Scaling Vision Transformers for Functional MRI with Flat Maps

Connor Lane[1,2] **Daniel Z. Kaplan**[1,2] **Tanishq M. Abraham**[1,2] **Paul S. Scotti**[1,2]
[1]Sophont [2]Medical AI Research Center (MedARC)

Abstract

A key question for adapting modern deep learning architectures to functional MRI (fMRI) is how to represent the data for model input. To bridge the modality gap between fMRI and natural images, we transform the 4D volumetric fMRI data into videos of 2D fMRI activity *flat maps*. We train Vision Transformers on 2.3K hours of fMRI flat map videos from the Human Connectome Project using the spatiotemporal masked autoencoder (MAE) framework. We observe that masked fMRI modeling performance improves with dataset size according to a strict power scaling law. Downstream classification benchmarks show that our model learns rich representations supporting both fine-grained state decoding across subjects, as well as subject-specific trait decoding across changes in brain state. This work is part of an ongoing open science project to build foundation models for fMRI data. Our code and datasets are available at https://github.com/MedARC-AI/fmri-fm.

1 Introduction

Functional MRI (fMRI) exploits properties of nuclear magnetic resonance to record a noisy 3D map of a person's brain activity every $\sim$1-2 seconds. A major goal of translational neuroscience is to extract clinically useful information from these remarkable but complicated data [1, 2]. In other domains, "foundation model" [3] approaches to analyzing complex scientific data have made significant progress [4–7]. These approaches, adapted from the broader deep learning community, e.g. [8–11], involve combining large scale data and compute together with flexible neural network architectures and self-supervised learning (SSL) paradigms. *Can we unlock novel clinical applications for brain and mental health by similarly applying this foundation model strategy to fMRI?*

There is growing interest in training foundation models on large-scale fMRI data [12–20]. One of the major considerations when adapting the foundation model paradigm to fMRI is how to format or "tokenize" the data for model input (see also Azabou et al. [21]). Modern neural network architectures such as transformers expect a sequence of embedding vectors as input. Most approaches for tokenizing fMRI first reduce each 3D fMRI volume to a fixed dimension vector by averaging the activity within a set of non-overlapping regions of interest (ROIs) from a standard brain parcellation [22, 23]. The parcellated fMRI time series is then transformed into an input embedding sequence using a linear token embedding. This is a computationally tractable approach leveraging the inductive bias that local cortical neighborhoods are functionally integrated. However, parcellating the native fMRI time series is lossy, reducing the dimensionality by $\sim 100\times$.

At the other extreme, a few works tokenize the native 4D fMRI volume data directly. Both Kim et al. [16] and Wang et al. [20] use an initial 4D convolution to transform the high-resolution 4D time series to a lower resolution 4D grid of embedding vectors, which are then input to a transformer encoder with local window attention [24]. This approach preserves the full information content of the fMRI data, but is more computationally expensive than parcellation-based approaches. Furthermore, the native 4D input representation places a greater burden on the model to learn the intrinsic structure of the data from scratch (e.g. localization of fMRI signal to gray matter, cortical folding, anatomical

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

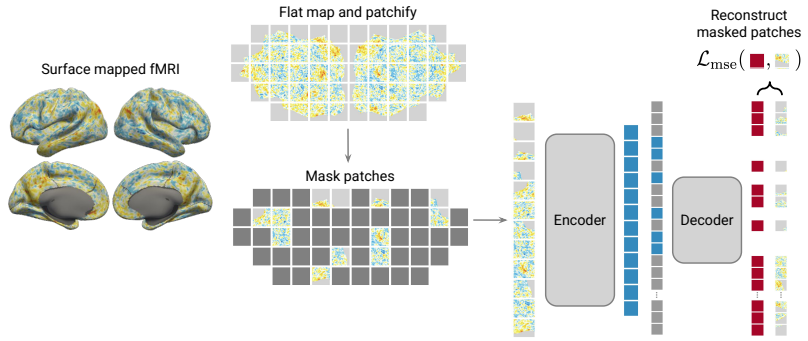


Figure 1: Our flat map MAE (fm-MAE) architecture. Surface-mapped fMRI activity patterns are projected to a flattened cortical mesh [30], resampled as 2D images, and tokenized into patches. We train a standard ViT [31] on temporal sequences of "patchified" flat maps using a spatiotemporal MAE [11, 32]. A large fraction of the image patches are first *masked*. The encoder computes embeddings for the remaining *observed* patches, which are passed to the decoder. The model is trained to minimize the MSE loss between the decoder output and pixel values for masked patches.

and functional networks [25–27]). While the *Bitter Lesson* [28] reminds us that more native, agnostic approaches like this ultimately prevail, they require more data and compute to do so [29].

In this work, we propose an intermediate tokenization strategy that preserves the full dimensionality of the data while eliminating the complexity of modeling fMRI in native 4D volumetric space. Specifically, we represent an fMRI activity time series as a series of 2D maps overlaid on a flattened cortical surface mesh (Figure 1). This flat map representation maintains the full cortical fMRI signal (like native 4D approaches), while also explicitly injecting the inductive bias of local cortical neighborhoods (like parcellation approaches). And crucially, since fMRI flat maps are standard 2D images, they can be tokenized by dividing into square non-overlapping patches ("patchifying"), and modeled using a standard vision transformer (ViT) [31].

