Visualizing and Controlling Cortical Responses Using Voxel-Weighted Activation Maximization

Supplementary Material

5. Supplementary Results

To evaluate the overall quality of the DNN-to-brain mapping, we computed the prediction accuracy of our Inception V3-based encoding model on the testing set of naturalistic movie clips. Prediction accuracy was noise-ceiling corrected and averaged across voxels and subjects for each region of interest (ROI). Shown in Fig. 4A and B, our results reveal high prediction accuracy across a wide range of visual cortical areas. High prediction accuracy in early visual areas (V1, V2, V3) suggests that our model captures selectivity to low-level visual features such as contrast and edges. High prediction accuracy in higher-level regions, including face-selective areas (OFA and FFA) and the body-selective Extrastriate Body Area (EBA), demonstrates that the same model also captures selectivity for more high-level visual features. Additionally, although slightly lower, prediction accuracy is consistently high in scene-selective regions such as the Occipital Place Area (OPA), Parahippocampal Place Area (PPA), and the Retrosplenial Complex (RSC). These results demonstrate that—by drawing from layers across Inception V3—a single encoding model can effectively model responses across visual areas.

Furthermore, Fig. 4A shows that this model also predicts responses accurately in many voxels outside of established regions. Taken together, these results show that our model can effectively predict fMRI responses to diverse visual characteristics across the visual system. This serves as a critical foundation for predicting fMRI responses to new stimuli, including our synthetic stimuli.

5.1. Experiment 3: Closed-Loop Optimization and Cross-Subject Generalization

Fig. 5 summarizes the key findings from the cross-subject analysis. Fig. 5A displays example synthetic images for each target ROI (V3, LO, FFA, EBA, and RSC) generated from the averaged weights. These images exhibit region-specific features consistent with known selectivity—for example, clear facial elements in face-selective regions and natural scene-like patterns in scene-selective areas. Fig. 5B presents flatmap visualizations of the contrasted responses for images optimized for EBA (red), FFA (green), and RSC (blue) for Subjects 1 and 2, demonstrating that the spatial distribution of the evoked responses aligns with the expected anatomical boundaries. Finally, Fig. 5C shows ROI-wise bar plots of the average fMRI responses across the four test subjects. In each ROI, the response to images optimized for that ROI was higher than the responses to images optimized for other regions in four out of five ROIs, thereby confirming that the synthetic images robustly drive the intended cortical responses across subjects.

These results indicate that our voxel-weighted activation maximization framework is not subject-specific but captures features that generalize across individuals. The robust activation patterns observed in the cross-subject evaluation further support the utility of our method in probing the functional selectivity of cortical regions.

6. Supplementary Methods

6.1. fMRI Data Acquisition and Preprocessing

For fitting our DNN-based encoding models, we used BOLD fMRI responses to a large set of naturalistic movie clips from Huth *et al*. [9]. This consisted of video clips depicting a wide variety of dynamic scenes, including people, animals, objects, and natural environments. These stimuli were divided into 120 minutes of training clips and 9 minutes of testing clips. We used BOLD responses to these stimuli for 6 subjects (4 male, 2 female); data for five of these subjects was previously collected and data for one subject was collected for this study. During scanning, participants maintained fixation on a central fixation cross. See Huth *et al*. [9] and Popham *et al*. [18] for more details.

Voxels were filtered based on a noise ceiling criterion that estimates response reliability across repeated stimulus presentations. Specifically, the noise ceiling was computed as the mean pairwise correlation between BOLD responses to repeated presentations of the naturalistic testing clips. Voxels with a noise ceiling significantly above zero ($p < 0.05$, uncorrected) were included in subsequent analyses.

In addition, for experiments 2 and 3 we collected BOLD responses to images synthesized via our activation maximization approach. We collected BOLD data for four subjects (2 male, 2 female) using a Siemens Skyra 3T scanner with a 32-channel head coil. Blood oxygenation level dependent (BOLD) fMRI data were collected with a repetition time (TR) of 1000 ms, an echo time (TE) of 31 ms, and a voxel resolution of $3.2 \times 3.2 \times 2.6$ mm. Each scanning run consisted of 100 stimulus images, preceded by 12 blank trials, followed by 16 blank trials, and interleaved with 10 additional blank (50% luminance gray screen) trials to ensure robust baseline estimation and to minimize adaptation effects. During BOLD data collection, participants were tasked with fixating on a central fixation dot while attending to the image contents.

