CVPR
#16804

Reanimating Images using Neural Representations of Dynamic Stimuli

Anonymous CVPR submission

Paper ID 16804

Supplementary Material: Reanimating Images using Neural Representations of Dynamic Stimuli

A. Supplemental

Sections

CVPR
#16804

A.1. HCP Parcellation Map

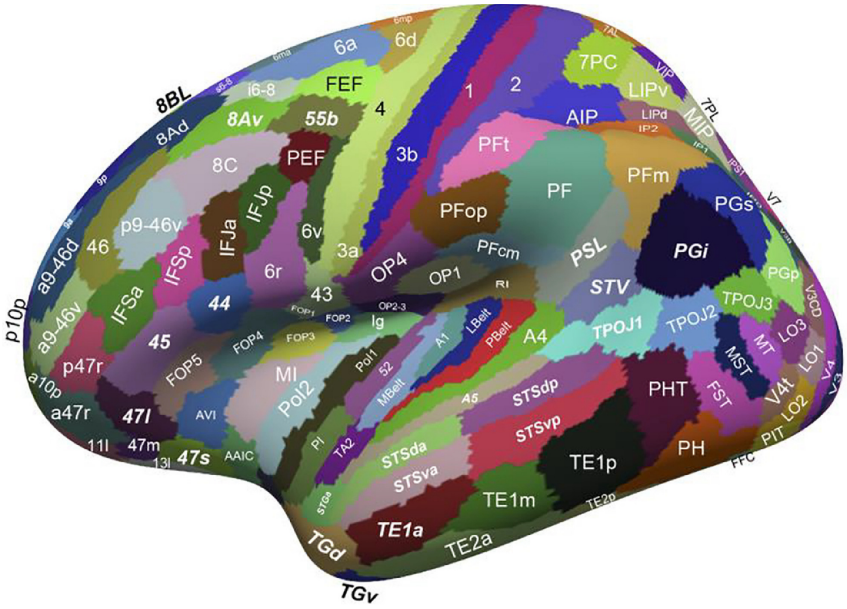


Figure S1. **HCP Parcellation Map.** The human cortex parceled into regions of interest as defined by the HCP-MMP [1] atlas and HCPex [2]. Regions were defined through multimodal imaging and take into account neural structure, connectivity, and task-related selectivity. This figure is a reproduction of a subset of Figure 1 from Rolls et al. [3].

A.2. Additional Dataset Information

We use the Dynamic Vision Dataset collected by Wen et al. [4]. The dataset is released under a CC0 1.0 universal license and can be found at https://purr.purdue.edu/publications/2805/1.

CVPR
#16804

A.3. Decoding Models: Visualizations

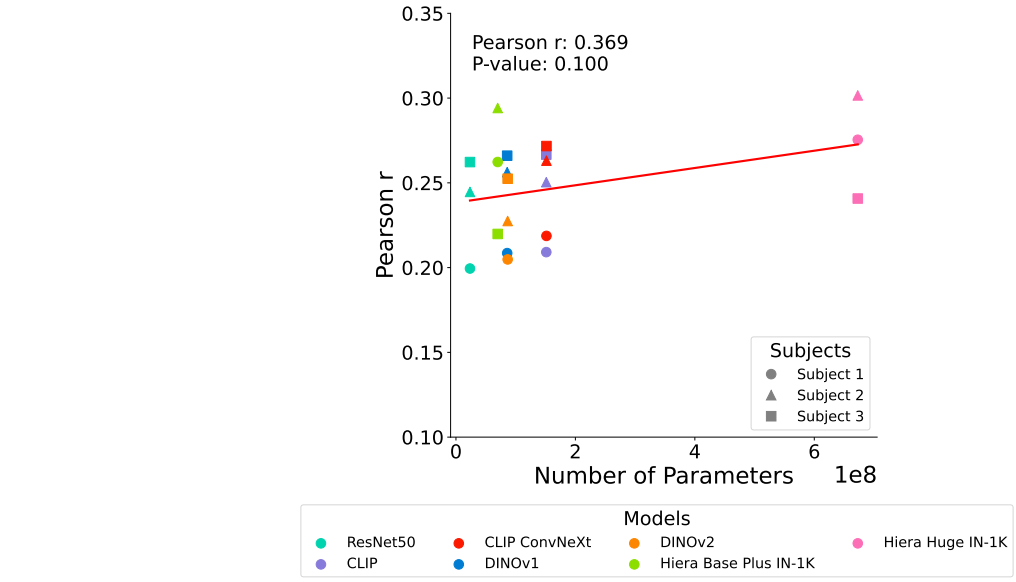
Attached is an HTML file, `index.html`, showing multiple reanimation examples as videos. To view the HTML file, open it in a browser. For each example, we show the ground truth video, the reanimated video using the flow predicted from the ground truth initial frame, and the reanimated video using the flow predicted from the initial frame generated by MindVideo [5]. The MindVideo initial frame represents the diffusion image as predicted by MindVideo from fMRI data. As such, this frame is often quite different in content from the ground truth, although the objects in each are in similar locations. However, the motion predicted by our model is still consistent with ground truth motion, for example, in the first video the jellyfish retracts backwards at the end clip. The same backwards motion can be seen in the reanimated videos for with both the ground truth initial frame and the MindVideo generated initial frame (which happens to be a boat). This establishes that the motion visualized using DragNUWA [6] is not derived solely from the diffusion model, but rather, incorporates our predicted motion.

In addition, we display example generated video frames below. Within each figure the first row is the ground truth frame, the second row is our reanimated video using the flow predicted from the ground truth initial frame, and the final row is the reanimated video using the flow predicted from the initial frame generated by MindVideo.

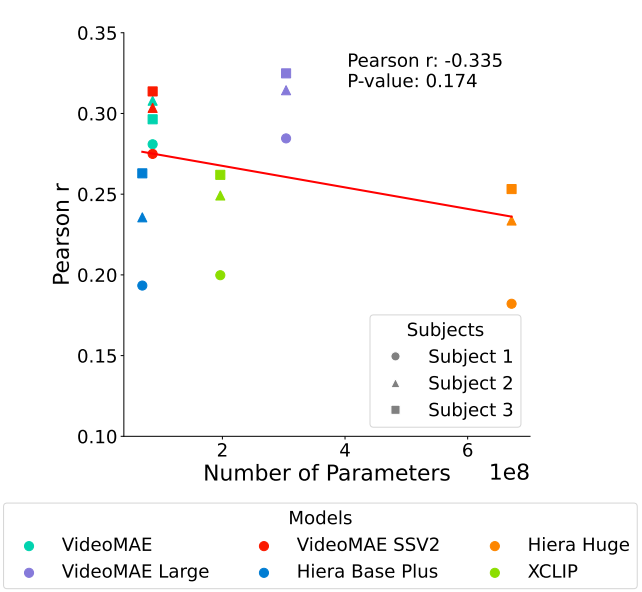
Figure S2. **Static image animation results.** Here we see an example of deer running to the right in the ground truth video (**first row**, "GT"). We show an example of animating the initial frame of the ground truth video by combining the brain conditioned motion prediction with DragNUWA [6] (**second row**, "Ours + GT"). We show an example of animating the initial frame obtained from fMRI data using MindVideo by combining the brain conditioned motion prediction with DragNUWA (**third row**, "Ours + MindVideo") and observe that the creatures run to the right.