To train ViTs on sequences of fMRI flat maps, we adopt the spatiotemporal masked autoencoder (MAE) framework [11, 32]. We pretrain our flat map MAE (**fm-MAE**) using 2.3K hours of publicly available preprocessed fMRI data from the Human Connectome Project (HCP) [33]. We find that masked signal reconstruction improves with increasing pretraining data according to a strict power scaling law—a hallmark of an effective foundation model. To our knowledge, this is the first time that *exact* power law scaling has been observed for an fMRI foundation model. In a preliminary evaluation of our model's downstream decoding performance, we observe "signs of life" that state of the art performance is attainable using this framework. The current work is part of an ongoing open project organized through the MedARC Discord[1], where we invite feedback and collaboration.

2 Method

Flat map data representation. To transform native 4D volume fMRI into sequences of 2D flat maps the data must first be preprocessed using a surface-based fMRI processing pipeline [34–37]. In this work, we use the official surface-preprocessed data provided by the dataset maintainers [33, 38, 39]. The outputs of preprocessing are fMRI data mapped to a group template cortical surface mesh (e.g. fsaverage, fsLR). We copy the surface-mapped data to a corresponding flat surface mesh created by pycortex [30], and resample to a regular image grid using linear interpolation. More details on flat map data generation are in Appendix B.1.

Model architecture. In principle, any modeling approach developed for natural images and video can be applied to fMRI flat maps. In this work, we experiment with the spatiotemporal masked autoencoder (MAE) [11, 32] (Figure 1). Briefly, an MAE consists of a large encoder and smaller decoder ViT [31]. An input image is first divided into a grid of square patches. The encoder receives a sparse subset of *observed* patches, while the remaining patches are removed as *masked*. The encoded latent embeddings for the observed patches are combined with `[MASK]` tokens and passed to the decoder, which predicts pixel values for the masked patches. The model is trained to minimize the

[1] https://discord.gg/tVR4TWnRM9

mean squared error (MSE) between the predicted and masked patches. After pretraining, the decoder is discarded and the encoder is applied to fully observed inputs. To extend from single images to video, the square $p \times p$ patches are expanded to $p_t \times p \times p$ "spacetime" patches, and the learned ViT position embedding is factorized into temporal plus spatial components [32].

One key difference between fMRI flat maps and natural images is the presence of all-zero background pixels that occupy $\sim 40\%$ of the image grid. We exclude entirely empty patches from both encoding and decoding, and compute the MSE loss only for valid, non-background pixels. This is the only significant change required to adapt MAEs to fMRI flat maps.

3 Experiments

3.1 Setup

Dataset. We pretrain our fm-MAE model using the minimally preprocessed data from the Human Connectome Project (HCP) [33, 36]. The dataset includes 21633 fMRI runs collected from 1096 subjects spanning task, resting-state, and movie watching conditions (total scan time 2291 hours). We preprocess the surface-mapped HCP data by normalizing each vertex time series to zero mean unit variance, and temporally resampling to a fixed repetition time (TR) of 1s. We then resample the data to a flat map grid of size 224×560 (1.2mm pixel resolution, 77K valid non-background pixels). To reduce global signal variation [40], we further normalize each *frame* to zero mean unit variance across the spatial grid. The total number of resulting flat map frames is 8.2M. We split the dataset by subject into training (7.4M frames, 979 subjects), validation (0.4M frames, 59 subjects), and test (0.4M frames, 58 subjects) so that family related subjects are assigned to the same split.

Pretraining setup. Inputs are clips of 16 single-channel flat map frames. Our default spacetime patch size is $p_t \times p \times p = 16 \times 16 \times 16$. This means each patch covers the full temporal sequence length ("temporal depth"). We use a default masking ratio of 0.9 (48 visible patches per sample). To prevent the model from interpolating across time, we adopt tube masking from VideoMAE [41]. More details on pretraining are in Appendix B.2.

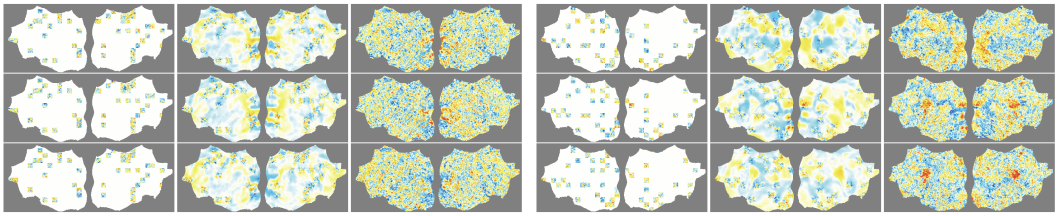
Downstream evaluation tasks. We evaluate our model using two previously used benchmarks: HCP 21 class cognitive state decoding [42–44] and UK Biobank (UKBB) sex classification [16, 18]. We also implement a new CLIP classification benchmark using the Natural Scenes Dataset (NSD) [38]. NSD is a dataset of 8 subjects viewing natural images from MS-COCO [45]. The task is to predict a global image label assigned by CLIP [46] from a set of 41 alternatives (e.g. "photo of dog", see Appendix B.4). Each dataset consists of 16s fMRI flat map clips generated using the same pipeline as for pretraining. For each evaluation, we construct small training, validation, and test sets ($\sim$60K/10K/10K samples). For HCP, we use the same subject splits as in pretraining. For UKBB, we select small random subsets of independent subjects (train: 1645, validation: 248, test: 272). For NSD, we hold out subject 4 for testing and use the remaining 7 subjects for training and validation.