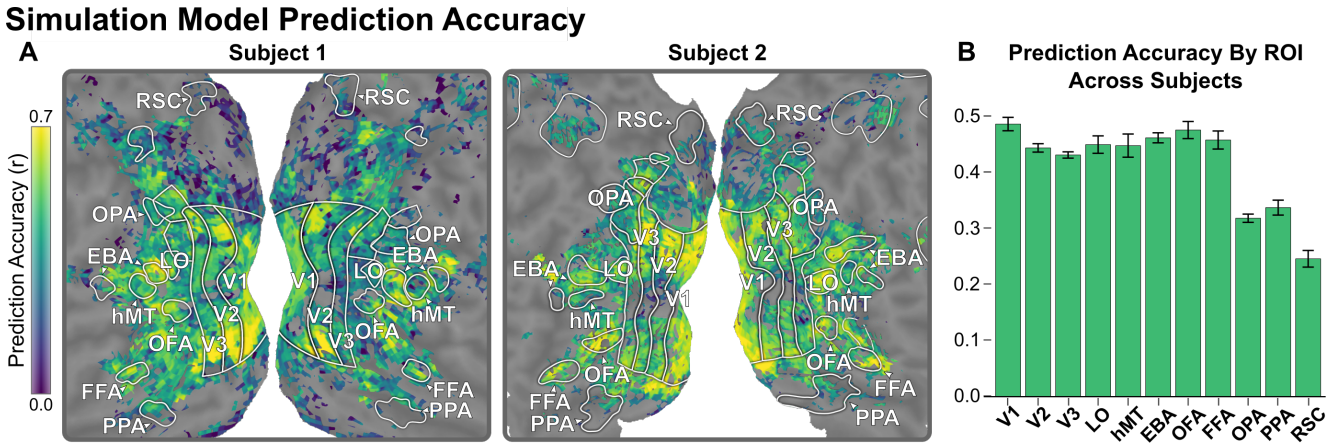

Figure 4. **Prediction Accuracy of the Inception V3-Based Encoding Model. (A)** Flatmaps of voxelwise prediction accuracy for subjects 1 and 2 in natural movie data, illustrating the encoding model prediction accuracy across the visual cortex. **(B)** Bar chart displaying ROI-wise prediction accuracy, noise-ceiling corrected and averaged across voxels and subjects, for early and higher-level regions. Error bars show SEM across subjects.

