Supplemental for Brain Netflix: Scaling Data to Reconstruct Videos from Brain Signals

Camilo Fosco*[1], Ben Lahner*[1], Bowen Pan[1], Alex Andonian[1], Emilie Josephs[1], Alex Lascelles[1], and Aude Oliva[1]

Massachusetts Institute of Technology, MA, USA
[camilolu, blahner, bpan, andonian, ejosephs, alexlasc, oliva]@mit.edu

1 fMRI Preprocessing and Data Preparation

This work achieves dynamic reconstruction by leveraging data-hungry computational techniques and diverse human visual responses. To achieve the necessary scale of neural data, we aggregate fMRI responses across four large-scale datasets - BMD [16], HAD [31], HCP [25], and CC2017 [30] - that encompass responses from resting-state, short video, and long form movie tasks. This aggregation of fMRI datasets leads to preprocessing challenges. Specifically, the masked brain modeling [13] [29] approach learns spatial patterns in brain activity, thus requiring all fMRI data, regardless of subject or dataset, to share a common spatial space. Furthermore, alignment between neural responses and video representations necessitates short and lognform video movie fMRI data be transformed into discrete fMRI response - stimuli pairs.

To overcome these challenges, we register all datasets to the fsLR32k grayordinate cortical surface using either the full HCP preprocessing pipeline [10] (as in HCP and CC2017) or Ciftify [4] (as in BMD and HAD). In this way, all data share a common space where each grayordinate vertex corresponds to the same spatial cortical location irrespective of the dataset's acquisition voxel size or other acquisition parameters. Short video responses from event-related experimental designs (BMD and HAD) were estimated with a general linear model (GLM) to achieve fMRI response - stimuli pairs. fMRI response - stimuli pairs were achieved for the long form movie dataset (CC2017) by segmenting the movie every 2 seconds (corresponding to the acquisition TR of CC2017) and using the time-series value offset by 4 seconds (to account for the temporally delayed BOLD response), as done in [27]. Across all datasets, we use the vertices defined by the 41 ROIs [9] in Appendix table 1. Additional dataset specific details are provided below, and please refer to the original manuscript for acquisition and preprocessing protocols.

HCP Preprocessing We use the resting state fMRI data from the 1200-subject release of the Human Connectome Project [25] to train the Masked-brain Model (MBM). In this release, 1084 subjects had resting state data available. We use the released "rfMRI_RESTX_LR_Atlas_MSMAll_hp2000_clean.dtseries.nii" files for each subject, where X is the run number between 1 and 4. The resting state

time series was normalized across the temporal dimension and averaged over a 10 second window for training the MBM.

BOLD Moments Dataset Preprocessing and Preparation We obtained permission from the authors of BMD [16] to use the data corresponding to Version B of the BMD release. The data was registered to the fsLR32k cortical mesh using Ciftify [4] and tools from the Human Connectome preprocessing pipeline [10]. The preprocessed data in fsLR32k grayordinate space was temporally interpolated from an acquisition TR of 1.75s to an interpolated TR of 1s to time-lock stimulus onset to image acquisition (e.g., 1.75s does not evenly go into the inter-trial interval of 4s). In this way, we acquire fMRI scans at different timepoints along the BOLD signal (with respect to stimulus presentation) and, after interpolating, achieve a regular sampling of the BOLD signal time-locked to stimulus onset for easier analysis. The interpolated fMRI time series, stimuli onsets, and stimuli durations (modeled as a 0s impulse) for each session separately were input to the general linear model. GLMsingle [22] estimated single trial beta values by (1) fitting an optimal HRF to each voxel from a library of HRFs, (2) identifying nuisance regressors from a noise pool that maximally explain variance, and (3) implementing fractional ridge regression to improve estimates in a rapid event-related design. Responses to testing and training videos were estimated separately.

In this way, we obtained one beta estimate for each stimulus presentation for each subject. This resulted in a total of 4,020 beta estimates per subject (3 beta estimates x 1,000 training videos and 10 beta estimates x 102 testing videos).

For the training and testing data separately, the beta estimates were then z-scored across video conditions such that the response profile for each stimulus presentation at each voxel (i.e., a vector of length 1000 for training data or of length 102 for testing data) had a mean value of 0 and standard deviation of 1.

Human Actions Dataset Preprocessing and Preparation We use the data made available by the HAD authors [31] that was converted to Cifti format using Ciftify [4]. In this way, the data was preprocessed with fMRIPrep [5] and registered to the fsLR32k cortical mesh. For each subject, we use a General Linear Model with GLMsingle [22] to estimate single trial beta responses. Since no stimuli were repeated, we use the GLMsingle typeB to fit an optimal HRF (step 1) but do not perform the additional nuissance regressor identification (step 2) or fractional ridge regression (step 3). We model the duration of the stimulus as an impulse of zero seconds. We the z-score the resulting beta values across stimuli conditions for each subject separately.