Attentive probe evaluation. We use an *attentive probe* to evaluate the quality of our learned representations [47, 48]. The input to the attentive probe is a sequence of feature embeddings from our pretrained fm-MAE encoder. The attentive probe classifier pools the embeddings into a single global representation by cross-attention with a single learned query vector. The pooled embedding is then passed to a standard linear classifier. Importantly, the encoder is frozen for probe training.

Baseline models. We compare our fm-MAE against two simple baseline models. The first is a *connectome* baseline [49–51]. Given an input clip of fMRI activity, we compute a functional connectivity matrix using the Schaefer 400 parcellation [22] and extract the flattened upper triangle as a feature embedding for a linear classifier. The second is a *patch embedding* baseline. As with our fm-MAE, an input sequence of flat maps is transformed into a grid of embeddings using a learned patch plus position embedding. The embedded patches are then passed directly to an attentive probe.

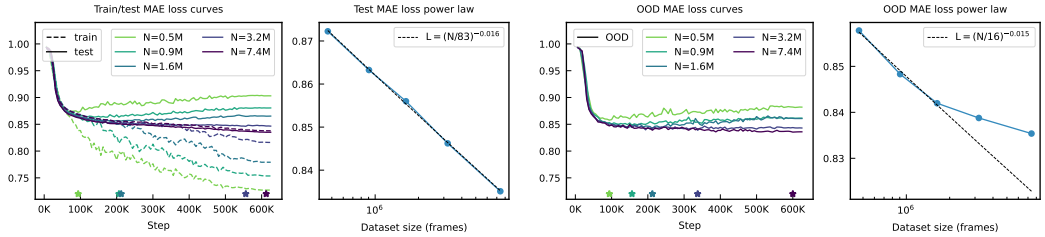
3.2 Masked reconstruction performance

In Figure 2 we visualize the masked reconstructions of our default fm-MAE model (ViT-B, spacetime patch size $16 \times 16 \times 16$) on examples from the HCP and NSD validation sets. Our fm-MAE is able to reconstruct precise fMRI activity patterns given limited context. The predictions are notably



(a) HCP validation set (in distribution) (b) NSD validation set (out-of-distribution)

Figure 2: Visualization of MAE predictions. Within each panel of 3×3 images, we show the masked input (left), MAE prediction (middle), and target data (right). We show predictions for 3 frames spaced 4s apart from top to bottom. The model is a ViT-B with a spacetime patch size of $16 \times 16 \times 16$. RGB color mapping is for visualization only, model inputs and predictions are single channel.



(a) HCP validation set (in distribution) (b) NSD validation set (out-of-distribution)

Figure 3: fMRI modeling performance scales with dataset size. The model is a ViT-B trained on varying size subsets of HCP from N = 500K to 7.4M frames (59 to 979 subjects). Stars indicate epochs with lowest test loss selected for power law estimation. Power law parameters in (b) are fit using only the first 3 loss values to illustrate the deviation from prediction. In-distribution reconstruction obeys a strict power law, whereas OOD reconstruction shows signs of saturating.

smoother compared to the noisy target data. This illustrates how MAEs can function as implicit denoisers [11, 52]. Structured signal can be reconstructed while unstructured noise cannot.

Scaling laws. In Figure 3, we show how masked reconstruction performance scales with pretraining dataset size. We pretrain our default ViT-B on varying size subsets of the HCP training set. In Figure 3a, we observe the expected pattern of greater train/test divergence for smaller subsets, indicating that the over-parameterized ViT-B is able to strongly overfit the undersized datasets. Most importantly, we find that fMRI masked reconstruction performance obeys a strict power law relationship (i.e. "scaling law") with dataset size. This is consistent with now classic work showing that language modeling performance scales log-linearly with the amount of pretraining data [53, 54].

Interestingly, we observe a similar but weaker scaling effect for the out-of-distribution NSD validation set (Figure 3b). Masked reconstruction performance on NSD improves monotonically with more HCP pretraining data, but the rate of improvement slows compared to the power law prediction. This raises the possibility that HCP is *insufficiently diverse* to support learning truly generalizable representations (see also Oquab et al. [55] for discussion of the importance of data diversity).

