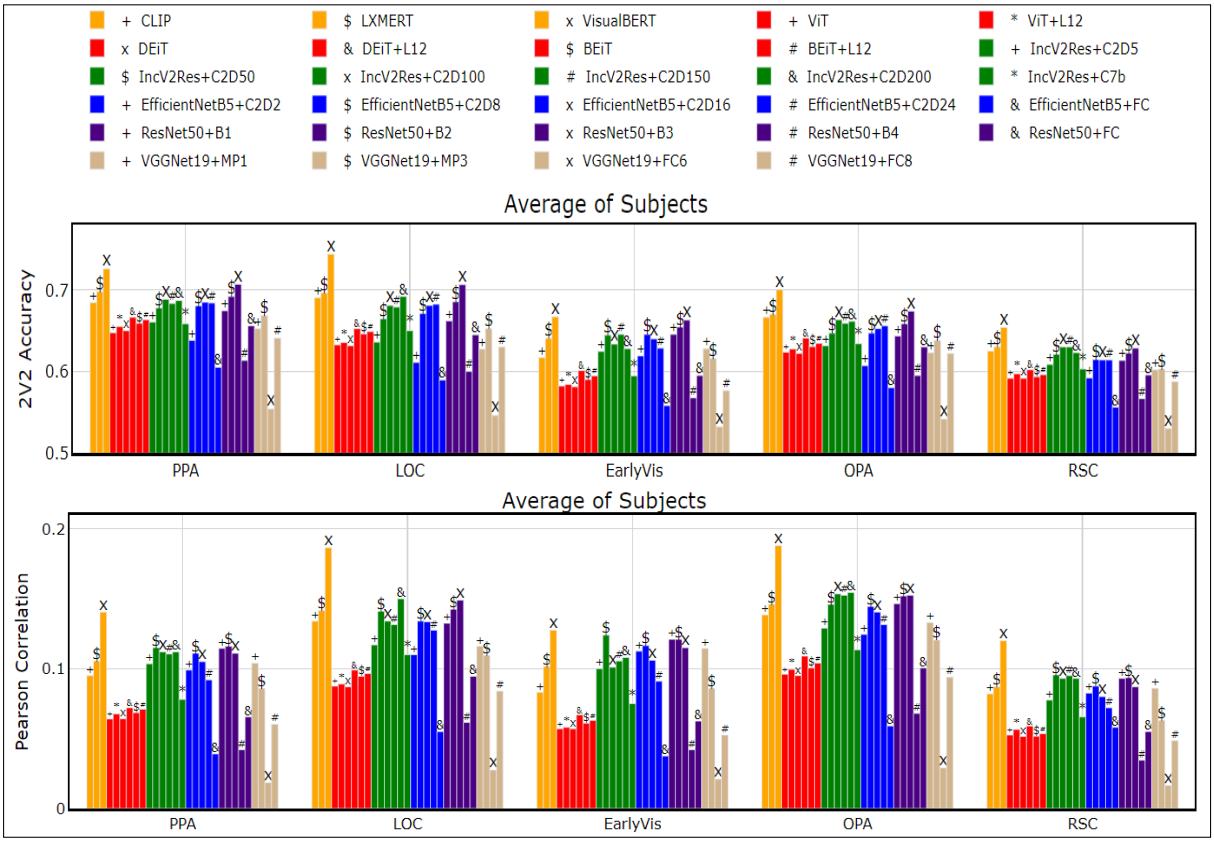


Fig. 9: BOLD5000: 2V2 (top Fig.) and Pearson correlation coefficient (bottom Fig.) between predicted and true responses across different brain regions using variety of models. Results are averaged across all participants. Pretrained CNN results are shown for all layers while multi-modal Transformer results are shown for last layers only.