This procedure resulted in 720 beta estimates per subject for a total of 21,600 beta estimates for 21,600 unique videos in the dataset.

CC2017 Dataset Preprocessing and Preparation We use a publicly available preprocessed version of the CC2017 in Cifti format made available by the

authors [30]. This data version was preprocessed with the HCP preprocessing pipeline [10]. We divide each 8 minute test and train segment into 2 second clips for a total of 5,497 unique clips (after accounting for technical scanner errors as described in the original report). In order to obtain fMRI-stimulus pairs from the continuous movie presentation, we index the time series value 4 seconds after the onset on the corresponding 2 second clip as done in [27]. The 4 second offset roughly corresponds to the peak of the BOLD signal supposedly evoked at the clip's first frame but is also influenced by the BOLD signal of subsequent frames.

The values for each 2 second clip were then z-scored across stimuli for each subject separately. While the time series estimates used here are fundamentally different measurements than the beta estimates used in BMD and HAD, they both capture relevant aspects of the BOLD signal and are z-scored into the same range.

1.1 Regions of interest definition

We use the region of interest (ROI) definitions of the Glasser atlas, indexed from the hcp-utils python package [9]. The Glasser atlas [9] used structural, task-based functional, and resting-state connectivity neural data to parcellate the human cortex into a set of 180 ROIs across 22 major sections. We selected a subset of 41 ROIs to constrain analyses to regions that are likely to respond to dynamic stimuli [26] [18] [17] [8] [23] [24] [20] [21] [28] and reduce the computational load on analyses. The 41 ROI names, their ROI ID, and ROI Group Number are listed in Table 1.

These 41 ROIs were selected in an unbiased manner to thoroughly sample visual regions and beyond that likely contribute to dynamic visual perception. This broad sampling recognizes that the brain is comprised of distributed interconnected networks, but it also takes advantage of the fact that most networks contributing to visual perception reside in and around the visual cortex [7] [11] [6]. This point is further evidenced by the within-subject correlations shown in supplementary Figure 2, Figure 3, and various reliability analyses in other large fMRI datasets [1] [3] [14] [31]. Thus, using all vertices across the whole brain would likely introduce noisy signal at a large computational cost (computation in attention layers scales quadratically).

We demonstrate the tradeoff between regression MSE (a proxy for reconstruction accuracy) and computational resources in four ROI groupings:

- our Group41 set (13,156 vertices) $\rightarrow$ 0.721
- core vision (6,549 vertices from Glasser atlas Group Numbers 1-4) $\rightarrow$ 0.755
- the average of 10 randomly selected sets of 41 ROIs (average of 12,390 vertices) $\rightarrow$ 0.983
- whole brain (59,412 vertices) $\rightarrow$ 0.717

The whole brain ROI shows a slight improvement over the Group41 ROI set for a 5x greater computational cost. Furthermore, this analysis shows our Group41 ROI set performs significantly better than both a random sampling of ROIs and efficiently samples informative signal outside the core visual ROIs.

ROI	ID	Group Number
V1	1	1
MST	2	5
V6	3	3
V2	4	2
V3	5	2
V4	6	2
MT	23	5
V8	7	4
V3A	13	3
RSC	14	18
POS2	15	18
V7	16	3
IPS1	17	3
FFC	18	4
V3B	19	3
LO1	20	5
LO2	21	5
PIT	22	4
PCV	27	18
STV	28	15
7m	30	18
POS1	31	18
23d	32	18
v23ab	33	18
d23ab	34	18
31pv	35	18
LIPv	48	16
VIP	49	16
MIP	50	16
PH	138	5
TPOJ1	139	15
TPOJ2	140	15
TPOJ3	141	15
IP2	144	17
IP1	145	17
IP0	146	17
VMV1	153	4
VMV3	154	4
LO3	159	5
VMV2	160	4
VVC	163	4

Table 1: ROI name, group number, and index of the Glasser Atlas for the ROIs used in this work.

1.2 Qualitative Evaluation by Human Users

Qualitative evaluations were collected using the Prolific online experiment platform (www.prolific.com). Consent was collected and payment was awarded according to procedures approved by the institution's IRB, and participants were paid according to an hourly rate of 11.25/hr. Our qualitative evaluations measure whether the reconstructed video captured the semantic meaning of the original video. Participants were shown one real video as a reference, and were instructed to identify (by clicking on it) which of 6 reconstructed videos best matched the real video. One of the 6 videos was always a reconstruction of the reference video. To ensure they evaluated the semantic content of the videos (rather than the visual quality), participants were instructed to "attend to the objects, setting and context".