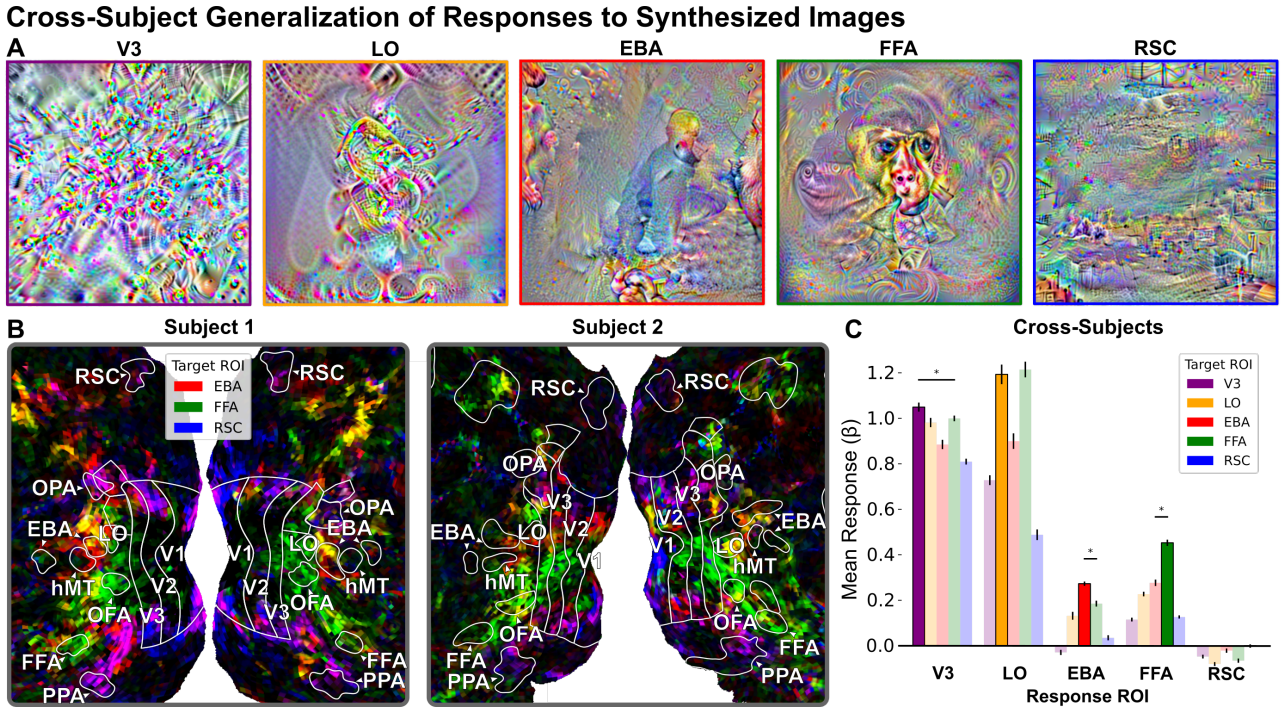

Figure 5. **Cross-Subject Generalization of ROI-Level Synthetic Images. (A)** Example synthetic images for each ROI, generated using predicted responses from a group of four subjects and then presented to a separate set of four subjects. These images capture region-specific features consistent with known selectivity. **(B)** Flatmap visualizations of contrast responses for subjects 1 and 2 illustrate the spatial distribution of activation differences evoked by the synthetic images. **(C)** A single bar plot shows the average fMRI response across subjects for the synthetic images, with the highlighted bars corresponding to responses evoked by images optimized for each target ROI. Together, these results demonstrate robust cross-subject generalization of our voxel-weighted activation maximization framework.

Estimation of responses to each image was performed using GLMsingle [19]. Standard preprocessing procedures (*e.g.*, motion correction) were applied to all BOLD data. Functional data for experiments 2 and 3 were temporally downsampled to an effective TR of 2.0 seconds to reduce physiological noise.

Experimental protocols were approved by the Institutional Review Boards of the University of Nevada, Reno

and the University of California, Berkeley.

6.2. DNN Activation Extraction and Downsampling

Image Preprocessing: Naturalistic movie frames and synthesized images were preprocessed using the standard transformations expected by the Inception V3 model. Each image was resized to 299 pixels, center-cropped to produce a 299×299 image, and normalized using the ImageNet mean values [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225]. These operations are fully differentiable, which allows us to optimize higher-resolution images (*e.g.*, 500×500 pixels) despite the network receiving inputs at a resolution of 299×299.

Layer Selection: Activations were extracted from 21 layers of the Inception V3 network [20]:

1. Conv 1a (3×3)
2. Conv 2a (3×3)
3. Conv 2b (3×3)
4. MaxPool 1
5. Conv 3b (1×1)
6. Conv 4a (3×3)
7. MaxPool 2
8. Inception 5b
9. Inception 5c
10. Inception 5d
11. Reduction 6a
12. Inception 6b
13. Inception 6c
14. Inception 6d
15. Inception 6e
16. Reduction 7a
17. Inception 7b
18. Inception 7c
19. AvgPool
20. Dropout
21. FC

This selection spans from low-level features to high-level semantic representations.

Adaptive Spatial Downsampling: To manage the high dimensionality of activations from convolutional layers, we employ adaptive spatial pooling. We use PyTorch's adaptive pooling functions (similar to those described in He *et al*. [8]) to reduce each feature map to a fixed output size such that approximately 5,000 features are retained per layer. The target output size for each spatial dimension is computed by

$$S_i = \left\lfloor \left(\frac{F_{\max}}{C} \right)^{\frac{1}{n}} \right\rfloor \tag{1}$$

where $F_{\max} = 5000$, C is the number of channels, and n is the number of spatial dimensions. Since n can vary, this formulation generalizes beyond 2D and is applicable to higher-dimensional feature maps (*e.g.*, those from 3D CNNs).