Figure S3. **Static image animation results.** Here we see an example of a soldier running to the left in the ground truth video (**first row**, "GT"). We show an example of animating the initial frame of the ground truth video by combining the brain conditioned motion prediction with DragNUWA [6] (**second row**, "Ours + GT"). We show an example of animating the initial frame obtained from fMRI data using MindVideo by combining the brain conditioned motion prediction with DragNUWA (**third row**, "Ours + MindVideo") and observe that the child pedals its feet in a similar motion to that of the soldier running.

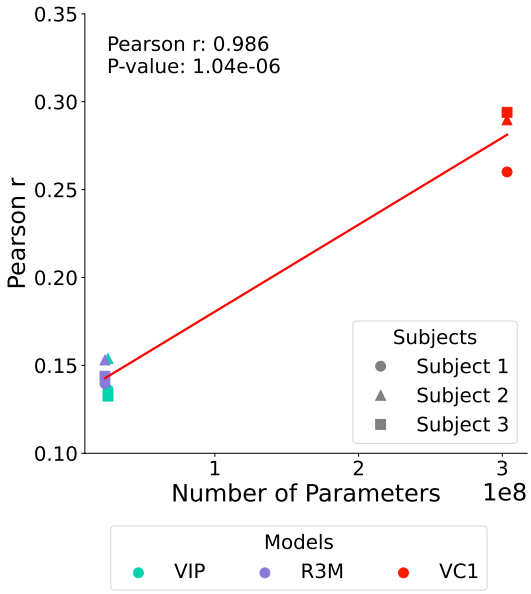
A.4. Encoding Models: Controls and Baselines



(a) **Image Models**

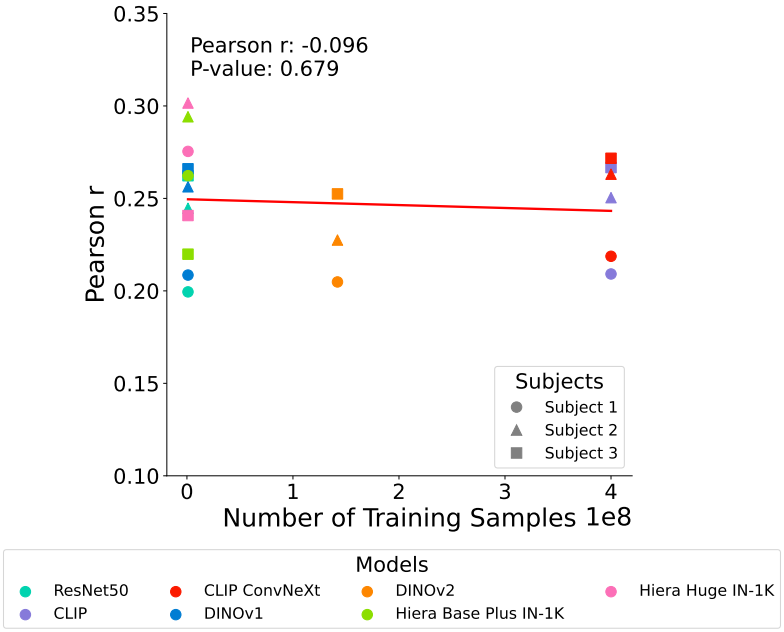


(b) **Video Models**

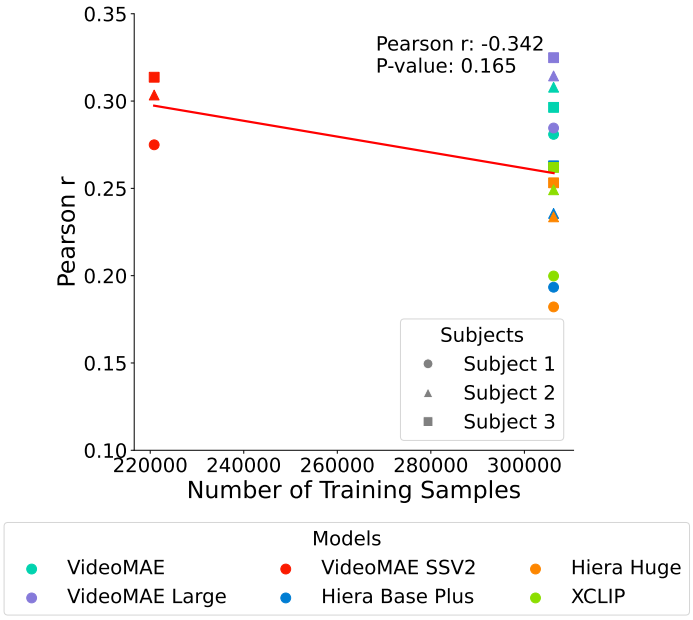


(c) **Embodied AI Models**

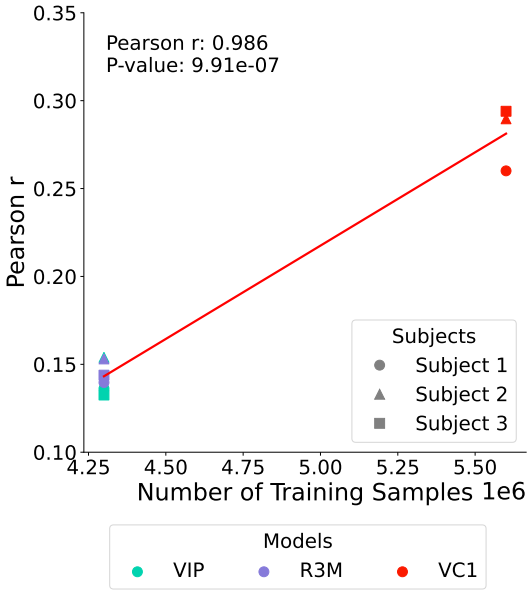
Figure S4. **Model size and encoding prediction performance.** Encoding model features of the viewed stimuli are used to predict voxel-wise fMRI brain activity [7]. The average Pearson-r is plotted across all voxels for each model and with respect to the number of parameters in each model. **(a)** Encoding performance for models trained on static images. **(b)** Encoding performance for models trained on videos. **(c)** Encoding performance for models trained to align representations of single frames across time for embodied AI visuomotor manipulation. Model size and encoding prediction performance are not significantly correlated (by statistical test) for image and video models. This indicates that the model size is not a confound for the encoding performance of image and video models. In contrast, model size and encoding performance are significantly correlated ($p < 0.001$) for the embodied AI visual models. Encoding performance per model is plotted for each of the three participants – the circle marker refers to S1, the triangle marker refers to S2, and the square marker refers to S3. Comparisons between models and model sizes are a critical step in building better and more interpretable models for understanding the human brain. Model architectures and training tasks, as well as the number of parameters and other model characteristics, instantiate varying constraints that have functional implications for brain prediction [8, 9].



(a) **Image Models**



(b) **Video Models**



(c) **Embodied AI Models**

Figure S5. **Training data size and encoding prediction performance.** Encoding model features of the viewed stimuli are used to predict voxel-wise fMRI brain activity [7]. The average Pearson-r is plotted across all voxels for each model and with respect to the number of samples each model is trained on. **(a)** Encoding performance for models trained on static images. **(b)** Encoding performance for models trained on videos. **(c)** Encoding performance for models trained to align representations of single frames across time for embodied AI visuomotor manipulation. Training data size and encoding prediction performance are not significantly correlated (by statistical test) for image and video models. This indicates that the training data size size is not a confound for the encoding performance of image and video models. In contrast, model size and encoding performance are significantly correlated ($p < 0.001$) for the embodied AI visual models. Encoding performance per model is plotted for each of the three participants – the circle marker refers to S1, the triangle marker refers to S2, and the square marker refers to S3.