We calculated the percent of trials where the best matching video (as evaluated by human subjects) corresponded to the reconstruction of the reference video. Pilot results showed strong consensus among participants, so a sample size of 6 participants was collected on videos from each reconstruction method we explored.

A noise ceiling, representing the best possible performance given our model backbone (zeroscope V2) was calculated by feeding human-generated captions into zeroscope v2. Specifically, the dataset used here contained 5 different captions per video, collected from independent subjects. Reconstructed videos were generated for each caption, and to ensure that we were capturing variation due to different model seeds, we reconstructed each using 5 different seeds (for a total of 25 reconstructions per video)

2 BMD and CC2017 Statistics

We analyze low level spatial and temporal statistics between BMD and CC2017 that may impact the success of reconstructions. We show the results in Figure 1. We observe that broadly, BMD tends to contain videos with higher spatial frequency, hinting at additional movement and complexity. When analyzing the distribution of average TVL1 optical flow across videos, we see that around 20% of CC2017 videos are static (falling in the first bin when using a bin width of 0.05), while only 10% of BMD videos exhibit that property.

3 Within and Between Subject Correlations

We perform pairwise correlation of brain responses both within and between subjects to quantify the similarity of brain activity in response to videos. This procedure differs slightly for BMD and CC2017 due to the format of their brain responses (i.e., beta values in BMD and fMRI time series in CC2017).

In BMD, single trial brain responses corresponding to each presentation of a 3s video from the BOLD Moments Dataset were obtained by estimating a beta value at each cortical vertex using using a General Linear Model (GLM) (Figure

4A). In this way, 3000 beta values (1000 videos x 3 repetitions) and 1020 beta values (102 videos x 10 repetitions) were estimated across the training and testing sets, respectively, at each vertex. Within subject (intra-subject) correlations were computed by correlating (Pearson) the vector of 1000 training set betas between each pair of repetitions (i.e., three unique pairs from three repetitions) and averaging the resulting correlations over the pairs (Figure 3, left). Between subject (inter-subject) correlations were computed pairwise between subject N's vector of 1000 training set betas (averaged over repetitions) and each other subject's vector of 1000 training set betas. The average of these pairwise correlations is regarded as subject N's between subject correlation (Figure 3, right).

Brain responses corresponding to a 2s clip from the CC2017 dataset were obtained by sampling the fMRI time series at a 4 second offset (2 TRs) from the beginning of the 2s clip (Figure 5A). This procedure resulted in 4,302 unique training clips and corresponding brain responses with two repetitions. Within subject (intra-subject) correlations were computed as the correlation (Pearson) between the vectors of time series estimates from repetitions 1 and 2 at each vertex in the brain (Figure 2, left). Correlations were computed within each of the 18 training segments then averaged together. The between subject (inter-subject) correlation for subject N was computed by pairwise correlating subject N's repetition-averaged vector of time series estimates with the remaining two subjects and averaging (Figure 2, right). Correlations were first computed within each of the 18 training segments and averaged. Since the brain responses used here are regular samples of the fMRI time series, the between subject correlation is nearly identical to pairwise intersubject correlation (ISC) analyses [12] [19] [15].

Note that the correlation values should only be compared within their respective dataset, as correlating vectors of beta values (as in BMD) and fMRI time series (as in CC2017) is not equivalent. Within subject correlations are higher than between subject correlations in both BMD and CC2017 datasets, and both measures show a high degree of similarity throughout visual and parietal cortices.

4 Zero-shot reconstruction

We analyze the possiblity of reconstructing videos from a subject that the model *hasn't seen during training*. We show how MSE decreases for a representative subject (subject 01 in both BMD and CC2017) as more subjects are added into model training, indicative of better reconstruction accuracy (Figures 4 and 5 panel C). The MSE does not reach the same value as the the baseline case where the model was trained and tested on subject 01. The within and between subject correlations in Figures 4 and 5 panel B show that the fMRI activity itself is more correlated within than between subjects but still show similar patterns of high correlations in visual cortex and beyond.