3.3 Downstream decoding

Effect of dataset size. In Section 3.2, we observed a strong effect of dataset size on masked reconstruction performance, particularly for in-distribution data. For downstream decoding, the effect is weak (Figure 4, left column). The models pretrained on the two largest subsets outperform the three smaller data models. However, the overall trend is not monotonic (let alone log-linear). Notably, the full 7.4M frame model performs the best only for the *in-distribution* HCP state decoding benchmark. The 3.2M frame model performs better for the two OOD benchmarks. This reinforces the possibility that increasing data *scale* without increasing *diversity* does not lead to better representations.

Effect of model size. Surprisingly, we find that relatively small models are sufficient to learn performant representations (Figure 4, middle column). We pretrain fm-MAE ViTs of increasing size on the full HCP training dataset. We find that the 12.4M parameter model performs about as well as

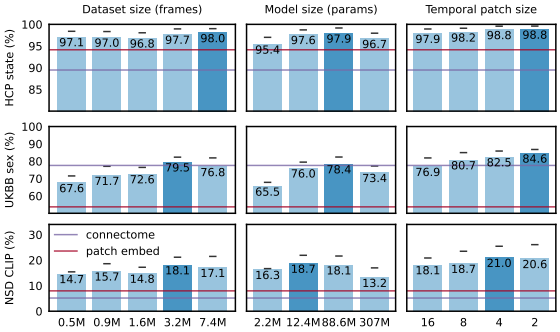


Figure 4: Downstream decoding performance as a function of dataset size (left column), model size (middle column), and temporal patch size p_t (right column). Smaller temporal patch size corresponds to larger effective sequence length (tokens per input $= 364 \cdot 16/p_t$). Black dashes indicate performance on independent validation sets used for classifier parameter tuning.

the 88.6M (ViT-B) model, despite 7× fewer parameters. The largest model (ViT-L) performs notably worse. At the other extreme, we do see a drop for the *very small* 2.2M parameter model.

Effect of temporal patch size. In all previous experiments, the temporal patch size p_t was fixed to 16 frames (the full temporal depth). In Figure 4 (right column) we examine the performance of smaller temporal patch size. Reducing temporal patch size increases the granularity of the model, resulting in more tokens per input. We find that this improves performance across all three benchmarks, suggesting that as with standard ViTs, there is a speed/accuracy tradeoff for smaller patches [56].

HCP state decoding. Due to variation in dataset splits and evaluation protocol, it is difficult to determine a definitive state of the art for this task. To our knowledge, the best reported performance using our same 21-state prediction setup is 93.4% accuracy [43]. NeuroSTORM reports 92.6% accuracy for 23-state prediction [20], while Thomas et al. [13] report 94.8% accuracy on 20-state prediction. We match the performance of these prior methods with just our patch embedding baseline (94.1%), while our best fm-MAE performs notably better, approaching ceiling with 98.8%.

UKBB sex classification. As with HCP state decoding, it is not straightforward to compare UKBB sex classification performance across prior works. Arguably, the current state of the art is Brain-JEPA (88.6%) followed by BrainLM (86.5%) [18]. Our best current model (84.6%) is approaching this performance, while outperforming the model trained from scratch in Dong et al. [18] (82.6%). Importantly, these prior works pretrain on UKBB and fine-tune specifically for UKBB sex classification. By contrast, we pretrain on HCP and use only a small subset of UKBB (60K samples, 1.6K subjects) for training the shallow attentive probe (while the main encoder is kept frozen). Furthermore, prior works use long input sequences (>320s), whereas we use short 16s clips.

NSD CLIP classification. This is a challenging new decoding benchmark without direct comparison, but the current results are nonetheless promising. NSD uses complex natural scene images capturing multiple objects, animals, and people. Predicting a single global label such as "photo of dog" is therefore an ambiguous, ill-posed task. Yet our model performs >8× better than chance and >2× better than our baselines (which themselves are competitive on the other two tasks). Most importantly, this performance is for *zero-shot* visual decoding on an unseen subject (subject 4), taken from an *out-of-distribution* dataset not used for model pretraining. Remarkably, the gap relative to held out data for the training subjects (subjects 1-3, 5-8) is only 4%. This result represents another step toward the long-standing goal of general-purpose cross-subject visual decoding [57–59].

4 Conclusion

In this work, we propose flat maps as a high fidelity yet structured representation for training fMRI foundation models. We train masked autoencoder vision transformers on 2.3K hours of flat-mapped fMRI data from HCP. We observe robust power law scaling with dataset size, and promising early results in downstream decoding evaluations. The current work is a work in progress. Active research directions include incorporating more diverse pretraining data, evaluating the robustness of our initial scaling result, implementing direct comparisons to alternative parcellation and volume based modeling approaches, experimenting with alternative SSL objectives, interrogating the models' learned representations, and expanding the set of downstream evaluation benchmarks. We invite open feedback and collaboration: https://discord.gg/tVR4TWnRM9.