CVPR
#16804

VideoMAE Models (Pearson r)		
Model	**Dataset**	**Encoding Performance**
VideoMAE Base	K-400	0.2964
VideoMAE Large	K-400	0.3248
VideoMAE Base	SSV2	0.3137

Table S1. **Control for dataset distribution: encoding performance of VideoMAE.** VideoMAE trained on SSV2 predicts fMRI brain activity at the same level as VideoMAE trained on Kinetics-400. All three VideoMAE models are better at predicting fMRI brain activity as compared to all other tested models as listed in Figure 7 in the main text. This combined with the results in Figure S4 indicates that model size and training data size are unlikely potential confounding factors. These results suggest that the architecture and training paradigm of VideoMAE lead to better fMRI brain activity prediction.

Hiera Models (Pearson r)		
Model	**Dataset**	**Encoding Performance**
Hiera Base Plus	K-400	0.2629
Hiera Base Plus	IN-1K	0.2198
Hiera Huge	K-400	0.2532
Hiera Huge	IN-1K	0.2407

Table S2. **Control for dataset distribution: encoding performance of Hiera.** Hiera models of the same size trained on Kinetics-400 (K-400) are better at predicting fMRI brain activity as compared to Hiera models trained on Imagenet-1K (IN-1K). This result supports the claim that fMRI brain activations encode dynamic visual representations, as modeling temporal dynamics improves fMRI brain activity prediction when architecture is held constant.

A.5. Encoding Models: Voxel-wise Prediction Performance on Inflated Cortical Maps of the Brain

Inflated brain maps showing voxel-wise fMRI prediction performance, quantified as the Pearson correlation (r) between measured and predicted responses, for all visual encoding models not shown in the main text. Maps for each model are shown in alphabetical order. 'AF' in the model name denotes a model using *A*verage *F*rames. All maps are shown using the same scale for ease of comparison between models.

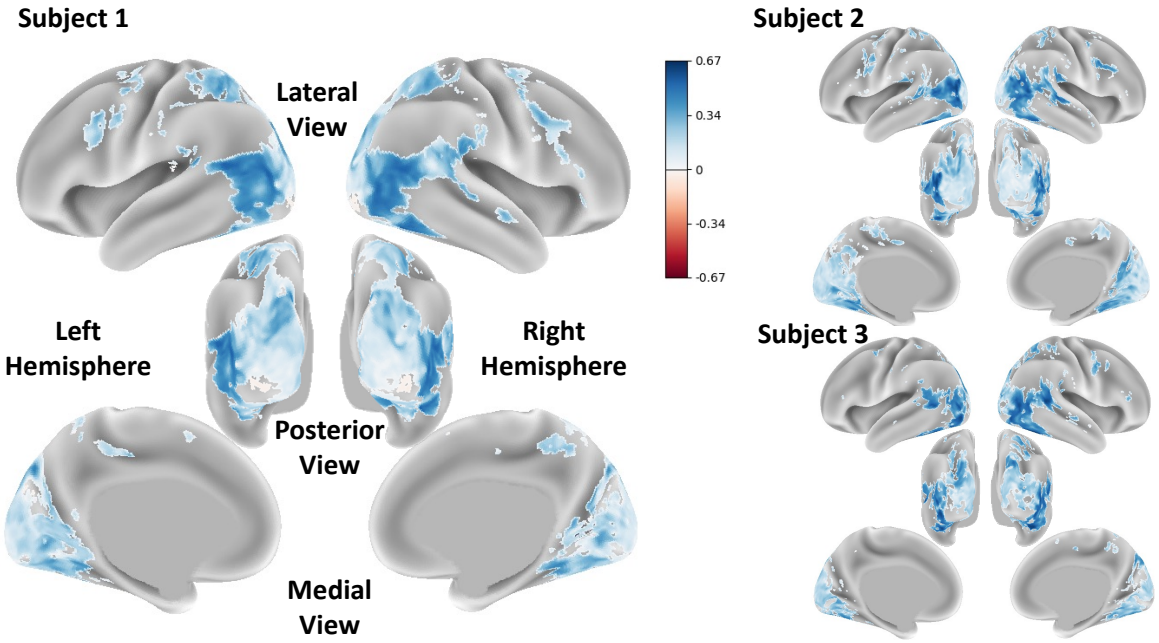


Figure S6. **CLIP Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of CLIP quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

CVPR
#16804

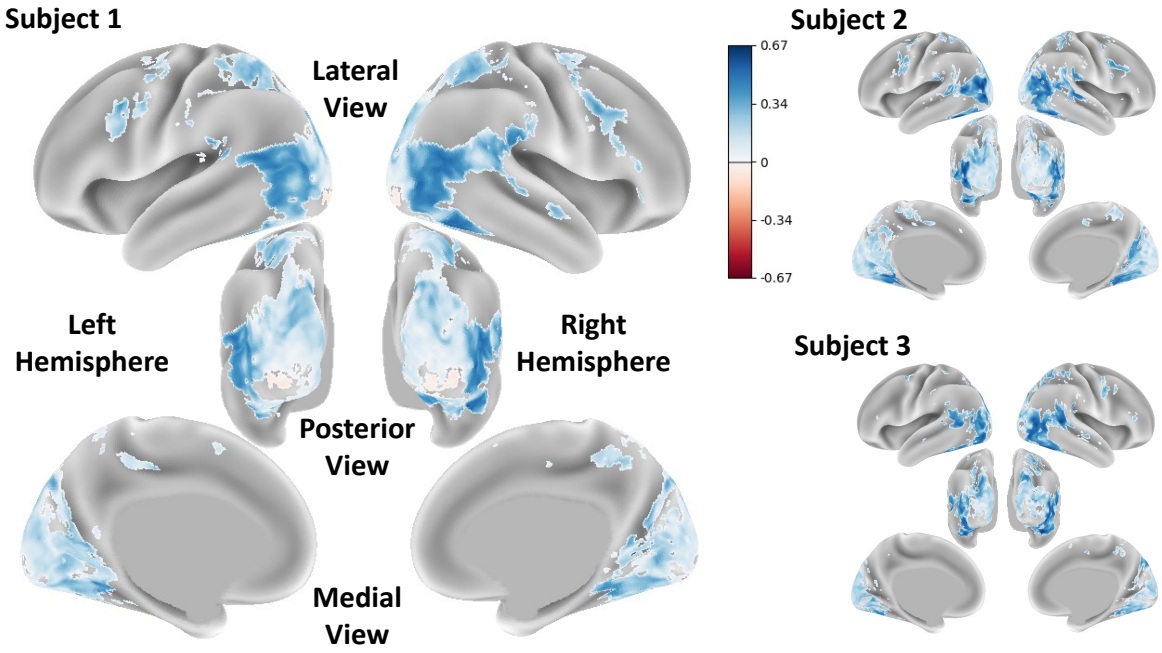


Figure S7. **CLIP AF Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of CLIP AF quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