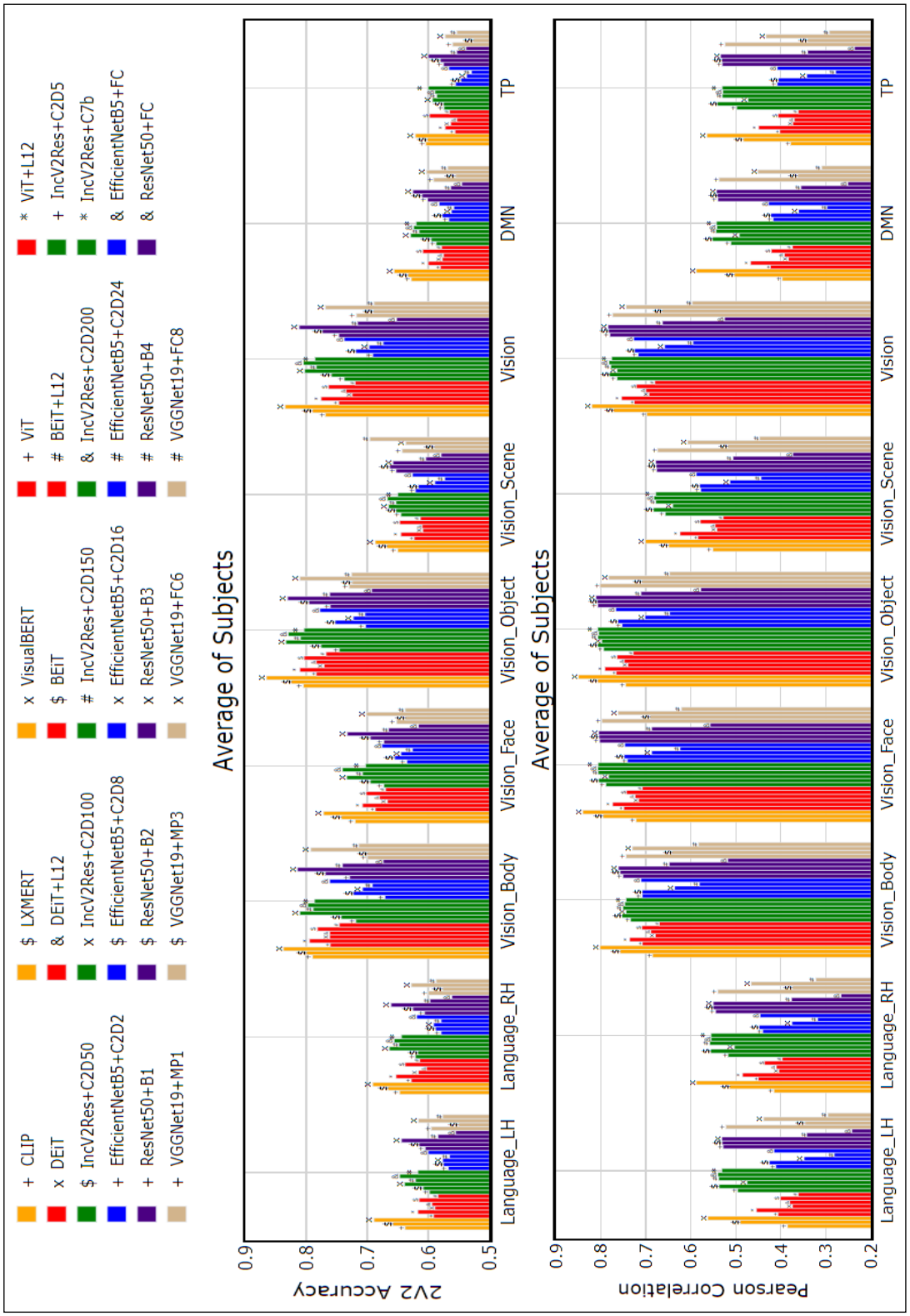


Fig. 10: Pereira dataset: 2V2 (top Fig.) and Pearson correlation coefficient (bottom Fig.) between predicted and true responses across different brain regions using variety of models. Results are averaged across all participants. Pretrained CNN results are shown for all layers while multi-modal Transformer results are shown for last layers only.