Acknowledgements

We are grateful to fal AI for providing the compute used for this work. We thank MedARC contributors Debojyoti Das, Ratna Sagari Grandhi, Leema Krishna Murali, Manish Ram, Harshil Shah, Utkarsh Singh, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Yuxiang Wei, and Shamus Sim Zi Yang for their active contributions to the ongoing project. We thank MedARC contributors Melvin Selim Atay, Mohammed Baharoon, Atmadeep Banerjee, Uday Bondi, Pierre Chambon, Alexey Kudrinsky, Souvik Mandal, Ashutosh Narang, Alex Nguyen, Yashvir Sabharwal, Kevin Son, and Dingli Yu for contributing to an earlier version of this project. We thank Zijao Chen, Gregory Kiar, and Florian Rupprecht for helpful discussions on an earlier version of this work. We thank the two anonymous workshop reviewers for helpful comments.

References

[1] John DE Gabrieli, Satrajit S Ghosh, and Susan Whitfield-Gabrieli. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, 85(1):11–26, 2015.

[2] Choong-Wan Woo, Luke J Chang, Martin A Lindquist, and Tor D Wager. Building better biomarkers: brain models in translational neuroimaging. *Nature neuroscience*, 20(3):365–377, 2017.

[3] Rishi Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

[5] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

[6] Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, pages 1–8, 2025.

[7] Eric Y Wang, Paul G Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, et al. Foundation model of neural activity predicts response to new stimulus types. *Nature*, 640(8058):470–477, 2025.

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

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460, 2020.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[12] Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599, 2022.

[13] Armin Thomas, Christopher Ré, and Russell Poldrack. Self-supervised learning of brain dynamics from broad neuroimaging data. *Advances in neural information processing systems*, 35:21255–21269, 2022.

[14] Itzik Malkiel, Gony Rosenman, Lior Wolf, and Talma Hendler. Self-supervised transformers for fmri representation. In *International Conference on Medical Imaging with Deep Learning*, pages 895–913. PMLR, 2022.

[15] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. 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*, pages 22710–22720, 2023.

[16] Peter Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. Swift: Swin 4d fmri transformer. *Advances in Neural Information Processing Systems*, 36:42015–42037, 2023.

[17] Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed A Rizvi, Matteo Rosati, Christopher Averill, James L Cross, Prateek Mittal, Emanuele Zappala, Rahul Madhav Dhodapkar, Chadi Abdallah, and David van Dijk. BrainLM: A foundation model for brain activity recordings. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RwI7ZEfR27.

[18] Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Chong, Fang Ji, Nathanael Tong, Christopher Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

[19] Mohammad Javad Darvishi Bayazi, Hena Ghonia, Roland Riachi, Bruno Aristimunha, Arian Khorasani, Md Rifat Arefin, Amin Darabi, Guillaume Dumas, and Irina Rish. General-purpose brain foundation models for time-series neuroimaging data. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024. URL https://openreview.net/forum?id=HwDQH0r37I.

[20] Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Changbae Bang, Lin Zhao, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, et al. Towards a general-purpose foundation model for fmri analysis. *arXiv preprint arXiv:2506.11167*, 2025.

[21] Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36: 44937–44956, 2023.

[22] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.

[23] Kamalaker Dadi, Gaël Varoquaux, Antonia Machlouzarides-Shalit, Krzysztof J Gorgolewski, Demian Wassermann, Bertrand Thirion, and Arthur Mensch. Fine-grain atlases of functional modes for fmri analysis. *NeuroImage*, 221:117126, 2020.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[25] Olaf Sporns, Giulio Tononi, and Rolf Kötter. The human connectome: a structural description of the human brain. *PLoS computational biology*, 1(4):e42, 2005.

[26] BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011.

[27] James C Pang, Kevin M Aquino, Marianne Oldehinkel, Peter A Robinson, Ben D Fulcher, Michael Breakspear, and Alex Fornito. Geometric constraints on human brain function. *Nature*, 618(7965): 566–574, 2023.

[28] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

[29] Hyung Won Chung. Stanford cs25: V4. https://youtu.be/3gb-ZkVRemQ?si=7FXnklTS9X3FCuv1, 2024. YouTube video, Stanford University.

[30] James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, 9:23, 2015.

[31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[32] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

[33] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80: 62–79, 2013.

[34] Anders M Dale, Bruce Fischl, and Martin I Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999.

[35] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[36] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

[37] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

[38] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[39] Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018.

[40] Jonathan D Power, Mark Plitt, Timothy O Laumann, and Alex Martin. Sources and implications of whole-brain fmri signals in humans. *Neuroimage*, 146:609–625, 2017.

[41] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023.

[42] Yu Zhang, Loïc Tetrel, Bertrand Thirion, and Pierre Bellec. Functional annotation of human cognitive states using deep graph convolution. *NeuroImage*, 231:117847, 2021.

[43] Yu Zhang, Nicolas Farrugia, and Pierre Bellec. Deep learning models of cognitive processes constrained by human brain connectomes. *Medical image analysis*, 80:102507, 2022.