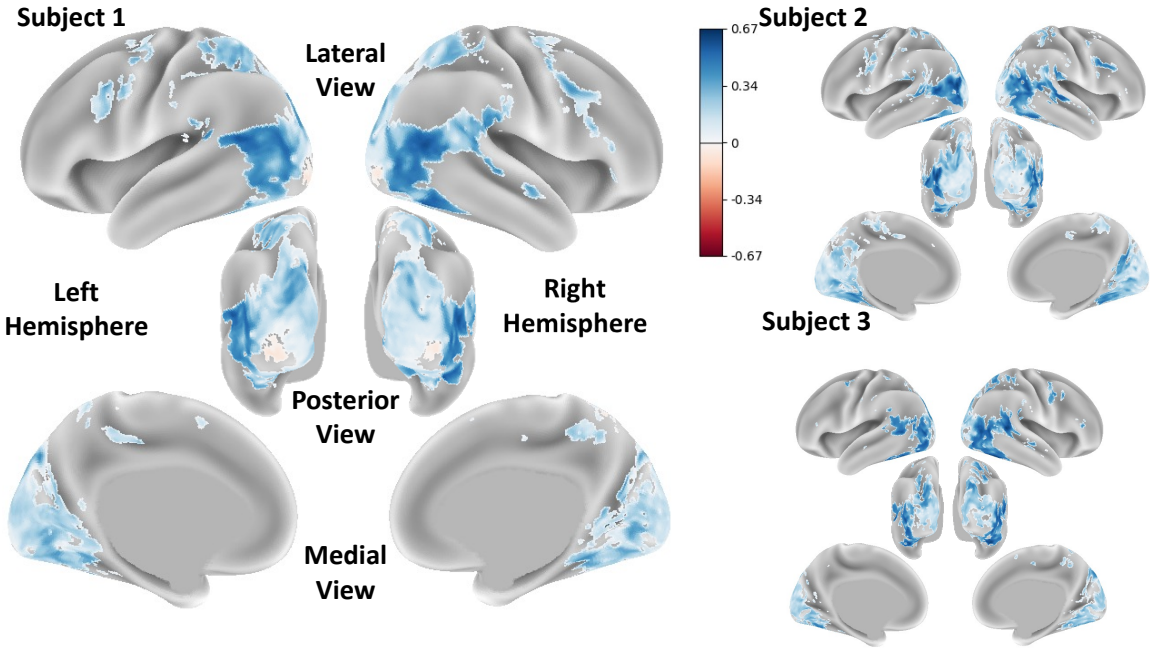


Figure S8. **CLIP ConvNeXt Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of CLIP ConvNeXt quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

CVPR
#16804

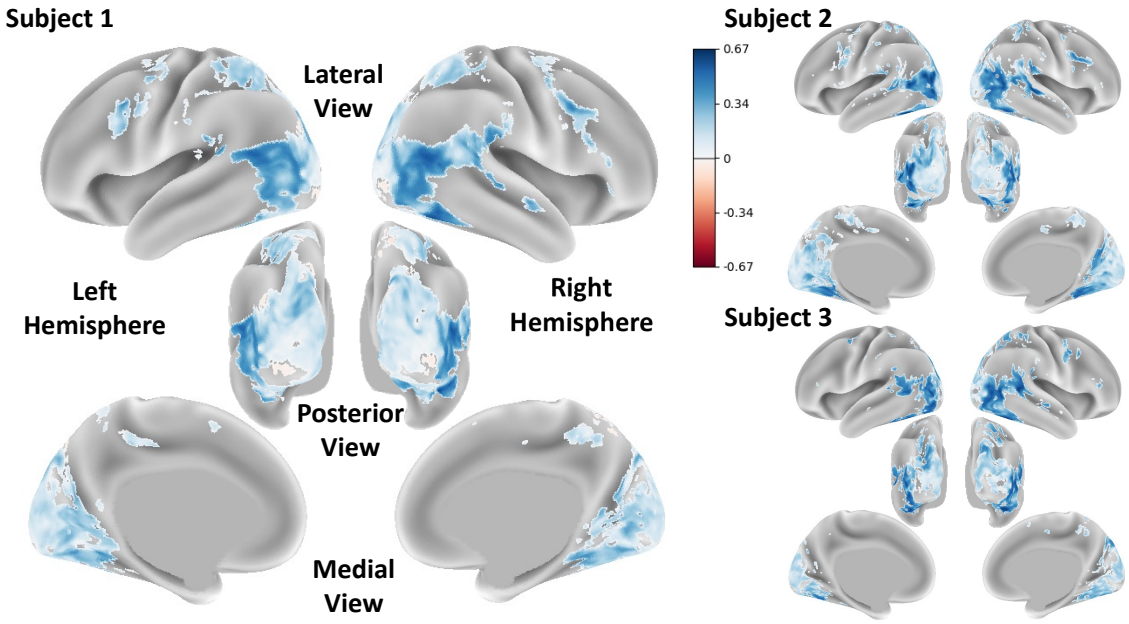


Figure S9. **CLIP ConvNeXt AF Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of CLIP ConvNeXt AF quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

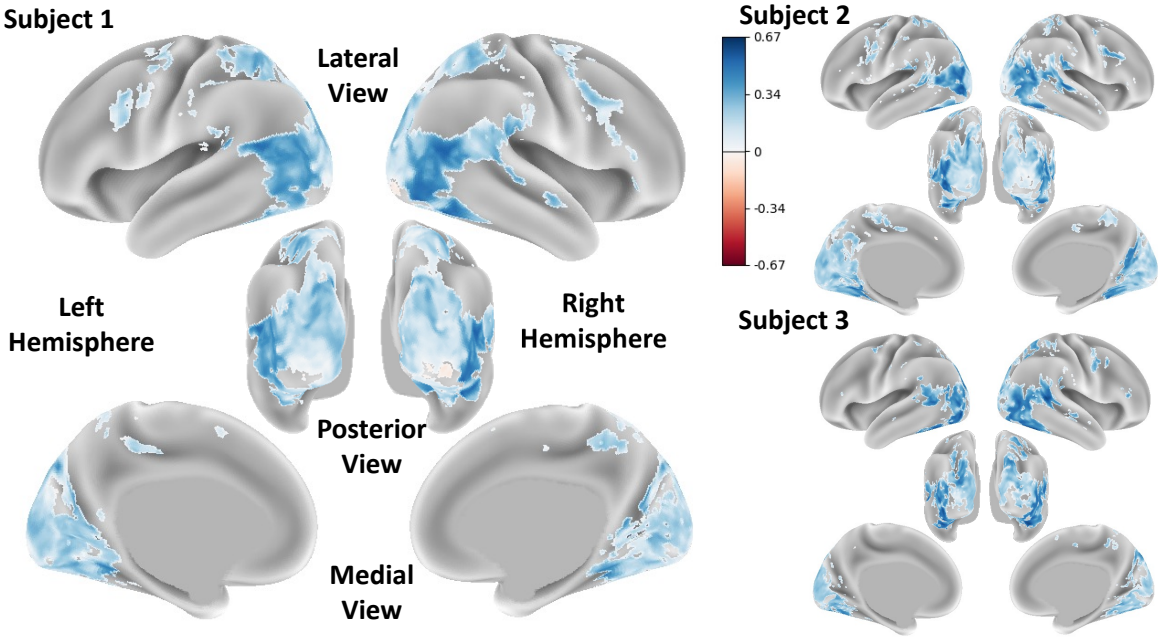


Figure S10. **DINOv1 Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of DINOv1 quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