Temporal Downsampling: Due to the mismatch between stimulus framerate (15 HZ) and fMRI repetition time (0.5 HZ), we average the downsampled features extracted from frames of the naturalistic movie clips over each 2-second window to match the temporal resolution of the BOLD signal.

6.3. Voxelwise Encoding Model Fitting

In order to generate a predictive mapping from extracted DNN activations and fMRI responses, we fit voxelwise encoding models [14]. To do this, we first extract downsampled layerwise activations (see above) to the training set of naturalistic movie clips. These downsampled activations from each layer are flattened and concatenated, yielding a feature vector of roughly 78,000 elements per stimulus. We then fit a ridge regression model to predict voxelwise fMRI responses from these features. Prior to regression, the fMRI responses are z-scored across TRs, and temporal lags of 2, 4, and 6 seconds are combined into a single design matrix.

For each voxel i, the regression model is formulated as

$$\hat{\beta}_i = \left(\mathbf{X}^\top \mathbf{X} + \alpha I \right)^{-1} \mathbf{X}^\top \mathbf{y}_i \tag{2}$$

where $\mathbf{X}$ is the design matrix containing the concatenated DNN features, $\mathbf{y}_i$ is the z-scored fMRI response vector for voxel i, α is the regularization parameter, and I is the identity matrix. The regularization parameter (α) is selected from a set of 15 values that are logarithmically interpolated between 10^0 and 10^{10}. For each voxel, the optimal α is chosen based on maximizing the cross-validated R^2 using 10 splits and 10 resamplings (n_splits = 10, n_resamps = 10). Regression fitting was performed using the tikreg package [2].

6.4. Activation Maximization

We perform activation maximization by optimizing an input image to maximize the predicted fMRI response in a target voxel or region. The optimization is carried out in the Fourier domain, which biases the solution towards smooth, interpretable patterns and avoids high-frequency artifacts [16].

The procedure is as follows:

- **Initialization:** The input is initialized as a neutral grayscale image (with a pixel value of 140) in Fourier space.
- **Inverse Fourier Transform:** The image is transformed to the spatial domain via an inverse FFT, taking the magnitude to obtain real-valued pixel intensities.
- **Image Augmentations:** To regularize optimization, we apply a series of random invariance transformations:
  - A random crop to 500×500 pixels with a padding of 5 pixels.

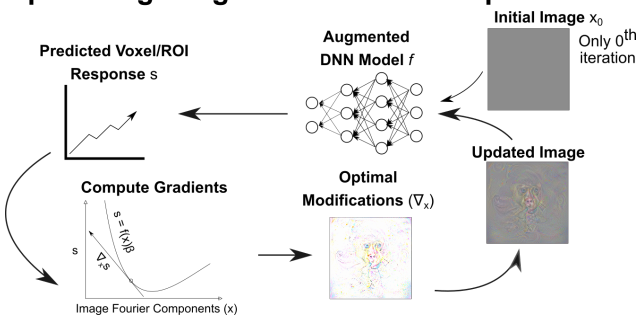

Figure 6. **Generation Process.** Diagram illustrating the activation maximization procedure. Starting from an initial neutral grayscale image in the Fourier domain, the image is transformed into the spatial domain using an inverse FFT. Random invariance transforms (*e.g.*, cropping, rotation within -5° to 5°, and resized cropping) are applied to regularize the optimization. The preprocessed image is then fed through the augmented network, and gradients are computed and backpropagated (using the Adam optimizer with a learning rate of 1×10^{-2}) to update the Fourier coefficients. This iterative gradient ascent process is repeated for 2,500 iterations to synthesize an image that maximizes the predicted fMRI response.

