

Rest2Visual: Predicting Visually Evoked fMRI from Resting-State Scans

Chuyang Zhou¹✉, Ziao Ji¹, Daochang Liu², Dongang Wang¹, Xiaochen Liu¹, Ken-Tye Yong¹, Chenyu Wang¹, and Chang Xu¹

¹ The University of Sydney, Sydney, Australia

² The University of Western Australia, Perth, Australia
chuyang.zhou@sydney.edu.au

Abstract. Understanding how intrinsic brain activity constrains stimulus-evoked responses is a central challenge in cognitive neuroscience. Task-based fMRI measures localized evoked activation but is costly and difficult to scale, whereas resting-state fMRI (rs-fMRI) is easier to acquire yet lacks direct stimulus interpretability. We study resting-conditioned synthesis of visually evoked fMRI (ve-fMRI): given an rs-derived intrinsic functional prior and a visual stimulus image, we predict the corresponding trial-wise activation volume. We propose *Rest2Visual*, a 3D encoder-decoder network that encodes the resting prior into hierarchical volumetric features and injects stimulus embeddings into the decoder via adaptive group normalization to produce stimulus-specific maps. Using large-scale rs-stimulus-evoked triplets, Rest2Visual improves voxel-wise and representational agreement over stimulus-agnostic and modality-ablated baselines, and preserves subject-specific response structure. We further use image reconstruction as a downstream probe to assess whether the synthesized activations retain stimulus information. The results suggest that individualized intrinsic organization can be transformed into stimulus-aligned activation patterns for scalable functional modeling.

Keywords: Deep Learning, Neuroscience, Brain, fMRI.

1 Introduction

Functional magnetic resonance imaging (fMRI) enables noninvasive mapping of stimulus-driven neural responses with high spatial resolution [20,18,21]. Visual paradigms reveal topographic organization and hierarchical processing across cortical areas [28,19]. However, acquiring visually evoked fMRI (ve-fMRI) requires synchronized protocols, sustained subject participation, and prolonged scanning [7,11], limiting scalability in clinical or low-compliance settings.

Resting-state fMRI (rs-fMRI) is easier to obtain but lacks an explicit link to external stimuli. Despite limited interpretability, rs-fMRI exhibits subject-specific, topologically stable connectivity [8,13] and correlates with task-evoked activations [27,6], suggesting that rs-fMRI encodes individualized priors that constrain stimulus-driven responses. This motivates the question: *can an intrinsic prior*

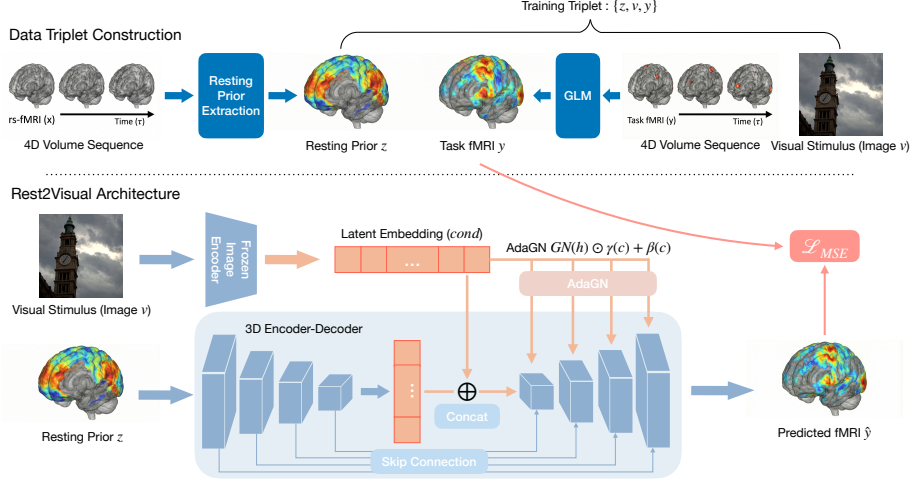


Fig. 1: Overview of Rest2Visual pipeline.

extracted from rs-fMRI be transformed into stimulus-aligned evoked activation when conditioned on visual input?