CVPR
#16804

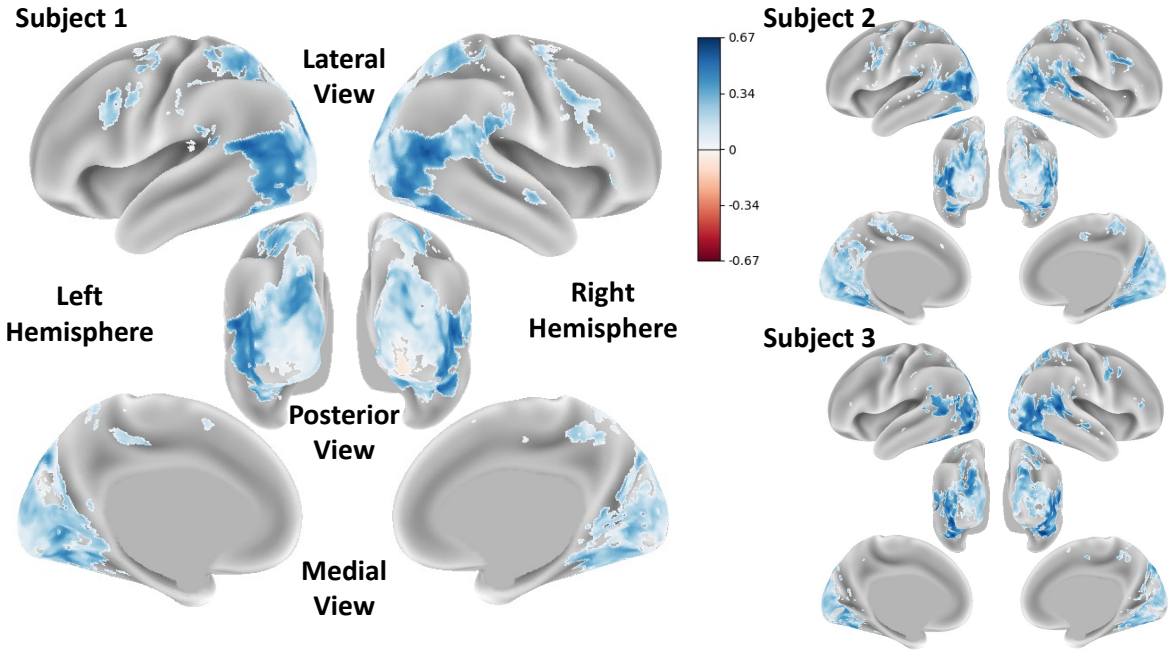


Figure S11. **DINOv2 Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of DINOv2 quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

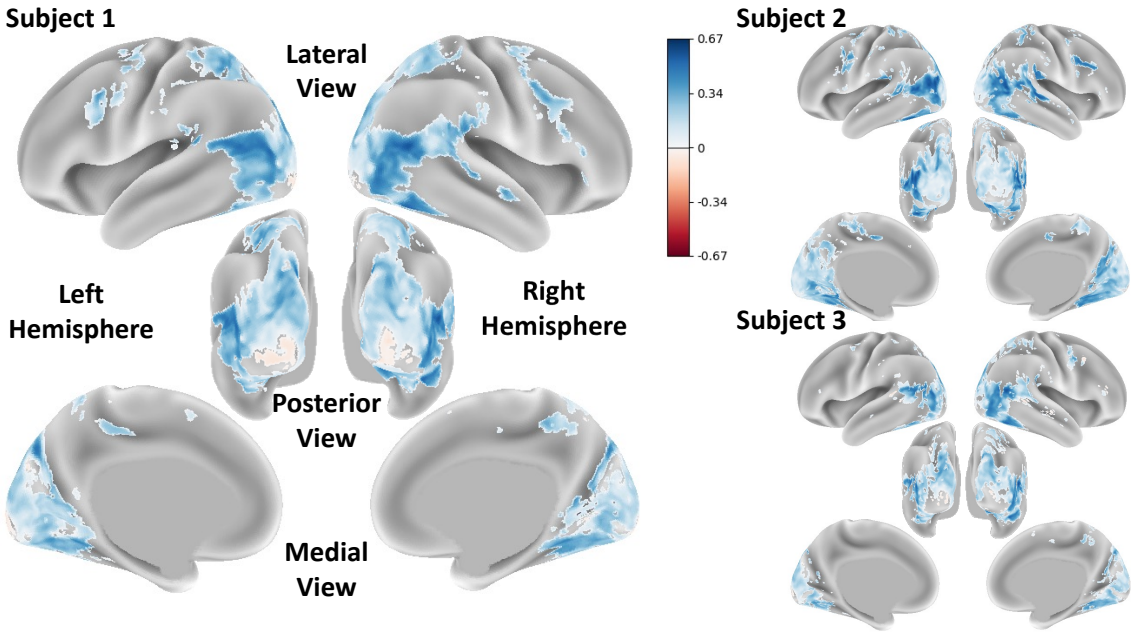


Figure S12. **Hiera Base Plus Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of Hiera Base Plus quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

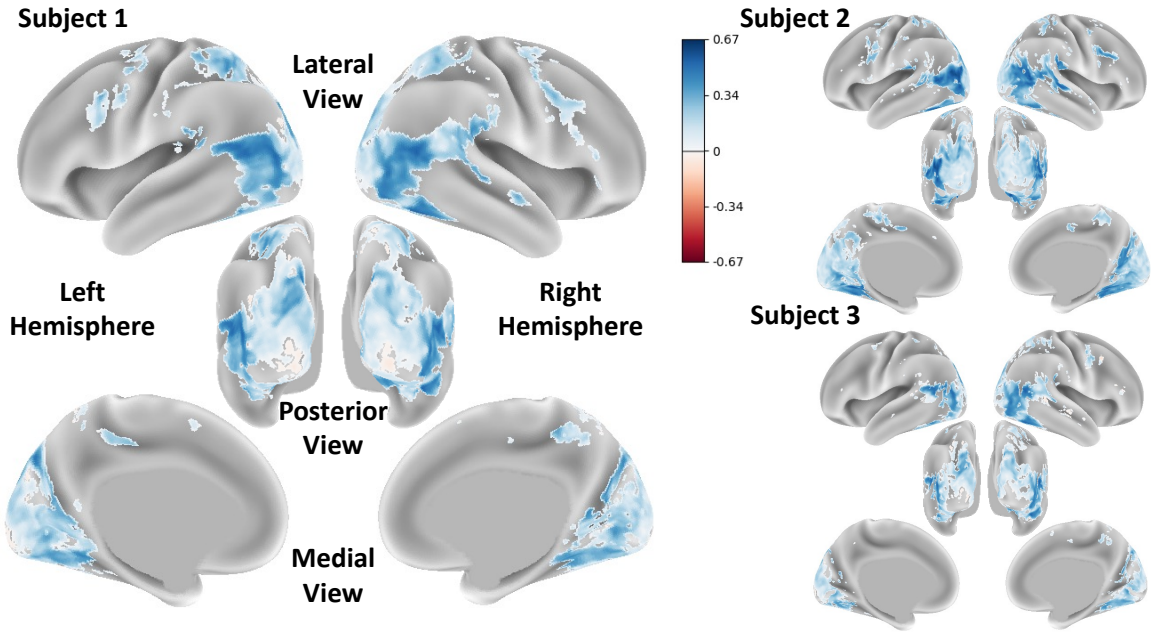


Figure S13. **Hiera Huge Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of Hiera Huge quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