5 Additional Reconstructions

We show additional reconstructions for BMD and CC2017 from our multi-subject Brainflix model in Figure 6. We observe strong semantic reconstruction performance in both datasets, with some clear limitations in terms of structure fidelity. The semantic fidelity is generally high, except in specific cases where a completely different semantic object is reconstructed (as shown in our main paper). In those cases, we hypothesize that the subject's mind wandered away from the task. To improve the quality of the reconstructions, we propose the following avenues for future work:

- Improve the reconstruction quality on the latent vectors (z), which hold most of the structural information;
- Train with more data, including image datasets, to incorporate more information about visual priors;
- Adopt a selection mechanism over reconstruction alternatives, to select instances with higher semantic match with the aligned fMRI embeddings.

6 Pseudocode for Reconstruction Procedure

```
Function train_mbm_encoder_decoder(fmri_data):
    encoder ← MBMEncoder()
    decoder ← MBMDecoder()
    masked_fmri_data ← mask_patches(fmri_data)
    latent_vectors ← encoder(masked_fmri_data)
    reconstructed_fmri ← decoder(latent_vectors)
    mse_loss ← calculate_mse_loss(reconstructed_fmri, fmri_data)
    optimize(encoder, decoder, mse_loss)
    return encoder
Function fine_tune_encoder_with_contrastive(encoder, fmri_data,
 clip_embeddings):
    latent_vectors ← encoder(fmri_data)
    contrastive_loss ← calculate_contrastive_loss(latent_vectors,
     clip_embeddings)
    optimize(encoder, contrastive_loss)
    return encoder
Function train_regressors(encoder_outputs):
    mlp_z ← MLP()
    mlp_b ← MLP()
    z ← mlp_z(encoder_outputs)
    b ← mlp_b(encoder_outputs)
    regression_loss_z ← calculate_mse_loss(z, true_latent_vector)
    regression_loss_b ← calculate_mse_loss(b, true_blip_embedding)
    optimize(mlp_z, mlp_b, regression_loss_z + regression_loss_b)
    return mlp_z, mlp_b
Function reconstruct_video(new_fmri, encoder, mlp_z, mlp_b):
    fmri_embedding ← encoder(new_fmri)
    latent_vector ← mlp_z(fmri_embedding)
    blip_embedding ← mlp_b(fmri_embedding)
    re_noised_z ← renoise(latent_vector)
    caption ← decode_blip(blip_embedding)
    video ← denoise_video(re_noised_z, caption)
    return video
Function main_pipeline(fmri_data, clip_embeddings, new_fmri):
    encoder ← train_mbm_encoder_decoder(fmri_data)
    encoder ← fine_tune_encoder_with_contrastive(encoder, fmri_data,
     clip_embeddings)
    encoder_outputs ← encoder(fmri_data)
    mlp_z, mlp_b ← train_regressors(encoder_outputs)
    video ← reconstruct_video(new_fmri, encoder, mlp_z, mlp_b)
    return video
```

