A Systematic Comparison of fMRI-to-video Reconstruction Techniques

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

Abstract

Recent advances in generative models and large-scale neural datasets have brought forth novel methods to reconstruct stimuli from brain activity. This rapidly evolving family of brain-to-stimuli reconstruction techniques has the opportunity to revolutionize fundamental brain sciences and human-computer interaction applications, yet systemic comparisons of these techniques are lacking. Here, we explore a novel method to reconstruct short videos from functional magnetic resonance imaging (fMRI) brain activity of human subjects that achieves state-of-the-art performance as assessed by a suite of evaluation metrics. We perform preliminary comparisons of reconstruction quality within our pipeline by testing different combinations of semantic encoders and video generation models. Lastly, we compare our pipeline's best reconstruction results with previous work. Together, this work comprehensively assesses state-of-the-art methodologies in the increasingly important discipline of brain-to-video reconstruction.

1. Introduction

Humans observe the visual world through a spatiotemporal stream of input. From this input humans must extract fine details of object shape, identity, and motion to effectively interact with their environment. Given current neuroimaging techniques, we wonder how can this visual information be best recovered? The ability to reconstruct a viewer's visual experience from their brain activity would increase researchers' understanding of biological intelligence and invite medical advancements in human-computer interaction technology.

[*]Equal contribution [1]Computer Science and Artificial Intelligence Lab (CSAIL), Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Camilo Fosco <camilolu@mit.edu>, Benjamin Lahner <blahner@mit.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

However, this line of brain-to-stimulus reconstruction work is often limited by the ability to obtain high quality generative models and large-scale neural datasets. High quality generative vision models are typically trained on billions of images or videos requiring massive amounts of computational resources often out of reach of researchers. Large-scale neural datasets are expensive and time-consuming to collect. While videos are more ecologically valid than images, videos demand even more computational resources for model training and are less often used in large-scale neural data collection efforts. Fortunately, recent work in generative AI and cognitive neuroscience disciplines are making progress against these barriers with the availability of high quality text-to-video generative models and large-scale video fMRI datasets.

In this work, we learn robust neural representations using a masked-brain modeling (MBM) technique. MBM extends masked training methods popular in natural language processing and computer vision to brain data by challenging an encoder-decoder architecture to reconstruct masked portions of the input brain signal (Chen et al., 2023b)(Chen et al., 2023a). From this brain representation, we leverage a class of generative models termed latent diffusion models (LDMs) (Rombach et al., 2022) to generate high quality naturalistic videos. LDMs accomplish these high fidelity video (or image) generations through a denoising process over latent pixel representations conditioned on an input (e.g., text).

Functional magnetic resonance imaging (fMRI) data, as opposed to other forms of brain data, offers the advantages of being non-invasive and high spatial resolution. fMRI's non-invasive data collection process facilitates large-scale efforts and application of reconstruction techniques to a wider population. fMRI's high spatial resolution throughout cortex captures a wide variety of visual representations (Fischer et al., 2016)(Silver & Kastner, 2009)(Konen & Kastner, 2008)(Haxby, 2012). Despite fMRI's temporal sluggishness, these representations have been shown to capture features of dynamic stimuli (Le et al., 2017)(Gazzola & Keysers, 2009)(Rizzolatti & Sinigaglia, 2010)(Rust et al., 2006)(Hasson et al., 2008).

We propose a 3-stage pipeline that combine these recent

advancements in generative AI and neuroscience to generate high fidelity short videos from brain data, following the progress of (Fosco et al., 2024). First, we use MBM to train an encoder-decoder architecture to recover masked portions of the fMRI signal while maintaining alignment with a video's CLIP representation. Second, we regress a visual and semantic conditioning vector from the learned encoder, and finally, reconstruct the video stimulus with the LDM. We use resting state (no visual stimulus) fMRI data from the Human Connectome Project (HCP) (Van Essen et al., 2013) and video-fMRI pairs from the BOLD Moments Dataset (BMD) (Lahner et al., 2024), Human Actions Dataset (HAD) (Zhou et al., 2023), and CC2017 dataset (Wen et al., 2018) to train our models.

Our experiments compare the reconstruction quality of the LDM and target semantic regressors used in the second and third stages, then compare our best approach with previous work. Our contributions are as follows:

1. We propose a novel fMRI-to-video reconstruction pipeline that aggregates fMRI data across subjects and datasets.
2. We vary core components of our pipeline to document their effect on reconstruction quality.
3. We compare our best reconstruction pipeline to previous work to contextualize our findings within the research community.

This work pushes the frontier of brain-to-video reconstruction research by studying the effect different modeling components have on reconstruction quality.