Prior work on visual brain modeling typically assumes task fMRI and focuses on stimulus-to-activation mapping, from hand-crafted encoding models to deep visual representations aligned with cortical responses [19,12,30,29]. In parallel, large-scale decoders can reconstruct stimuli from evoked activity [26,23,5]. While rs-fMRI has been linked to task activations [27,6], existing methods generally do not combine an rs-derived intrinsic prior with stimulus conditioning to synthesize individualized voxel-wise evoked maps. We instead study *resting-conditioned synthesis*, while exploit image reconstruction as an important downstream probe.

We instantiate this task with *Rest2Visual*, a 3D encoder-decoder where a volumetric encoder extracts hierarchical features from the resting prior and a conditional decoder injects stimulus embeddings via adaptive group normalization to produce stimulus-specific activation maps (Fig. 1). We construct large-scale rs-stimulus-evoked triplets from NSD [2], pairing each resting prior with natural images and GLM-derived evoked maps. Experiments show improved voxel-wise and representational agreement over stimulus-agnostic and modality-ablated baselines while preserving subject-specific response structure, supporting the feasibility of transforming intrinsic organization into stimulus-aligned activations with reduced reliance on task-based acquisition.

2 Methodology

2.1 Data Triplet Construction

We construct our dataset from the Natural Scenes Dataset (NSD) [2], which provides high-resolution 7T fMRI responses from 8 healthy adults viewing thou-

sands of natural scene images sourced from MS COCO [16]. Let $\mathcal{S} = \{s_1, \dots, s_8\}$ denote the set of human participants. For each subject-session pair (s, t) , we extract a resting-state fMRI scan $\mathbf{x}_{\text{raw}}^{(s,t)} \in \mathbb{R}^{T \times D_s \times H_s \times W_s}$ (with T time points, and D_s, H_s, W_s denoting the depth, height and width of the 3D brain volume for subject s , respectively), and a set of stimulus-evoked fMRI beta volumes $\{\mathbf{y}_{\text{raw}_i}^{(s,t)}\}_{i=1}^{750}$ paired with visual stimuli $\{\mathbf{v}_i^{(s,t)}\}_{i=1}^{750}$, where $\mathbf{y}_{\text{raw}_i} \in \mathbb{R}^{D_s \times H_s \times W_s}$ and $\mathbf{v}_i \in \mathbb{R}^{3 \times 425 \times 425}$.

Spatial Registration. To ensure spatial alignment across subjects, all volumes are registered to MNI152 space [9] via continuous interpolation and resampled to 2 mm isotropic resolution. The spatially normalized versions are denoted as $\mathbf{x}^{(s,t)}, \mathbf{y}_i^{(s,t)} \in \mathbb{R}^{D \times H \times W}$.

Resting-state processing. To obtain a subject-specific resting-state connectivity prior that captures intrinsic brain organization, we compute a seed-based GLM map from the posterior cingulate cortex (PCC), a key hub in the default mode network. We define $\Omega \subset \{1, \dots, D\} \times \{1, \dots, H\} \times \{1, \dots, W\}$ which denotes the set of voxels within the brain, as defined by a MNI152 whole-brain mask. Let $\mathcal{R} \subset \Omega$ denote the PCC region. For each session t , we compute the PCC mean time series $\mathbf{r}^{(s,t)}(\tau)$, where τ indexes fMRI time points, by averaging the BOLD signal across all voxels in the PCC region:

$$\mathbf{r}^{(s,t)}(\tau) = \frac{1}{|\mathcal{R}|} \sum_{v \in \mathcal{R}} \mathbf{x}^{(s,t)}(\tau, v), \quad \tau = 1, \dots, T. \quad (1)$$

The denoised PCC trace $\tilde{\mathbf{r}}^{(s,t)}$ is obtained by temporal bandpass filtering and convolving with the canonical hemodynamic response function (HRF). A voxel-wise GLM is then fitted:

$$\mathbf{x}^{(s,t)}(\cdot, v) = \beta^{(s,t)}(v) \cdot \tilde{\mathbf{r}}^{(s,t)} + \varepsilon(v), \quad \forall v \in \Omega, \quad (2)$$

where $\beta^{(s,t)}(v)$ is the regression coefficient for voxel v . We z-score the coefficient as $\mathbf{z}^{(s,t)}(v) = \frac{\beta^{(s,t)}(v) - \mu^{(s,t)}}{\sigma^{(s,t)}}$, where $\mu^{(s,t)}$ and $\sigma^{(s,t)}$ denote the mean and standard deviation of $\{\beta^{(s,t)}(v)\}_{v \in \Omega}$. The resulting map $\mathbf{z}^{(s,t)} \in \mathbb{R}^{D \times H \times W}$ serves as the resting-state intrinsic function map.