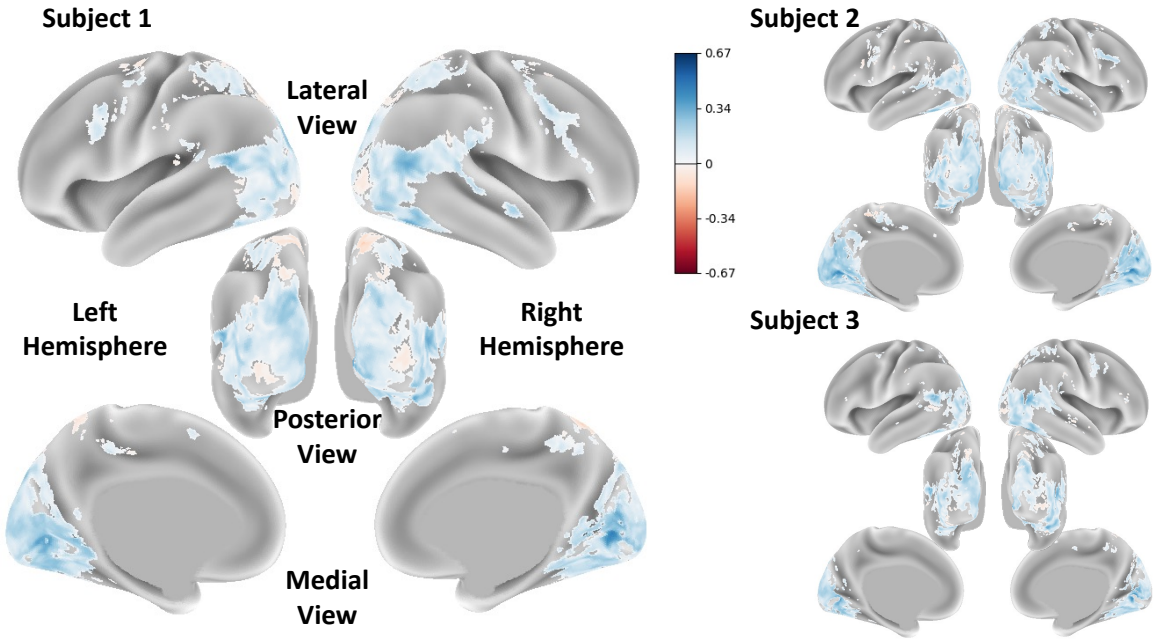


Figure S14. **R3M Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of R3M quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

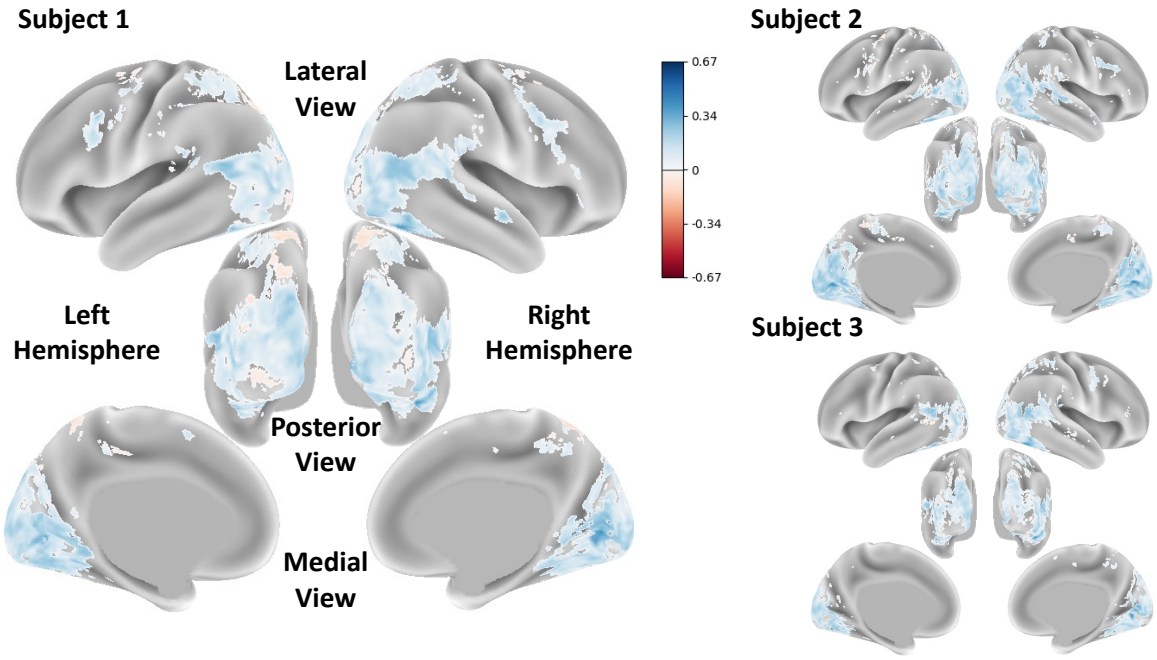


Figure S15. **R3M AF Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of R3M AF quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

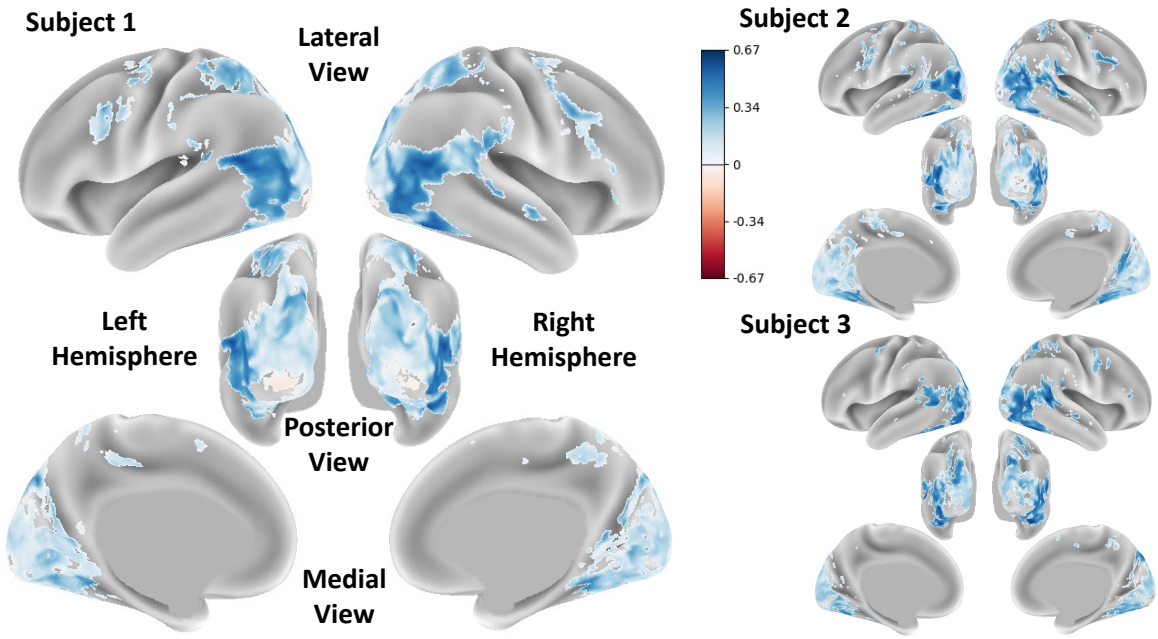


Figure S16. **ResNet50 Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of ResNet50 quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