[44] Shima Rastegarnia, Marie St-Laurent, Elizabeth DuPre, Basile Pinsard, and Pierre Bellec. Brain decoding of the human connectome project tasks in a dense individual fmri dataset. *NeuroImage*, 283:120395, 2023.

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

[47] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[48] Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latents patches for improved masked image modeling. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Ycmz7qJxUQ.

[49] Michelle Hampson, Naomi R Driesen, Pawel Skudlarski, John C Gore, and R Todd Constable. Brain connectivity related to working memory performance. *Journal of Neuroscience*, 26(51):13338–13343, 2006.

[50] Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664–1671, 2015.

[51] Tong He, Lijun An, Pansheng Chen, Jianzhong Chen, Jiashi Feng, Danilo Bzdok, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nature neuroscience*, 25(6):795–804, 2022.

[52] Dayang Wang, Yongshun Xu, Shuo Han, and Hengyong Yu. Masked autoencoders for low-dose ct denoising. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2023.

[53] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[54] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[55] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

[56] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14496–14506, 2023.

[57] Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *Forty-first International Conference on Machine Learning*, 2024.

[58] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11333–11342, 2024.

[59] Yuqin Dai, Zhouheng Yao, Chunfeng Song, Qihao Zheng, Weijian Mai, Kunyu Peng, Shuai Lu, Wanli Ouyang, Jian Yang, and Jiamin Wu. Mindaligner: Explicit brain functional alignment for cross-subject visual decoding from limited fMRI data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=1W2WlYRqOK.

[60] Daniel S Marcus, Michael P Harms, Abraham Z Snyder, Mark Jenkinson, J Anthony Wilson, Matthew F Glasser, Deanna M Barch, Kevin A Archie, Gregory C Burgess, Mohana Ramaratnam, et al. Human connectome project informatics: quality control, database services, and data visualization. *Neuroimage*, 80:202–219, 2013.

[61] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[62] Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, et al. Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23:S208–S219, 2004.

[63] Karthik Gopinath, Douglas N Greve, Sudeshna Das, Steve Arnold, Colin Magdamo, and Juan Eugenio Iglesias. Cortical analysis of heterogeneous clinical brain mri scans for large-scale neuroimaging studies. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2023.

[64] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[65] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[66] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020.

[67] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[68] Ken Shirakawa, Yoshihiro Nagano, Misato Tanaka, Shuntaro C Aoki, Yusuke Muraki, Kei Majima, and Yukiyasu Kamitani. Spurious reconstruction from brain activity. *Neural Networks*, page 107515, 2025.

A Author contributions

Connor Lane conceived and implemented the flat map strategy, developed the project framing, wrote the majority of the code, trained all the models, ran all the analyses, led the writing of the paper, and is leading the ongoing project. **Daniel Z. Kaplan** provided technical feedback and developed compute infrastructure. **Tanishq M. Abraham** provided technical advice, coordinated compute, and co-supervised the project. **Paul S. Scotti** proposed and organized the initial project, coded early implementations based around VideoMAE [41], coordinated data acquisition and compute, and co-supervised the project. All authors reviewed and edited the paper.

B Additional methods

B.1 Flat map construction

We use the precomputed `fsaverage` flat map distributed with pycortex [30], which we resample onto the `32k_fs_LR` template mesh using the connectome workbench [60, 36]. We exclude vertices with a non-zero z component in flat map coordinates, and intersect with the Schaefer-1000 parcellation mask [22] to yield a valid flat map mask of containing 58212 vertices across both cortical hemispheres. We fit a regular grid of size height $\times$ width $= 224 \times 560$ to the array of (x, y) points contained in the mask. The grid has a pixel resolution of 1.2mm in flat map coordinates, which equals the mean nearest neighbor distance. To project surface-mapped fMRI data onto the flat map grid, we extract the array of values corresponding to our flat map vertex mask and then resample using linear interpolation (`scipy.interpolate.LinearNDInterpolator`) [61]. After resampling, there are 77763 pixels contained in the flat map mask. The correspondence between surface and flat map space is illustrated in Figure 6 using the Yeo resting-state networks overlaid on the Schaefer 400 parcellation [26, 22].

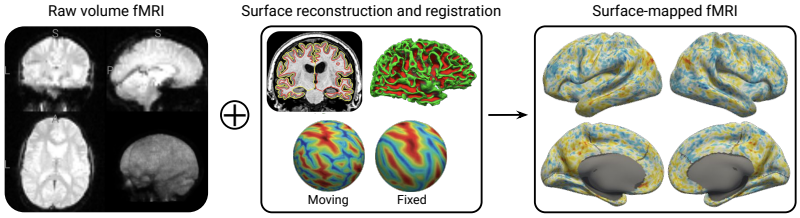


Figure 5: 4D fMRI time series are first preprocessed using standard methods [62]. The cortical surface mesh is reconstructed using structural MRI and aligned to a standard surface template [34, 35]. The fMRI data are then extracted for the cortical ribbon and resampled to the standard surface [36]. This processing was performed by the dataset providers [33, 39, 38]. Middle figure adapted from Gopinath et al. [63].