Stimulus-evoked data. Each stimulus image $\mathbf{v}_i^{(s,t)}$ is associated with a beta activation map $\mathbf{y}_i^{(s,t)}$ derived via trial-level GLM. To eliminate the extreme outliers introduced by NSD’s liberal masking policy, we apply the MNI152 whole-brain mask as follows $\tilde{\mathbf{y}}_i^{(s,t)}(v) = \mathbf{y}_i^{(s,t)}(v) \mathbf{1}_\Omega(v)$.

Final dataset. The full data pipeline transforms the raw data into a set of triplets of resting-state map, stimulus image, and beta map for model training:

$$\mathcal{D}(\mathbf{x}_{\text{raw}}, \mathbf{y}_{\text{raw}}, \mathbf{v}) = \left\{ \left(\mathbf{z}^{(s,t)}, \mathbf{v}_i^{(s,t)}, \tilde{\mathbf{y}}_i^{(s,t)} \right) \mid s \in \mathcal{S}, t, i \right\}. \quad (3)$$

2.2 Rest2Visual

Given a subject-session pair (s, t) and stimulus index i , Rest2Visual takes a resting-state connectivity prior $\mathbf{z}^{(s,t)}$ and a paired stimulus image $\mathbf{v}_i^{(s,t)}$, and predicts the corresponding evoked beta volume $\hat{\mathbf{y}}_i^{(s,t)}$. We denote the model as $f_\theta : (\mathbf{z}, \mathbf{v}) \mapsto \hat{\mathbf{y}}$, implemented as a conditional 3D U-Net with an image-conditioned decoder:

$$\hat{\mathbf{y}}_i^{(s,t)} = f_\theta(\mathbf{z}^{(s,t)}, \mathbf{v}_i^{(s,t)}) = \mathcal{D}_{\text{cond}}(\mathcal{E}_{\text{vol}}(\mathbf{z}^{(s,t)}), c_i^{(s,t)}), \quad c_i^{(s,t)} = \mathcal{E}_{\text{img}}(\mathbf{v}_i^{(s,t)}). \quad (4)$$

$\mathcal{E}_{\text{img}}$ is a frozen ResNet-50 encoder that maps $\mathbf{v}_i^{(s,t)}$ to a condition vector $c_i^{(s,t)} \in \mathbb{R}^{512}$. $\mathcal{E}_{\text{vol}}$ is a 4-stage 3D convolutional encoder with channel widths $\{32, 64, 128, 256\}$ and $2 \times$ max-pooling at each stage. Each stage consists of two $3 \times 3 \times 3$ convolutions with GroupNorm and GELU, producing hierarchical volumetric features. The decoder mirrors the encoder with transpose-convolution upsampling and concatenation skip connections, and outputs a single-channel volume via a $1 \times 1 \times 1$ convolution.

We inject $c_i^{(s,t)}$ through two pathways. First, at the bottleneck we concatenate it with the bottleneck feature map. Second, we modulate each decoder stage using adaptive group normalization (AdaGN). For an intermediate feature map $h \in \mathbb{R}^{B \times C \times D' \times H' \times W'}$, where $\gamma(\cdot)$ and $\beta(\cdot)$ are learned linear projections producing channel-wise affine parameters,

$$\text{AdaGN}(h; c) = \text{GN}(h) \odot \gamma(c) + \beta(c), \quad \gamma, \beta : \mathbb{R}^{512} \rightarrow \mathbb{R}^C, \quad (5)$$