Algorithm 1: Reconstruction Procedure

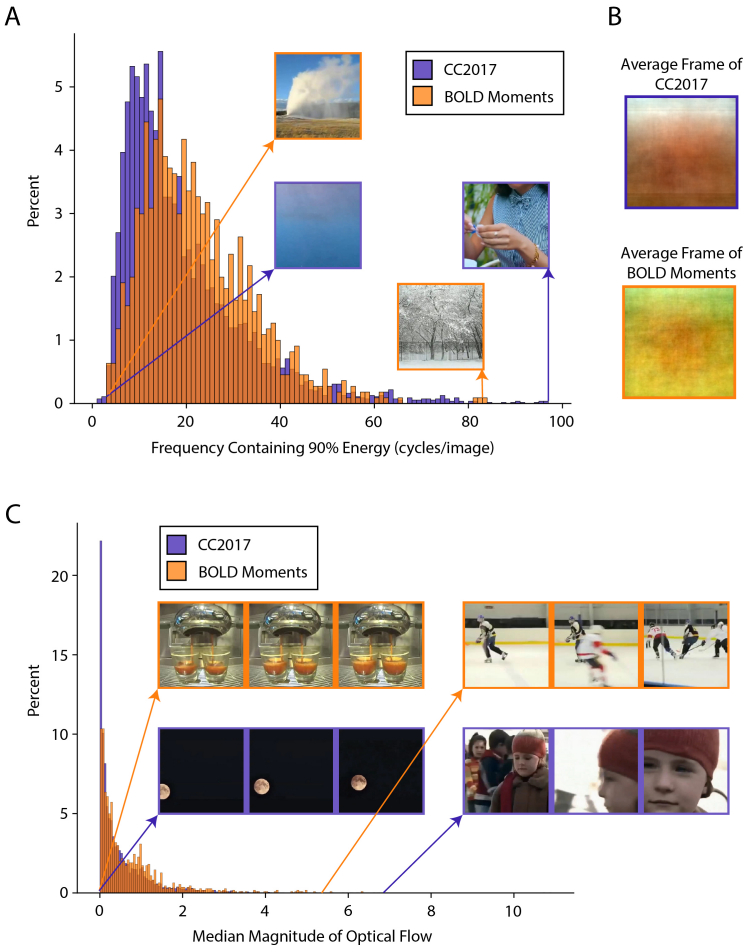

Fig. 1: We summarize low-level spatial and temporal statistics between the BOLD Moments (BMD) and CC2017 datasets. A) For both datasets, the middle frame of each clip was resized to 256 x 256 and the frequency that contains 90% of the energy was computed using the Natural Image Statistical Toolbox [2]. The histogram shows the percentage of clips (using the middle frame to approximate the spatial information in the clip) in each dataset at frequencies that capture 90% of the energy (bin width = 1). Higher frequencies correspond to high spatial frequency content of the middle frame of the clip, as shown in the frames with the checkered shirt (CC2017) and tree branches (BMD). BMD contains a higher percentage of video clips with higher spatial frequency frames compared to CC2017. B) The middle frame of each video clip was extracted, resized to 256 x 256, and averaged in each dataset separately using the Natural Image Statistical Toolbox [2]. The average of frames across each video clip in the BMD and CC2017 datasets highlights their differences in color and average low-level spatial content. C) The magnitude of the TVL1 optical flow was computed between each pair of consecutive frames (resized to 128 x 128) for each video clip in each dataset, separately. The median of these magnitudes across frames and pixels was computed to summarize the optical flow of each video clip and was plotted in a histogram (bin width = 0.05). Over 20% of CC2017 clips are highly static, compared to approximately 10% of clips in BMD.

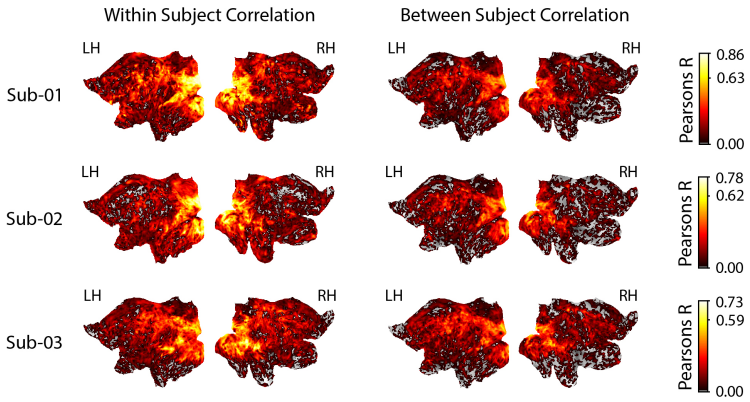

Fig. 2: The within subject (left) and between subject (right) correlations (Pearsons R) are shown on the left and right hemishperes of a flattened brain for all subjects in the CC2017 dataset. The maximum value of the colorbar corresponds to the maximum within subject correlation, and the middle colorbar value correponds to the maximum between subject correlation. All correlations are clipped at a threshold of 0.01.