2. Related Work

Diffusion models: Latent Diffusion models (LDMs) (Rombach et al., 2022) are a family of generative models that perform a denoising process on latent representations to generate high fidelity outputs. Generative outputs can be guided to incorporate desired features by conditioning the denoising process on other inputs, such as text. This process has exhibited impressive results in tasks as diverse as image generation, super-resolution, recoloring, and audio generation (Dhariwal & Nichol, 2021)(Saharia et al., 2022b)(Song et al., 2020)(Saharia et al., 2022c)(Saharia et al., 2022a)(Liu et al., 2023). Recently, LDMs have been used to generate short videos true to a text input (Blattmann et al., 2023b). In this work, we compare the brain-to-video reconstruction quality between different LDMs.

Masked Brain modeling: Masked-brain modeling (MBM) (Chen et al., 2023a) is self-supervised method to pre-train models on fMRI data based on principles of Masked Signal Modeling (MSM). MSM is a commonly used pre-training task for large language and vision models (Devlin et al., 2018; He et al., 2022; Wei et al., 2022; Xie et al., 2022) where, after masking a portion of the input, an autoencoder is tasked to recover the input in its entirety. Applied to MBM, masking some amounts of fMRI data allows the model to learn complex spatial relationships between vertices of the brain.

Video reconstruction from brain activity: Early brain-to-video reconstruction work used a voxelwise motion-energy encoding model (Adelson & Bergen, 1985) to model dynamic movie information in visual cortex and a Bayesian decoder to reconstruct the previously seen movie from a sampled natural movie prior (Nishimoto et al., 2011). In the space of deep learning, features from convolutional neural networks (CNNs) have showed promising video reconstructions that capture basic object shape and content (Wen et al., 2018)(Le et al., 2022)(Kupershmidt et al., 2022). Researchers have found that using dual spatial and temporal discriminators in generative adversarial networks (GANs) result in spatiotemporally accurate reconstructions (Wang et al., 2022), and variational autoencoders can leverage compressed latent spaces to reconstruct videos frame-by-frame (Han et al., 2019). In an approach most similar to ours, (Chen et al., 2023b) uses MBM to learn powerful latent representations of fMRI data that is then input into an LDM to reconstruct the previously seen video.

3. fMRI Datasets

We train and evaluate our pipeline on four large fMRI datasets diverse in subjects and stimulus (Van Essen et al., 2013)(Lahner et al., 2024)(Zhou et al., 2023)(Wen et al., 2018). Together, our pipeline uses over 1,000 hours of resting state data from 1,084 subjects and 28,100 short videos (over 123,000 fMRI brain response trials) from 43 subjects. To the best of our knowledge, the BOLD Moments Dataset (BMD) (Lahner et al., 2024) and Human Actions Dataset (HAD) (Zhou et al., 2023) have not been used in any prior reconstruction work. The CC2017 dataset (Wen et al., 2018) has been extensively used as a benchmark (Chen et al., 2023b) and thus serves as a good comparison for our proposed pipeline. Below we summarize each dataset's fMRI preprocessing and offer more details in Appendix section A "fMRI Preprocessing and Data Preparation."

To account for the anatomical differences between individual human brains, we first register each brain to a template cortical surface in fsLR32k space (if not already done) using the MSMSulc algorithm (Robinson et al., 2018)(Glasser et al., 2013). This algorithm accurately maps the gray matter voxels along the cortical ribbon to a shared surface mesh to establish a voxel-to-voxel correspondence between subjects. In this way, the model is able to learn spatial patterns across subjects and datasets.