The model is trained to minimize voxel-wise prediction error between the generated activation $\hat{\mathbf{y}}_i^{(s,t)}$ and the corresponding ground-truth beta volume $\tilde{\mathbf{y}}_i^{(s,t)}$. We use a standard mean squared error (MSE) loss over all brain voxels:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{|\Omega|} \sum_{v \in \Omega} \left(\hat{\mathbf{y}}_i^{(s,t)}(v) - \tilde{\mathbf{y}}_i^{(s,t)}(v) \right)^2. \quad (6)$$

During training, each sample $(\mathbf{z}^{(s,t)}, \mathbf{v}_i^{(s,t)}, \tilde{\mathbf{y}}_i^{(s,t)})$ is independently drawn from the dataset $\mathcal{D}$, and the loss is averaged over the batch. Model parameters θ are optimized to minimize $\mathcal{L}_{\text{MSE}}$ across the dataset:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(s,t,i) \sim \mathcal{D}} \left[\mathcal{L}_{\text{MSE}} \left(f_\theta(\mathbf{z}^{(s,t)}, \mathbf{v}_i^{(s,t)}), \tilde{\mathbf{y}}_i^{(s,t)} \right) \right]. \quad (7)$$

3 Experiments

We train Rest2Visual jointly on all 8 NSD participants, holding out 1,000 triplets uniformly sampled across subjects for testing and splitting the remainder into 90%/10% train/validation (across subjects and sessions). We compare against two non-parametric baselines: Random, spatial Gaussian noise matched to target volume statistics [10,17], and Resting, a stimulus-agnostic prediction using the

subject’s resting-state fMRI [24,6]. We report voxel-wise accuracy (Pearson, MAE, SSIM) and representational structure preservation via RSA, defined as the Pearson correlation between predicted and ground-truth RDMs computed from correlation-distance patterns [14]. Results are reported for whole brain (MNI152 mask; Nilearn [1]) and visual cortex (V1–V5; Julich-Brain atlas [3]). As a downstream probe, we reconstruct stimuli from predicted and ground-truth ve-fMRI using MindEye2 [23] and evaluate reconstructions with low-level (SSIM, AlexNet [15]) and high-level (Inception [25], CLIP [22], SwAV [4]) metrics.

3.1 Main Quantitative Results

Table 1 summarizes performance on the proposed resting-conditioned ve-fMRI synthesis task. Rest2Visual substantially outperforms both sanity-check baselines (Random and Resting) in whole brain and visual cortex across voxel-wise and representational metrics, indicating that the model recovers stimulus-evoked activation patterns beyond background noise and intrinsic topology alone. Improvements are strongest in V1–V5, consistent with these regions being most stimulus-responsive. Overall, the results demonstrate that combining resting priors with visual conditioning yields reliable, stimulus-aligned activation maps.

Table 1: Quantitative results in full brain and visual cortex (mean $\pm$ std).

ROI	Method	Pearson	Corr ($\uparrow$)	RSA ($\uparrow$)	MAE ($\downarrow$)	SSIM ($\uparrow$)
Full Brain	Random	0.0000 $\pm$ 0.0020	0.0016	1.4932 $\pm$ 0.0269	0.6022 $\pm$ 0.0861	
	Resting	0.0104 $\pm$ 0.0560	0.5472	2.0633 $\pm$ 0.0741	0.4134 $\pm$ 0.0947	
	Ours(ns-stim)	0.1752 $\pm$ 0.0662	0.1054	1.4750 $\pm$ 0.0720	0.6266 $\pm$ 0.0738	
	Ours(ns-fmri)	0.3357 $\pm$ 0.1164	0.6245	1.8644 $\pm$ 0.2538	0.4952 $\pm$ 0.1080	
	Ours	0.3861 $\pm$ 0.1238	0.6523	1.2653 $\pm$ 0.1143	0.6935 $\pm$ 0.0727	
Visual Cortex (V1–V5)	Random	0.0000 $\pm$ 0.0071	0.0024	2.1384 $\pm$ 0.2180	0.3087 $\pm$ 0.0902	
	Resting	0.1068 $\pm$ 0.1020	0.6864	2.1745 $\pm$ 0.2017	0.2707 $\pm$ 0.1008	
	Ours(ns-stim)	0.2788 $\pm$ 0.0961	0.0018	2.9324 $\pm$ 0.2037	0.3209 $\pm$ 0.0724	
	Ours(ns-fmri)	0.6089 $\pm$ 0.1527	0.6864	2.1036 $\pm$ 0.3039	0.4696 $\pm$ 0.1108	
	Ours	0.6541 $\pm$ 0.1560	0.8199	2.0350 $\pm$ 0.2991	0.5077 $\pm$ 0.1147	

3.2 ve-fMRI Prediction and Visual Stimulus Reconstruction

We further examine prediction quality beyond aggregate voxel metrics via qualitative neuroanatomical inspection and downstream stimulus reconstruction.