CVPR
#16804

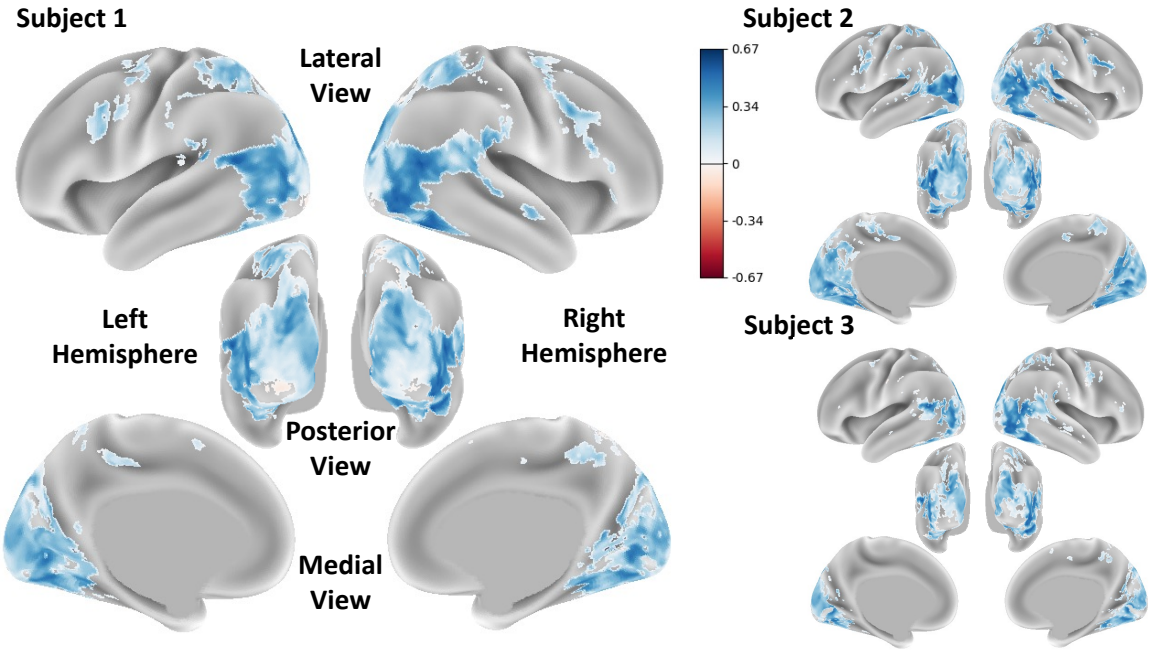


Figure S17. **VC1 Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of VC1 quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

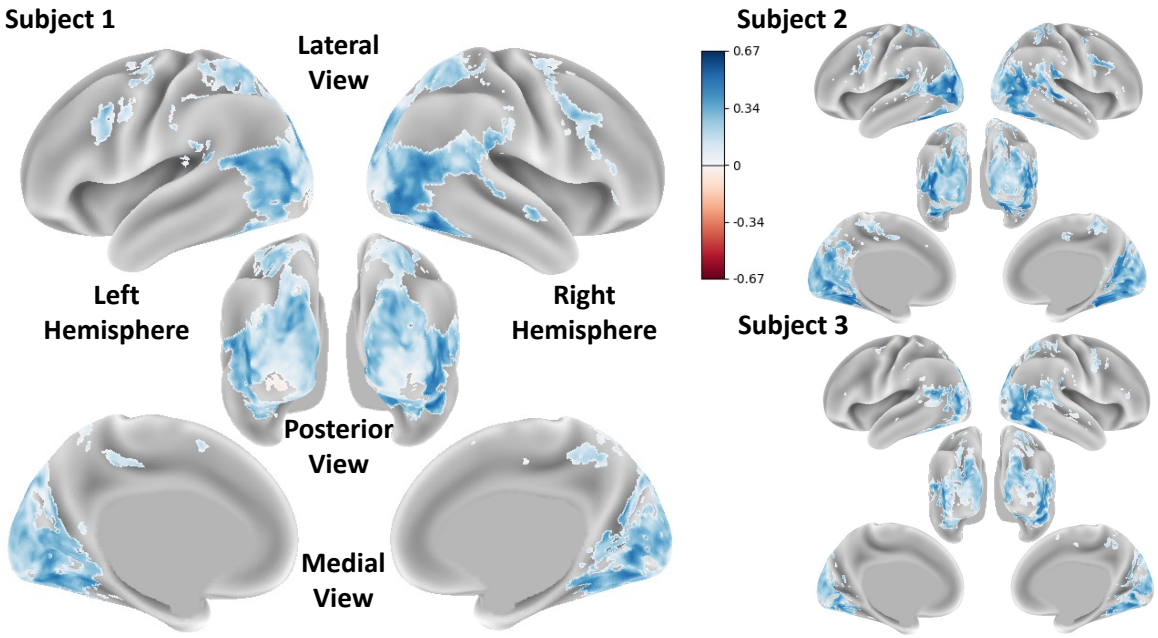


Figure S18. **VC1 AF Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of VC1 AF quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

CVPR
#16804

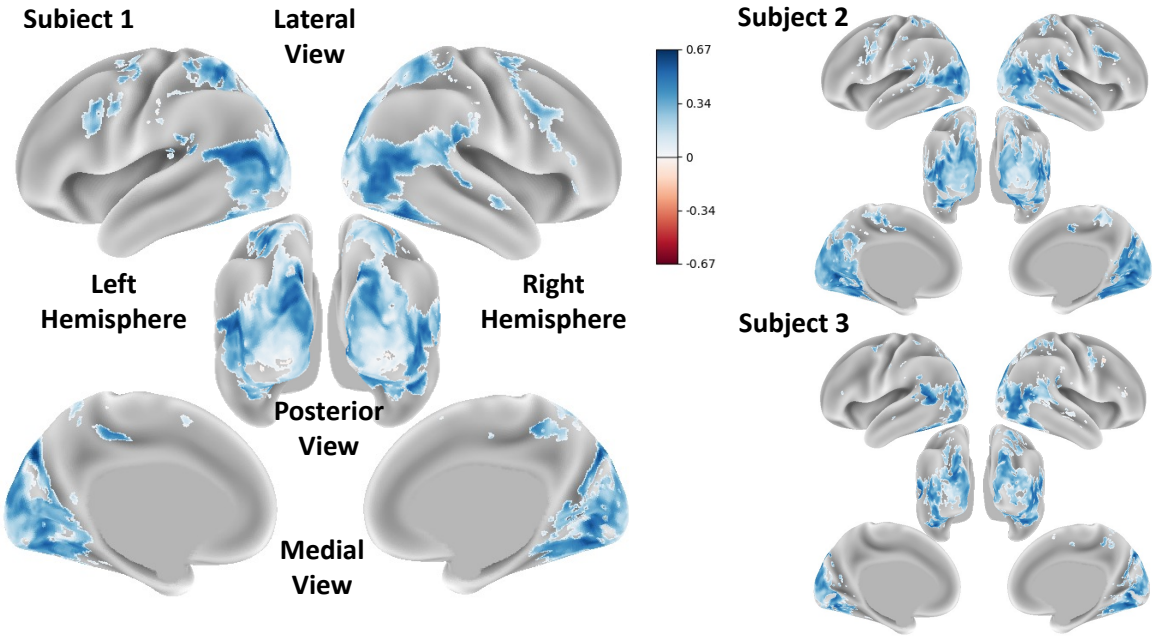


Figure S19. **VideoMAE Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of VideoMAE quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