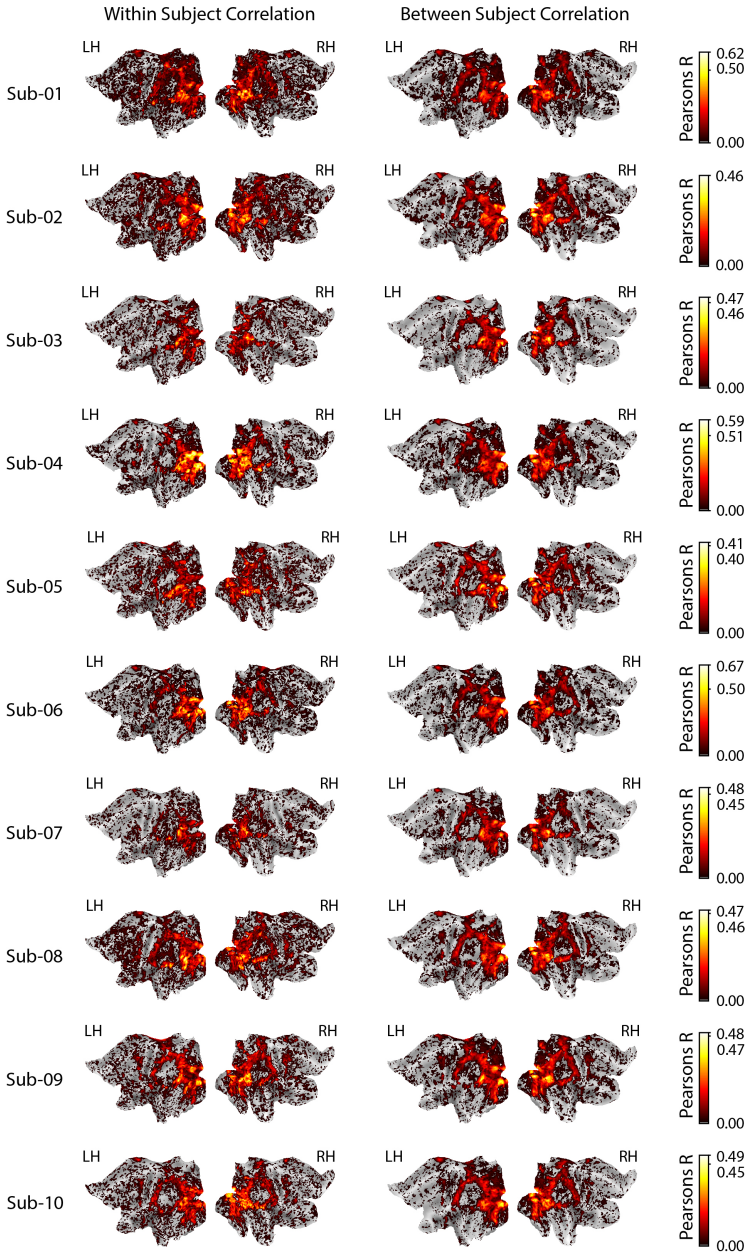

Fig. 3: The within subject (left) and between subject (right) correlations (Pearsons R) are shown on the left and right hemishperes of a flattened brain for all subjects in the BOLD Moments Dataset. The maximum value of the colorbar corresponds to the maximum within subject correlation, and the middle colorbar value correponds to the maximum between subject correlation. All correlations are clipped at a threshold of 0.01.

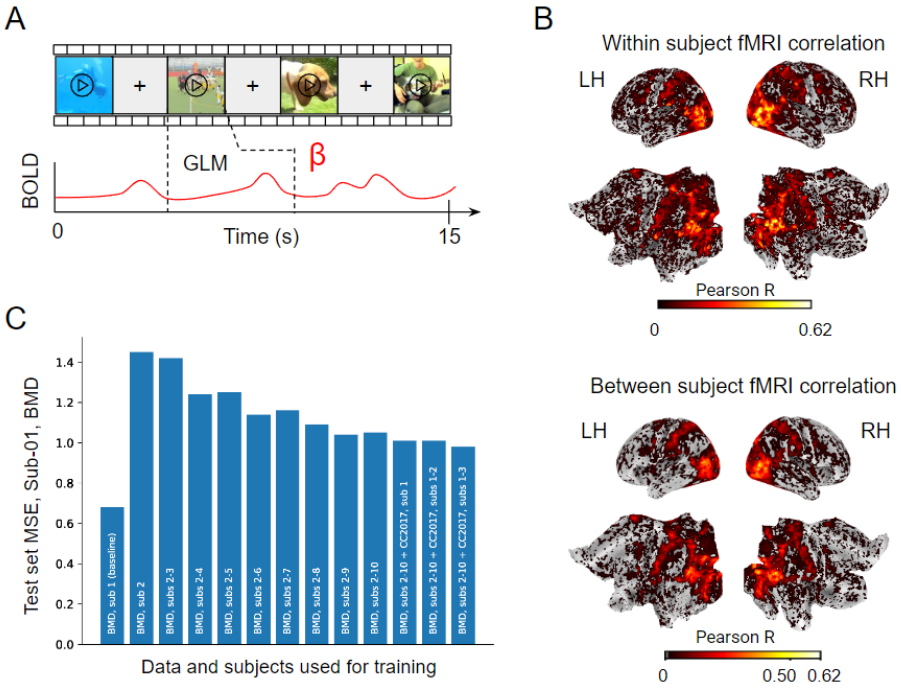

Fig. 4: A) In BMD, subjects viewed a series of discrete 3s videos, and brain responses (beta values) were estimated using a General Linear Model (GLM). B) Within subject pairwise correlations (top) between the three stimulus repetitions for representative subject 01 are plotted in an inflated and flattened brain. C) Between subject pairwise correlations (bottom) for representative subject 01 are plotted in an inflated brain and flattened brain. The colorbar is scaled to subject 01's maximum within subject correlation, and the middle colorbar tick reflects the maximum between-subject correlation. D) The test set MSE is plotted as subjects are increasingly included in model training.