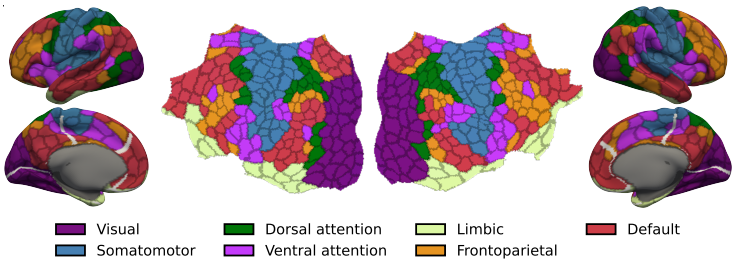


Figure 6: Schaefer 400 parcellation [22] with Yeo resting-state networks [26] on the cortical surface and flat map. Relaxation cuts required for flat map transformation [30] are marked in white.

B.2 Pretraining implementation details

We pretrain for 625K steps using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$) [64] with a batch size of 32, learning rate of 1.25e-4 (base learning rate 1e-3 scaled by `batch_size` / 256), and weight decay

0.05. We apply learning rate warmup for 31K steps followed by cosine decay [65]. In total, the model sees 320M fMRI frames during pretraining, which is ∼43 effective epochs over our HCP training set. We use repeated sampling [32, 66] to improve data loading throughput. Each time an fMRI run is loaded from disk, we extract $4 \cdot N_t/16$ random clips, where N_t is the length of the run. The clips are then appended to an in-memory shuffle buffer, which we sample from to construct training batches. One pretraining run (ViT-B, $p_t = 2$, 88.6M encoder params, 99.2M total) takes ∼27 hours using 1 NVIDIA H100 GPU (16GB memory usage, 130ms/step).

B.3 Probe evaluation implementation details

We use the same protocol to train both the attentive probe for our fm-MAE as well as the connectome and patch embedding baseline models. The protocol is adapted from Darcet et al. [48]. We train for 20 epochs using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$) with a batch size of 128 and base learning rate 5e-4. We apply learning rate warmup for 2 epochs followed by cosine decay [65]. We train a sweep of models over a grid of learning rate scale = [0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0] and weight decay [3e-4, 0.001, 0.01, 0.03, 0.1, 0.3, 1.0], and choose the best hyperparameter setting based on validation accuracy. The effective learning rate is set to be the learning rate scale × 5e-4.

B.4 NSD CLIP classifcation benchmark

To construct the NSD CLIP classification benchmark, we assign each seen NSD stimulus image a global label by CLIP (ViT-L/14) [46] nearest neighbor assignment over a set of 41 short captions (Table 1). The task is then to predict the assigned label from the fMRI activity. We constructed the list of target captions by clustering the CLIP embeddings for all NSD images and manual inspecting the UMAP projection [67], following Shirakawa et al. [68].

photo of zebra	photo of bear	photo of dog	photo of computer
photo of giraffe	photo of bike	photo of sweets	photo of umbrella
photo of horse	photo of toy	photo of sports	photo of baseball
photo of bedroom	photo of cow	photo of group of people	photo of pizza
photo of sky	photo of elephant	photo of fruits	photo of living room
photo of vehicle	photo of surfer	photo of hydrant	photo of stop sign
photo of train	photo of tennis	photo of cat	photo of bus
photo of bathroom	photo of soccer	photo of boat	photo of person eating
photo of food	photo of airplane	photo of skate	photo of sheep
photo of clocktower	photo of flower	photo of ski	photo of bird
photo of a person			

Table 1: List of 41 label categories for NSD CLIP classification.