Anatomical consistency. Fig. 2 compares predicted and ground-truth activation maps across multiple anatomical views. Predictions reproduce characteristic occipital activation patterns consistent with visually driven responses and show coherent spatial organization across slices. Qualitatively, the synthesized maps appear less dominated by isolated outlier voxels than the raw targets, suggesting that the model captures structured signal rather than background fluctuations.

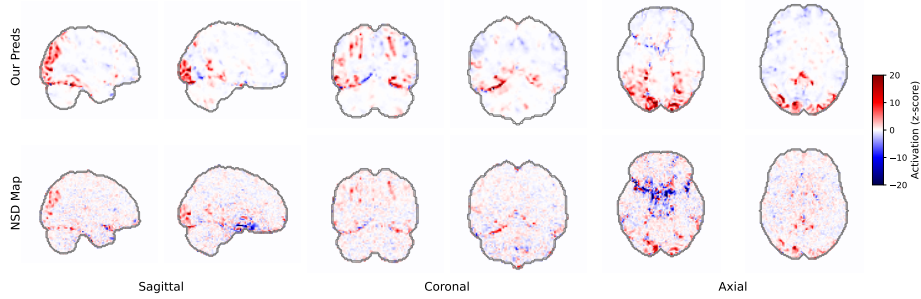


Fig. 2: Visual comparison between predicted visually evoked fMRI maps and the NSD-provided ground truth beta maps across multiple anatomical views.

Perceptual integrity of reconstructed images. We use a pretrained MindEye2 decoder [23] as a downstream probe by zero-shot reconstructing stimuli from predicted and ground-truth ve-fMRI. As shown in Fig. 3, reconstructions from our predictions preserve salient visual attributes. Table 2 quantifies reconstruction similarity in both low-level and semantic feature spaces: using predicted maps yields slightly lower scores than using NSD beta maps, which is expected since MindEye2 is trained on NSD targets and thus favors their statistics.

Table 2: Similarity between reconstructed images and original stimuli, using either NSD-provided ve-fMRI or our predicted ve-fMRI as input.

ve-fMRI	Low-Level			High-Level		
	SSIM ($\uparrow$)	Alex(2) ($\uparrow$)	Alex(5) ($\uparrow$)	Incep ($\uparrow$)	CLIP ($\uparrow$)	SwAV ($\downarrow$)
NSD	0.224 ± 0.156	0.703 ± 0.110	0.555 ± 0.145	0.647 ± 0.116	0.631 ± 0.126	0.444 ± 0.113
Ours	0.192 ± 0.156	0.673 ± 0.115	0.512 ± 0.141	0.601 ± 0.114	0.582 ± 0.122	0.496 ± 0.109

3.3 Individuality Preservation

To assess whether predictions preserve subject-specific response structure, we compute inter-subject correlation matrices over the 125 stimuli shared across all participants (Fig. 4). Each entry is the Pearson correlation between two subjects’ responses aggregated over the shared stimuli. Our predicted matrix closely matches the ground-truth pattern, with low off-diagonal correlations, indicating that predictions do not collapse to a population-average response. In contrast, the stimulus-only baseline (no resting prior) exhibits a different correlation structure, consistent with reduced subject-specific differentiation. These results suggest that conditioning on rs-fMRI enables Rest2Visual to retain individualized functional signatures in synthesized ve-fMRI.



Fig. 3: Stimulus images (top row), reconstruction from ground truth fMRI (middle row) and zero-shot reconstruction from our predicted fMRI (bottom row).