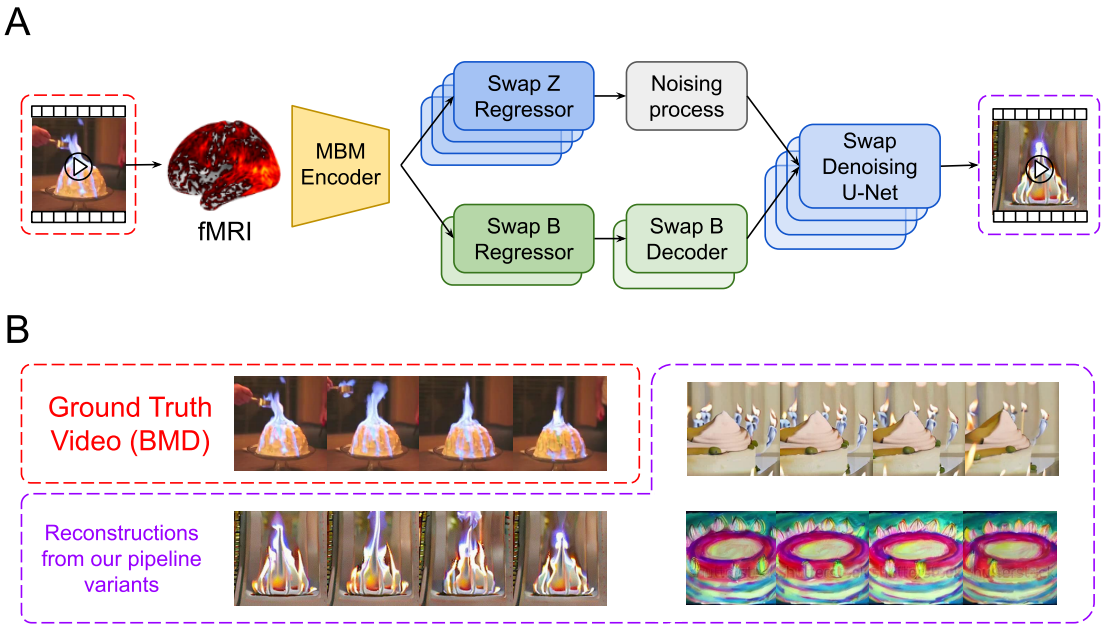

Figure 1. **Reconstruction of short videos from fMRI activity** A) Brain responses of subjects viewing video clips (red dotted outline) was measured with fMRI and then used to reconstruct the seen video (purple dotted outline). The video reconstruction quality was compared between pipelines using different diffusion U-Net models (blue), Z regressors (light blue), semantic B decoders (green), and B regressors (light green). B) We show examples of a ground truth video (red dotted outline) and our reconstructions from various pipeline combinations (purple outline).

We train and evaluate our pipeline on brain activity from 41 regions of interest (ROIs) defined in the Glasser Atlas (Glasser et al., 2016). These 41 ROIs (see 2) balance a broad sampling of cortex with computational efficiency of model training (see Appendix section B "Region of interest definition" for an analysis on this tradeoff). The ROI selection samples from visual and visual adjacent cortices and covers approximately 22% of the whole brain. Defining ROIs from the Glasser Atlas (Glasser et al., 2016) (as opposed to subject-specific functional definitions) further facilitates computational modeling by ensuring each subject's brain activity is derived from the same set of voxels of the same size.

Human Connectome Project Dataset (HCP). We use nearly an hour of resting state scans from 1,084 subjects from the 1200-subject release of the Human Connectome Project (HCP) (Van Essen et al., 2013). During the resting state scans, the subjects were instructed to fixate on a cross-hair and remain awake. They did not perform any task or view any other visual stimulus. Resting state fMRI captures natural fluctuations of brain activity over time, and temporal correlations have been used to understand cortical organization in the brain (Smith et al., 2013).

BOLD Moments Dataset (BMD). In the BOLD Moments Dataset (BMD) (Lahner et al., 2024), 10 subjects viewed multiple repetitions of 1,102 3-second videos in an event-related design. Each subject viewed a 1,000 video training set 3 times and a 102 video testing set 10 times for a total of 40,200 fMRI responses. For each trial, we use a general linear model (GLM) to estimate a beta value at each voxel. We maintain this train/test split in our pipeline's training and testing phases. The stimulus set was sampled from the Moments in Time dataset (Monfort et al., 2019) and depicts naturalistic, amateur-shot videos (e.g., home videos) that may or may not contain humans. Each video in BMD is also annotated with 5 text descriptions that were used in training our pipeline's semantic regressor.

Human Actions Dataset (HAD). In the Human Actions Dataset (HAD) (Zhou et al., 2023), 30 subjects viewed a single presentation of 720 2-second videos in an event-related design. No stimuli were repeated within or across subjects for a total of 21,600 individual fMRI responses. We use a GLM to estimate single-trial beta responses to each video. The stimuli were sampled from the Human Action Clips and Segments (HACS) dataset (Zhao et al., 2019) and depict naturalistic human-centered actions. The authors of HAD did not define a train/test split, and we only use HAD in our pipeline's training.

CC2017 Dataset (CC2017). The CC2017 dataset (Wen et al., 2018) consists of three subjects each viewing 18

eight minute training segments and 5 eight minute testing segments. In contrast to BMD and HAD's event-related design, the training and testing segments were presented longform and were composed of shorter video clips (10-15 second duration) concatenated together. The training segments were composed of 374 video clips and the testing segments were composed of 598 videos clips. The video clips depict mostly naturalistic content shot by professionals for cinematic movies or stock advertising footage. Each training and testing segment was repeated two and five times per subject.