  - A random rotation between -5° and 5°.
  - A random resized crop to 500×500 pixels with a scaling factor between 0.95 and 1.05, maintaining a 1:1 aspect ratio.
  - A second random crop to 500×500 pixels with a padding of 3 pixels.
- **Preprocessing:** The augmented image is resized to 299×299 pixels, center-cropped, and normalized using the standard Inception V3 transformations.
- **Forward Pass and Loss Computation:** The preprocessed image is passed through the augmented network to obtain downsampled activations. The predicted fMRI response is computed as the dot product between the flattened activations and the target weight vector. We define the loss as the negative predicted response:

$$\mathcal{L}(\mathbf{x}) = -\left(\mathbf{f}(\mathbf{x}) \cdot \boldsymbol{\beta}\right) \tag{3}$$

- **Gradient Update:** Gradients are backpropagated to the Fourier domain representation, and the Adam optimizer (with a learning rate of 1×10^{-2}) updates the Fourier coefficients. The optimization runs for 2,500 iterations.

For synthetic images optimized for single-voxel responses, we incorporated two modifications to enhance interpretability. First, the image initialization was set to be completely black with a small amount of noise added. We found that this initialization yields a clearer distinction between the emergent image content and the background. Second, we applied a color channel decorrelation step based on the approach described in Olah *et al*. [16]. This decorrelation reduces redundancy across the red, green, and blue channels, encouraging the emergence of distinct color features in the synthesized images. While these modifications improve the qualitative interpretability of the generated images, their effects on actual fMRI responses remain a subject for future investigation.

6.5. Optimization Objectives

For all three experiments, we create an optimization objective by computing contrasts in the space of regression weights. This contrastive approach minimizes the impact of shared visual selectivity among voxels or regions while amplifying their differences. Concretely, let $\boldsymbol{\beta} \in \mathbb{R}^d$ denote the regression weight vector for a given voxel or region. First, we normalize the weights by performing z-scoring across features:

$$\mathbf{z} = \frac{\boldsymbol{\beta} - \mu(\boldsymbol{\beta})}{\sigma(\boldsymbol{\beta})} \tag{4}$$

where $\mu(\boldsymbol{\beta})$ and $\sigma(\boldsymbol{\beta})$ are the mean and standard deviation of $\boldsymbol{\beta}$, respectively. This normalizes the scale of the weights (which is influenced by factors such as fMRI signal strength).

Next, we subtract an estimate of the mean feature selectivity—computed either across all cortical voxels (for individual voxel optimization) or across the five selected regions (for ROI-level optimization)—from the z-scored weights:

$$\boldsymbol{\beta}_{\text{contrast}} = \mathbf{z} - \bar{\mathbf{z}} \tag{5}$$

with $\bar{\mathbf{z}}$ representing the average z-scored weight vector. This step effectively removes common components of visual selectivity across voxels, allowing the optimization to focus on features that are uniquely pronounced in the target voxel or region.

Finally, to ensure a consistent rate of image optimization during activation maximization, we normalize the contrast weights by their L2 norm:

$$\boldsymbol{\beta}_{\text{final}} = \frac{\boldsymbol{\beta}_{\text{contrast}}}{\|\boldsymbol{\beta}_{\text{contrast}}\|} \tag{6}$$

This stabilizes the gradient ascent process by ensuring that the optimization target has a consistent scale, independent of the magnitude of the regression weights.

During image optimization, the normalized contrast weights are multiplied by the feature vector $\mathbf{f}(\mathbf{x})$ derived from the DNN layers, yielding an estimated contrast value:

$$s = \boldsymbol{\beta}_{\text{final}}^{\top} \mathbf{f}(\mathbf{x}) \tag{7}$$

This estimated contrast value is then used as the optimization target. Specifically, we define the loss function as

$$\mathcal{L}(\mathbf{x}) = -s \tag{8}$$

and apply gradient descent to minimize $\mathcal{L}(\mathbf{x})$, which is equivalent to performing gradient ascent on s.

6.6. Software and Implementation

Regression fitting was performed using the tikreg package [2]. Beta estimation of the fMRI response was conducted using GLMsingle [19] and ROIs and cortical surface visualizations were generated using Pycortex [4]. Custom PyTorch code for DNN feature extraction and synthetic image generation is available at https://github.com/MShinkle/VWAM.