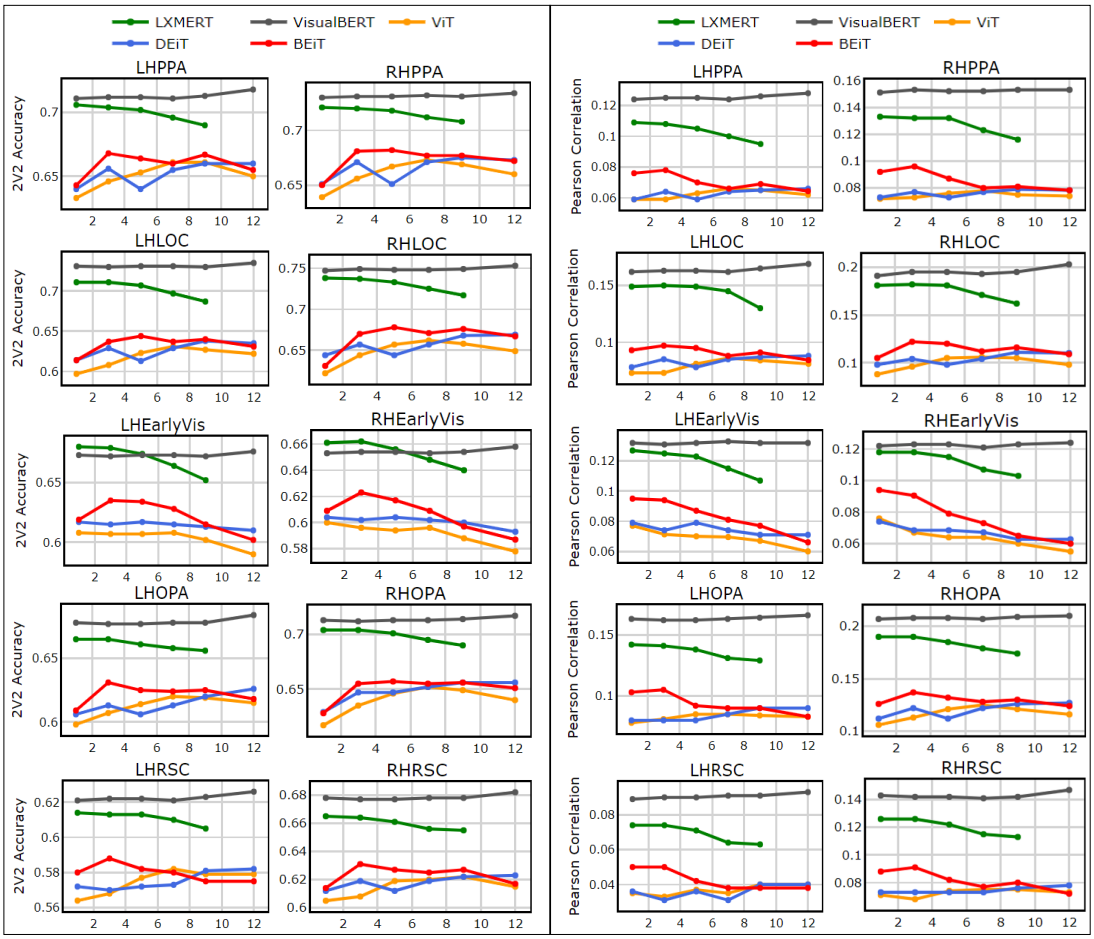


Fig. 11: BOLD5000: 2V2 (left) and Pearson correlation coefficient (right) between predicted and true responses across different brain regions using Transformer models. Results are averaged across all participants. The results are shown for all layers of image and multi-modal Transformers. Note that LXMERT has only 9 layers.