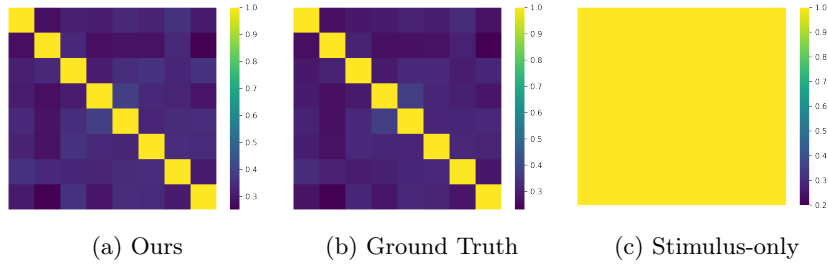


Fig. 4: Inter-subject Pearson correlation of our predicted (a), ground truth (b) and stimulus-only (c) ve-fMRI activation maps, averaged across 125 shared stimuli.

3.4 Ablation Study

We ablate each modality by replacing it with spatial Gaussian noise (Fig. 5). Noisy rs-fMRI (**ns-fMRI**) substantially degrades performance (Table 1) and collapses predictions, confirming that the resting prior provides essential subject-specific spatial organization. Noisy stimuli (**ns-stimuli**) retain coarse spatial structure but remove semantic content in reconstructions, suggesting that stimulus information is needed for stimulus-specific variation beyond the resting scaffold.

Table 3: Generalization evaluation for unseen stimulus and subjects.

Method	Pearson	RSA	MAE	SSIM
Ours (Normal)	0.3861 ± 0.1238	0.6523	1.2653 ± 0.1143	0.6935 ± 0.0727
Ours (Unseen stimuli)	0.3804 ± 0.1243	0.6438	1.2639 ± 0.1130	0.6950 ± 0.0726
Ours (Unseen subject)	0.1361 ± 0.0430	0.2590	1.4714 ± 0.0485	0.6992 ± 0.0377



Fig. 5: Reconstruction under input ablations with spatial Gaussian noise. Left: noisy rs-fMRI (valid stimulus). Right: noisy stimulus (valid rs-fMRI).

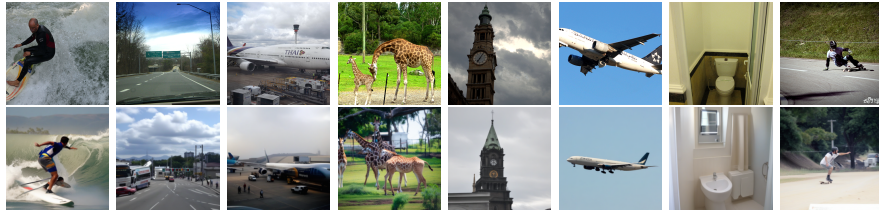


Fig. 6: Stimuli images (top) and reconstructions (bottom) for unseen stimulus.

3.5 Generalization Analysis

Table 3 summarizes generalization to unseen stimuli and a held-out subject. Performance on unseen stimuli remains close to the in-distribution setting across metrics, suggesting Rest2Visual learns transferable stimulus-conditioned mappings rather than memorizing specific images. In contrast, holding out an entire subject leads to a substantial drop, consistent with the subject-conditioned formulation and reliance on individualized resting priors. With only eight NSD participants, learning a fully subject-agnostic mapping remains challenging and is a key direction for future work.

4 Conclusion

We proposed resting-conditioned synthesis of visually evoked fMRI: predicting trial-wise beta maps from an rs-fMRI derived intrinsic prior and a visual stimulus. Rest2Visual combines a 3D encoder-decoder with stimulus conditioning to produce stimulus-aligned activation maps, improving over noise and resting-only baselines, showing strongest gains in visual cortex, and preserving inter-subject individuality. Using zero-shot visual reconstruction as a downstream probe, reconstructions from predicted maps retain stimulus-related content, suggesting that the synthesized activations preserve information useful for downstream decoding. These results support a practical route toward task-reduced functional mapping when task-based acquisition is difficult or infeasible. Several limitations define the current scope of the study. Rest2Visual targets GLM-derived trial-wise beta maps because they summarize stimulus-associated activation patterns while

reducing temporal and acquisition-related variability present in raw BOLD time series. Cross-subject generalization remains limited because the task requires paired resting-state fMRI, visual stimuli, and evoked responses; NSD is among the few suitable public resources but includes only eight subjects. Future studies could further extend this framework by improving subject-adaptive modeling and training on larger and more diverse cohorts.