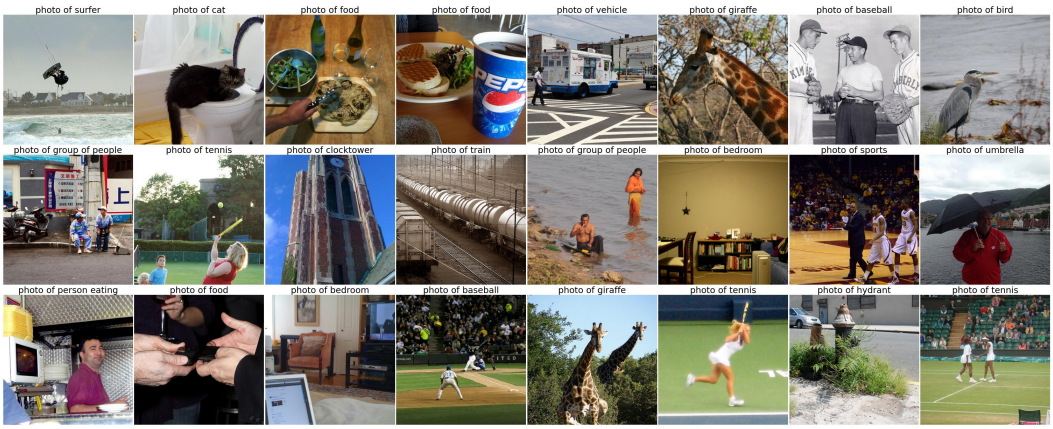
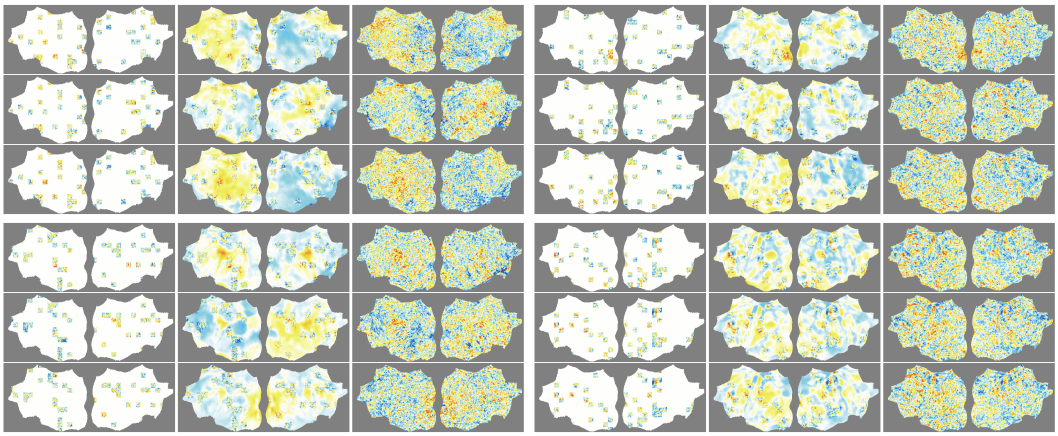


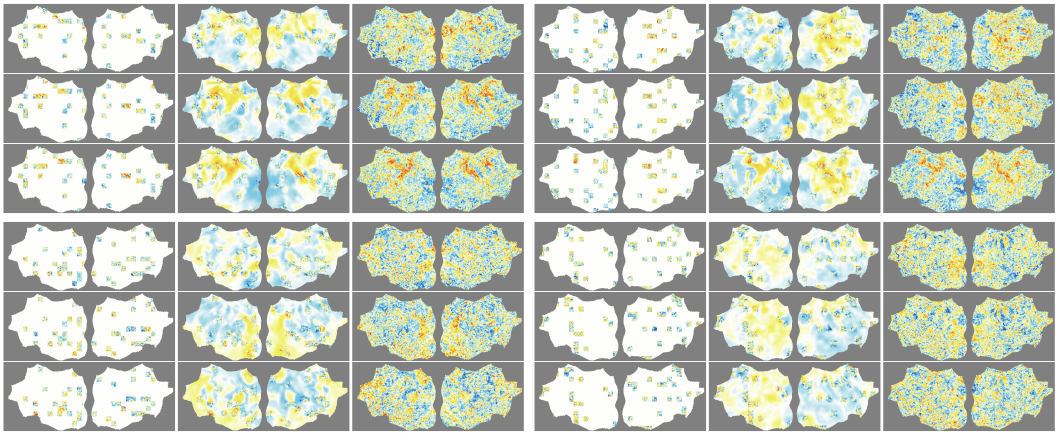
Figure 7: Example NSD images with CLIP assigned labels.

C Additional results

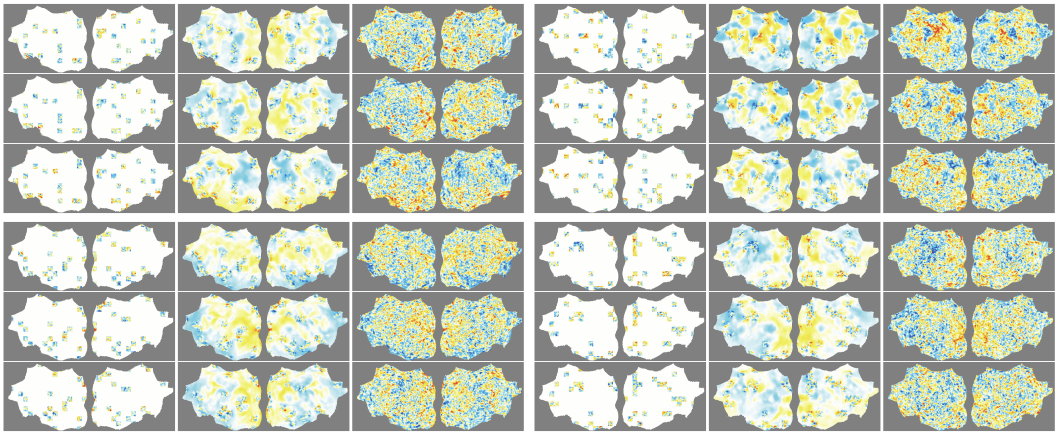
Figure 8 shows additional MAE masked reconstructions for randomly sampled data from HCP, NSD, and UKBB.



(a) HCP validation set (in distribution)



(b) NSD validation set (out-of-distribution)



(c) UKBB validation set (out-of-distribution)

Figure 8: Additional MAE predictions for randomly sampled data.