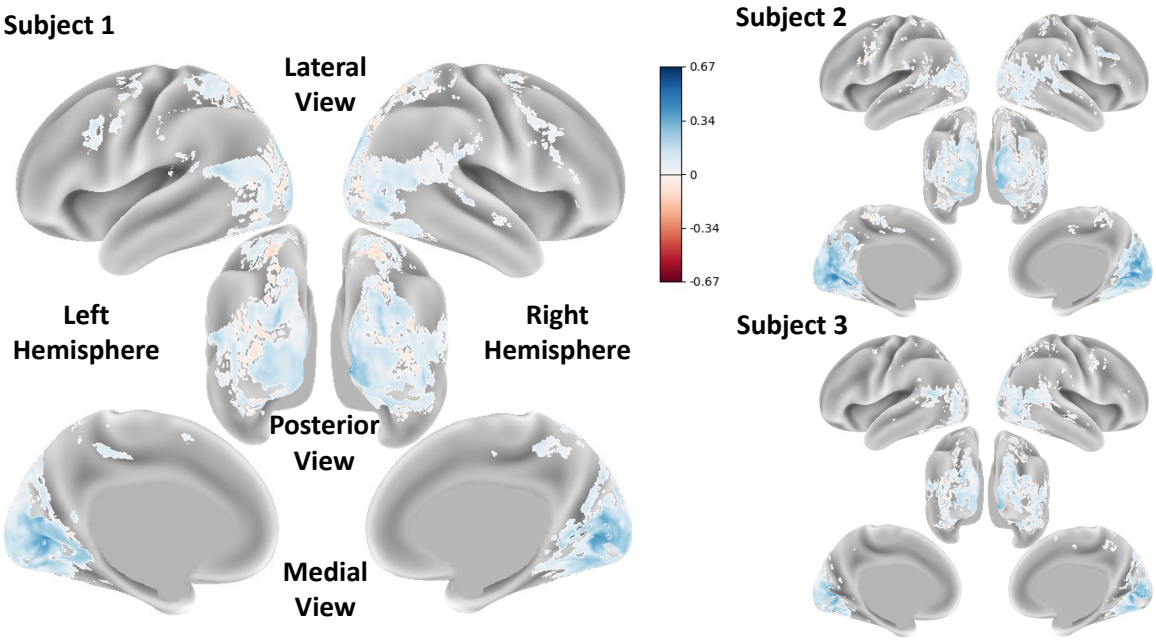


Figure S20. **VIP Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of VIP quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

CVPR
#16804

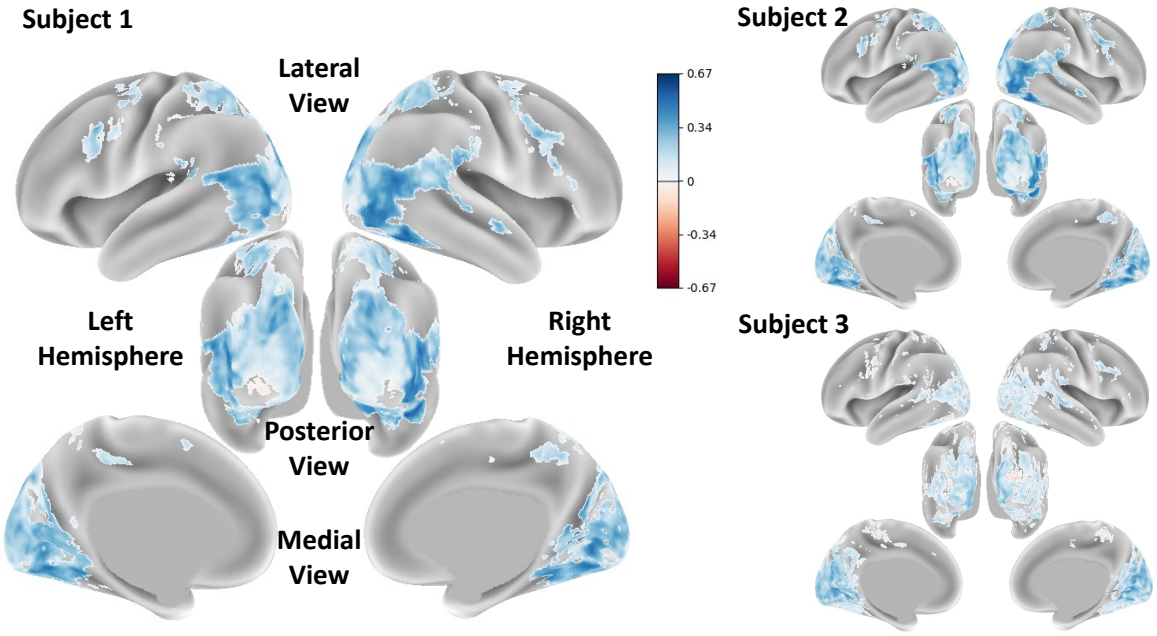


Figure S21. **VIP AF Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of VIP AF quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

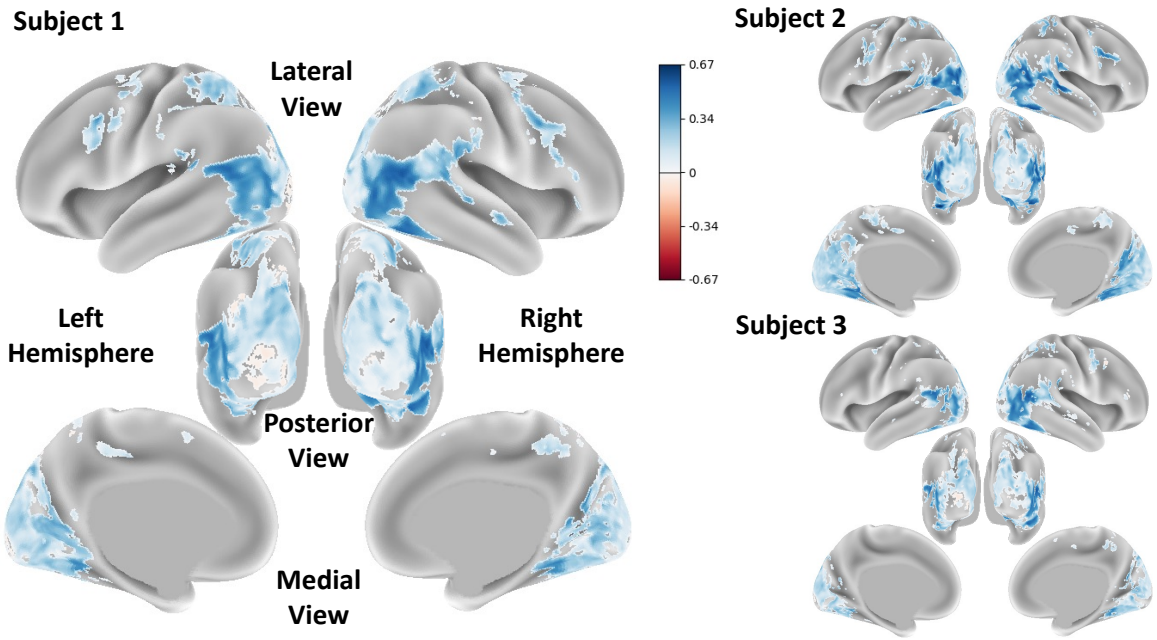


Figure S22. **XCLIP Voxel-wise Prediction Performance.** Voxel-wise fMRI encoding accuracy of XCLIP quantified as the Pearson correlation between measured and predicted responses. Refer to Figure S1 for a labeled parcellation of the human cortex.

References

[1] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. 2

[2] Chu-Chung Huang, Edmund T. Rolls, Jianfeng Feng, and Ching-Po Lin. An extended Human Connectome Project multimodal parcellation atlas of the human cortex and subcortical areas. *Brain Structure and Function*, 227(3):763–778, 2022. 2

[3] Edmund T. Rolls, Gustavo Deco, Chu-Chung Huang, and Jianfeng Feng. The human language effective connectome. *NeuroImage*, 258:119352, 2022. 2

[4] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*, 28(12):4136–4160, 2017. 2

[5] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[6] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. DragNUWA: Fine-grained control in video generation by integrating text, image, and trajectory. 2023. 3

[7] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410, 2011. 4, 5

[8] A. Y. Wang, K. Kay, T. Naselaris, M. J. Tarr, and L. Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, 2023. 4

[9] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, 2024. 4