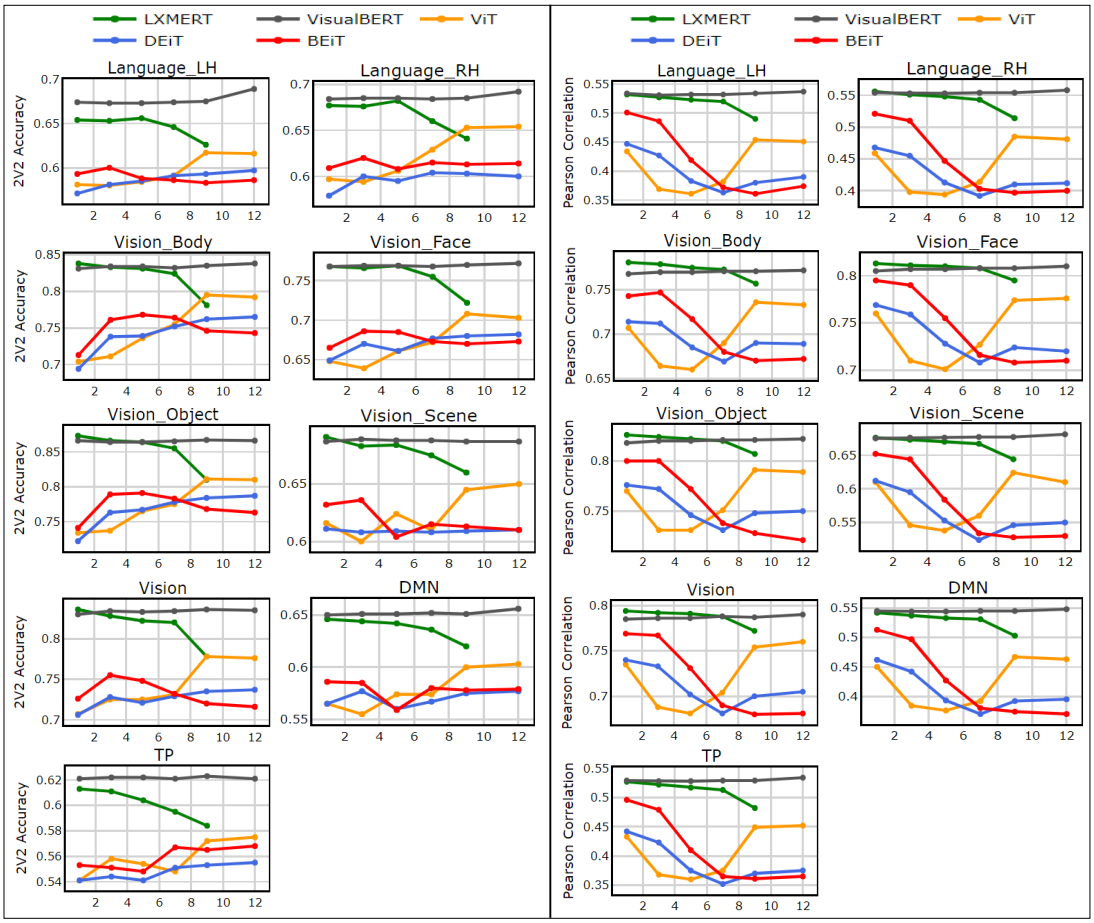


Fig. 12: Pereira: 2V2 (left) and Pearson correlation coefficient (right) between predicted and true responses across different brain regions using Transformer models. Results are averaged across all participants. The results are shown for all layers of image and multi-modal Transformers. Note that LXMERT has only 9 layers.