Acknowledgments. This work was supported by the University of Sydney Digital Sciences Initiative Ignite Grant.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G.: Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics* **8**, 14 (2014)
2. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al.: A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience* **25**(1), 116–126 (2022)
3. Amunts, K., Mohlberg, H., Bludau, S., Zilles, K.: Julich-brain: A 3d probabilistic atlas of the human brain’s cytoarchitecture. *Science* **369**(6506), 988–992 (2020)
4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems* **33**, 9912–9924 (2020)
5. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22710–22720 (2023)
6. Cole, M.W., Ito, T., Bassett, D.S., Schultz, D.H.: Activity flow over resting-state networks shapes cognitive task activations. *Nature neuroscience* **19**(12), 1718–1726 (2016)
7. Dubois, J., Adolphs, R.: Building a science of individual differences from fmri. *Trends in cognitive sciences* **20**(6), 425–443 (2016)
8. Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T.: Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience* **18**(11), 1664–1671 (2015)
9. Fonov, V.S., Evans, A.C., McKinsty, R.C., Almli, C.R., Collins, D.: Unbiased non-linear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, S102 (2009)
10. Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.: Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping* **2**(4), 189–210 (1994)
11. Greene, A.S., Gao, S., Scheinost, D., Constable, R.T.: Task-induced brain state manipulation improves prediction of individual traits. *Nature communications* **9**(1), 2807 (2018)

12. Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L.: Identifying natural images from human brain activity. *Nature* **452**(7185), 352–355 (2008)
13. Kong, R., Li, J., Orban, C., Sabuncu, M.R., Liu, H., Schaefer, A., Sun, N., Zuo, X.N., Holmes, A.J., Eickhoff, S.B., et al.: Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cerebral cortex* **29**(6), 2533–2551 (2019)
14. Kriegeskorte, N., Mur, M., Bandettini, P.A.: Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* **2**, 249 (2008)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*. pp. 740–755. Springer (2014)
17. Lindquist, M.A.: The Statistical Analysis of fMRI Data. *Statistical Science* **23**(4), 439 – 464 (2008). <https://doi.org/10.1214/09-STS282>, <https://doi.org/10.1214/09-STS282>
18. Logothetis, N.K.: What we can do and what we cannot do with fmri. *Nature* **453**(7197), 869–878 (2008)
19. Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L.: Encoding and decoding in fmri. *Neuroimage* **56**(2), 400–410 (2011)
20. Ogawa, S., Lee, T.M., Kay, A.R., Tank, D.W.: Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences* **87**(24), 9868–9872 (1990)
21. Poldrack, R.A., Mumford, J.A., Nichols, T.E.: *Handbook of functional MRI data analysis*. Cambridge University Press (2024)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
23. Scotti, P.S., Tripathy, M., Villanueva, C.K.T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K.A., et al.: Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207* (2024)
24. Smith, S.M., Vidaurre, D., Beckmann, C.F., Glasser, M.F., Jenkinson, M., Miller, K.L., Nichols, T.E., Robinson, E.C., Salimi-Khorshidi, G., Woolrich, M.W., et al.: Functional connectomics from resting-state fmri. *Trends in cognitive sciences* **17**(12), 666–682 (2013)
25. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1–9 (2015)
26. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14453–14463 (2023)
27. Tavor, I., Jones, O.P., Mars, R.B., Smith, S.M., Behrens, T.E., Jbabdi, S.: Task-free mri predicts individual differences in brain activity during task performance. *Science* **352**(6282), 216–220 (2016). <https://doi.org/10.1126/science.aad8127>, <https://www.science.org/doi/abs/10.1126/science.aad8127>

28. Wandell, B.A., Winawer, J.: Imaging retinotopic maps in the human brain. *Vision research* **51**(7), 718–737 (2011)
29. Wen, H., Shi, J., Zhang, Y., Lu, K.H., Cao, J., Liu, Z.: Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex* **28**(12), 4136–4160 (2018)
30. Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences* **111**(23), 8619–8624 (2014)