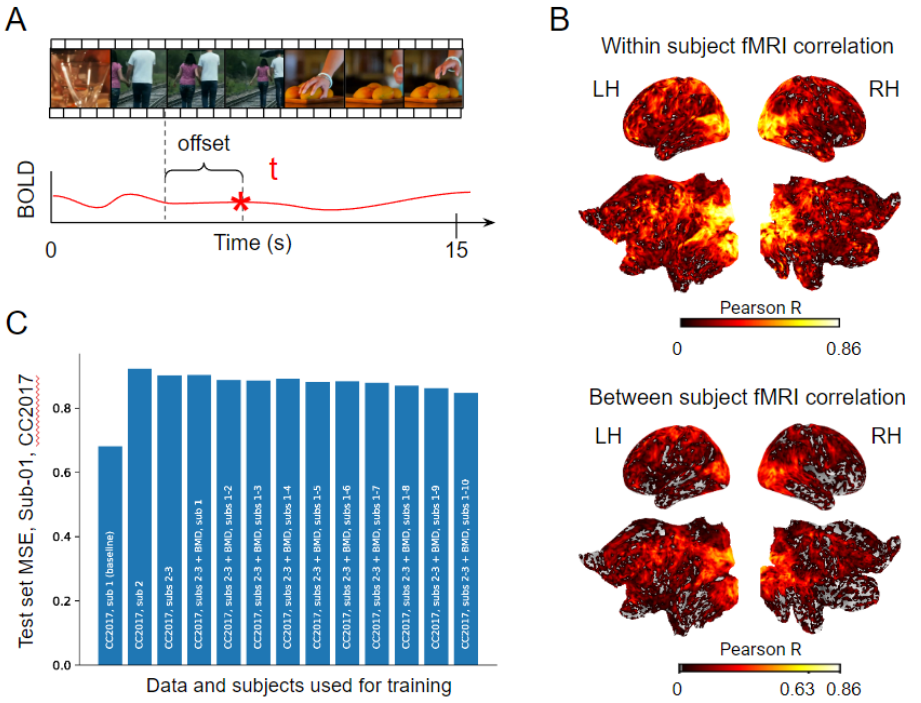

Fig. 5: A) In CC2017, subjects viewed a continuous longform movie composed of 10-15 seconds of professionally shot footage. Brain responses corresponding to a 2s clip was obtained by sampling the fMRI time series at a 4 second offset from the beginning of the 2s clip. B) Within subject pairwise correlations (top) between the two stimulus repetitions for representative subject 01 are plotted in an inflated and flattened brain. C) Between subject pairwise correlations (bottom) for representative subject 01 are plotted in an inflated and flattened brain. The colorbar is scaled to subject 01's maximum within subject correlation, and the middle colorbar tick reflects the maximum between-subject correlation. D) The test set MSE is plotted as subjects are increasingly included in model training.

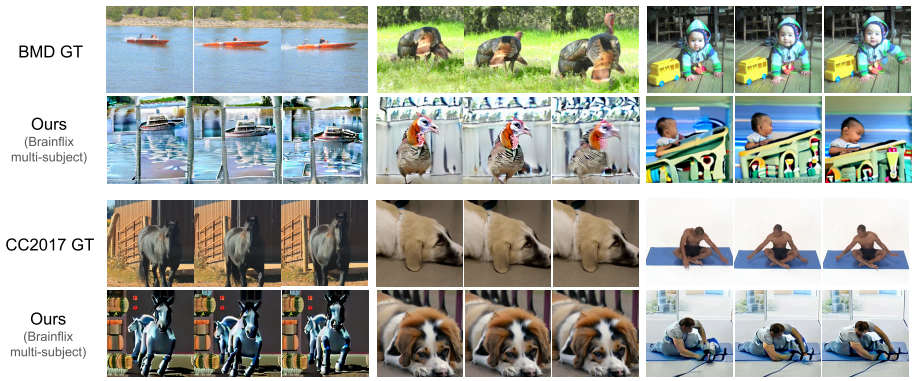

Fig. 6: Additional reconstructions from multi-subject Brainflix on BMD and CC2017. We observe reasonable semantic reconstructions, and some structural fidelity, although noise, as well and mismatched objects and positions are still perceivable.