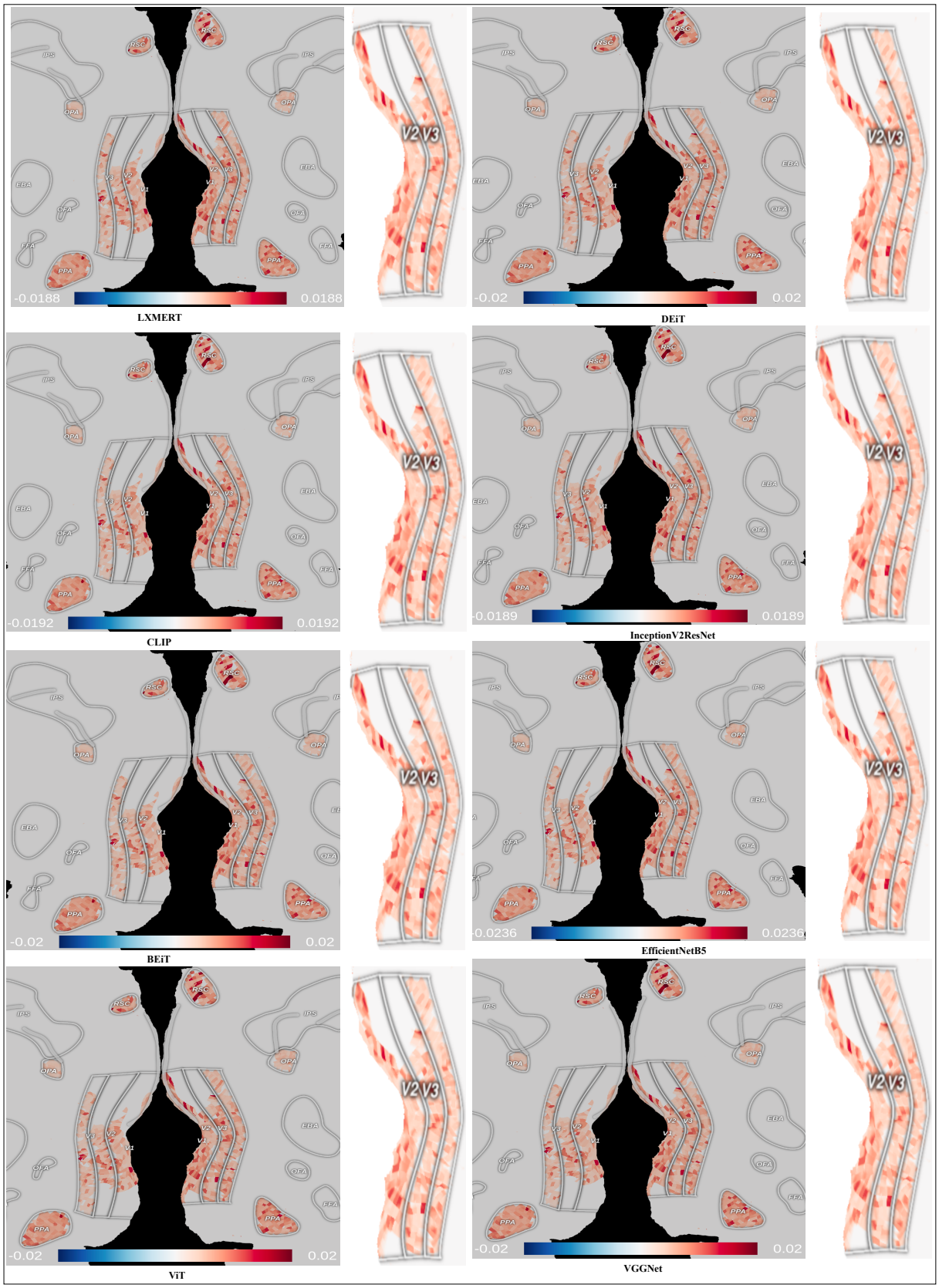


Fig. 13: MAE between actual and predicted voxels zoomed on V2 and V3 brain areas for various models. Note that V1 and V2 are also called EarlyVis area, while V3 is also called LOC area.