References

1. 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)
2. Bainbridge, W.A., Oliva, A.: A toolbox and sample object perception data for equalization of natural images. Data in brief **5**, 846–851 (2015)
3. Chang, N., Pyles, J.A., Marcus, A., Gupta, A., Tarr, M.J., Aminoff, E.M.: Bold5000, a public fmri dataset while viewing 5000 visual images. Scientific data **6**(1), 49 (2019)
4. Dickie, E.W., Anticevic, A., Smith, D.E., Coalson, T.S., Manogaran, M., Calarco, N., Viviano, J.D., Glasser, M.F., Van Essen, D.C., Voineskos, A.N.: Ciftify: A framework for surface-based analysis of legacy mr acquisitions. Neuroimage **197**, 818–826 (2019)
5. Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al.: fmriprep: a robust preprocessing pipeline for functional mri. Nature methods **16**(1), 111–116 (2019)
6. Etzel, J.A., Zacks, J.M., Braver, T.S.: Searchlight analysis: promise, pitfalls, and potential. Neuroimage **78**, 261–269 (2013)
7. Felleman, D.J., Van Essen, D.C.: Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex (New York, NY: 1991) **1**(1), 1–47 (1991)
8. Gazzola, V., Keysers, C.: The observation and execution of actions share motor and somatosensory voxels in all tested subjects: single-subject analyses of unsmoothed fmri data. Cerebral cortex **19**(6), 1239–1255 (2009)
9. Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., et al.: A multi-modal parcellation of human cerebral cortex. Nature **536**(7615), 171–178 (2016)
10. Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., et al.: The minimal preprocessing pipelines for the human connectome project. Neuroimage **80**, 105–124 (2013)
11. Grill-Spector, K., Malach, R.: The human visual cortex. Annu. Rev. Neurosci. **27**(1), 649–677 (2004)
12. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R.: Intersubject synchronization of cortical activity during natural vision. science **303**(5664), 1634–1640 (2004)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
14. Hebart, M.N., Contier, O., Teichmann, L., Rockter, A.H., Zheng, C.Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., Baker, C.I.: Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. Elife **12**, e82580 (2023)
15. Kauppi, J.P., Jääskeläinen, I.P., Sams, M., Tohka, J.: Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Frontiers in neuroinformatics **4**, 669 (2010)
16. Lahner, B., Dwivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., Gifford, A.T., Pan, B., Jin, S., Murty, N.A.R., Kay, K., Oliva, A., Cichy, R.: Modeling short visual events through the bold moments video fmri dataset and

metadata. Nature Communications (Jul 2024), received: 14 August 2023; Accepted: 4 July 2024
17. Le, A., Vesia, M., Yan, X., Crawford, J.D., Niemeier, M.: Parietal area ba7 integrates motor programs for reaching, grasping, and bimanual coordination. Journal of Neurophysiology (2017)
18. Logothetis, N.K., Sheinberg, D.L.: Visual object recognition. Annual review of neuroscience **19**(1), 577–621 (1996)
19. Pajula, J., Kauppi, J.P., Tohka, J.: Inter-subject correlation in fmri: method validation against stimulus-model based analysis (2012)
20. Peeters, R., Simone, L., Nelissen, K., Fabbri-Destro, M., Vanduffel, W., Rizzolatti, G., Orban, G.A.: The representation of tool use in humans and monkeys: common and uniquely human features. Journal of Neuroscience **29**(37), 11523–11539 (2009)
21. Peeters, R.R., Rizzolatti, G., Orban, G.A.: Functional properties of the left parietal tool use region. Neuroimage **78**, 83–93 (2013)
22. Prince, J.S., Charest, I., Kurzawski, J.W., Pyles, J.A., Tarr, M.J., Kay, K.N.: Improving the accuracy of single-trial fmri response estimates using glmsingle. Elife **11**, e77599 (2022)
23. Rizzolatti, G., Sinigaglia, C.: The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nature reviews neuroscience **11**(4), 264–274 (2010)
24. Silver, M.A., Kastner, S.: Topographic maps in human frontal and parietal cortex. Trends in cognitive sciences **13**(11), 488–495 (2009)
25. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.M.H., et al.: The wu-minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
26. VanRullen, R., Thorpe, S.J.: The time course of visual processing: from early perception to decision-making. Journal of cognitive neuroscience **13**(4), 454–461 (2001)
27. Wang, C., Yan, H., Huang, W., Li, J., Wang, Y., Fan, Y.S., Sheng, W., Liu, T., Li, R., Chen, H.: Reconstructing rapid natural vision with fmri-conditional video generative adversarial network. Cerebral Cortex **32**(20), 4502–4511 (2022)
28. Wang, L., Mruczek, R.E., Arcaro, M.J., Kastner, S.: Probabilistic maps of visual topography in human cortex. Cerebral cortex **25**(10), 3911–3931 (2015)
29. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14668–14678 (2022)
30. 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)
31. Zhou, M., Gong, Z., Dai, Y., Wen, Y., Liu, Y., Zhen, Z.: A large-scale fmri dataset for human action recognition. Scientific Data **10**(1), 415 (2023)