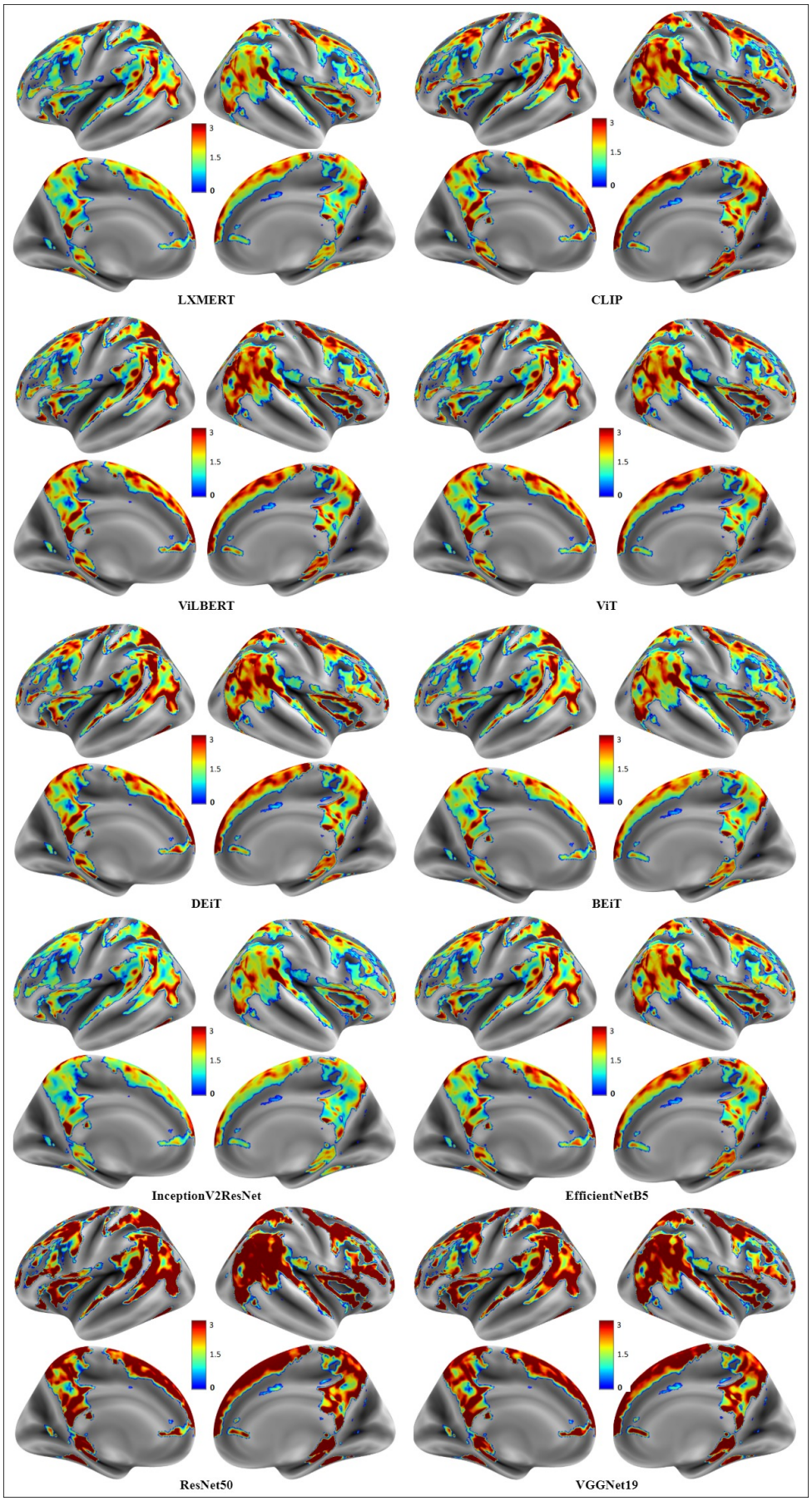


Fig. 14: MAE between actual and predicted voxels zoomed on V2 and V3 brain areas for various models. Note that V1 and V2 are also called EarlyVis area, while V3 is also called LOC area.