In this work, we identify discrete fMRI-video pairs by splitting the longform movie segments into 2 second clips (CC2017's acquisition TR) and identify the corresponding fMRI response at an offset of 4 seconds (accounting for the BOLD responses lag). We maintain the originally proposed train/test splits for our pipeline's training and testing.

4. Reconstruction Pipeline

Approach Overview. Our general three stage reconstruction pipeline first transforms the fMRI signal into a compressed representation then regresses a predominantly visual (z) and semantic vector (b) for a latent and conditioning input, respectively, into a video generation model (Figure 2). In the first **alignment stage**, an encoder-decoder network is pre-trained using masked-brain modeling on resting state fMRI data and further finetuned on the task-based fMRI data. The encoder is additionally aligned to the video's CLIP representation to learn the video's semantic properties in addition to the fMRI signal's spatial structure. In the second **regression stage**, the output from the encoder in stage 1 is used to train a Z MLP regressor and a B MLP regressor from the video's ground truth z and b vectors. Finally, the third **reconstruction stage** takes as input only a raw fMRI signal and freezes all other models. It first uses the encoder trained in stage 1 to encode a compressed representation of the signal and the Z and B MLP regressors from stage 2 to output a latent vector and conditioning semantic vector for input into the denoising U-Net.

In our experiments we perform reconstructions using different video generation models and semantic encoders. Thus, stage 1 stays fixed but the Z and B MLP regressors in stage 2 are trained for each model.

Stage 1: Alignment. The goal of this stage is to train an encoder that learns robust, compressed representations of the raw fMRI signal. We begin by pretraining an encoder-decoder architecture using masked-brain modeling (MBM). The fMRI input is divided into patches and some patches are masked. The encoder processes the masked input into a 1024-dimensional latent vector, and the decoder reconstructs the 1024-dimensional latent vector back into the original unmasked fMRI input. This framework is supervised by a simple reconstruction loss (MSE) over the masked patches. This approach is similar to the masked image modeling task (He et al., 2022) and learns the fMRI signal's spatial structure.

The encoder-decoder architecture is first pretrained on large-scale resting state fMRI data from the Human Connectome Project (HCP) (Van Essen et al., 2013) then finetuned on task-based data from the Human Actions Dataset (HAD) (Zhou et al., 2023) and BOLD Moments Dataset (BMD) (Lahner et al., 2024). Finally, we add an additional contrastive learning loss between CLIP embeddings of BMD and CC2017 (Wen et al., 2018) stimuli and their corresponding (encoded) fMRI signal to learn semantic information. Following (Radford et al., 2021), this contrastive loss promotes high cosine similarity between positive pairs of fMRI embeddings and its associated CLIP embedding while discouraging cosine similarity between negative pairs of the same fMRI embedding other CLIP embeddings:

$$\mathcal{L}_{contrastive} = -\sum_{i=1}^{N} log\Big(\frac{\exp(\frac{f_i * c_i}{\tau})}{\sum_{j=1}^{N}\exp(\frac{f_j * c_j}{\tau})}\Big)$$

where f_i is the fMRI embedding (output from the MBM encoder), c_i is the CLIP-text embedding, and τ is a temperature hyperparameter. We follow (Chen et al., 2023a) and (Scotti et al., 2024) for training parameters of the MBM model and alignment, respectively. Specifically, we use a 24-layer MBM encoder with patch size of 16 and hidden dimension of 1024. The temperature was set to 0.9 and masking ratio to 75%.

Stage 2: Regression. We freeze the encoder trained in stage 1 and use its 1024-dimensional output vector to train two multi-target MLP regressors with regularization: one for the z vector and the other for the b vector. The ground truth z and b targets are generated from the stimuli and models directly. The MLPs are composed of a linear layer, 3 residual blocks, and an output layer. The linear layer and 3 reisdual blocks each contain 2048 units. We additionally use dropout regularization ($p = 0.3$), GELU (Hendrycks & Gimpel, 2016) activations, and Batch Normalization. The MLPs are trained with an MSE loss and predict flattened output vectors. The output is then reshaped for input into the denoising U-Net in stage 3. Note that near-perfect regression at this stage would result in near-perfect reconstruction in stage 3, since we would be obtaining the exact latent and conditioning vectors.

Stage 3: Reconstruction. In this final stage, we reconstruct a video previously seen by the human participant purely from their fMRI signal. We first transform the raw fMRI signal into a 1024-dimensional vector using the encoder in stage 1. From this vector, we regress the latent z vector and conditioning b vector using the MLP regressors trained

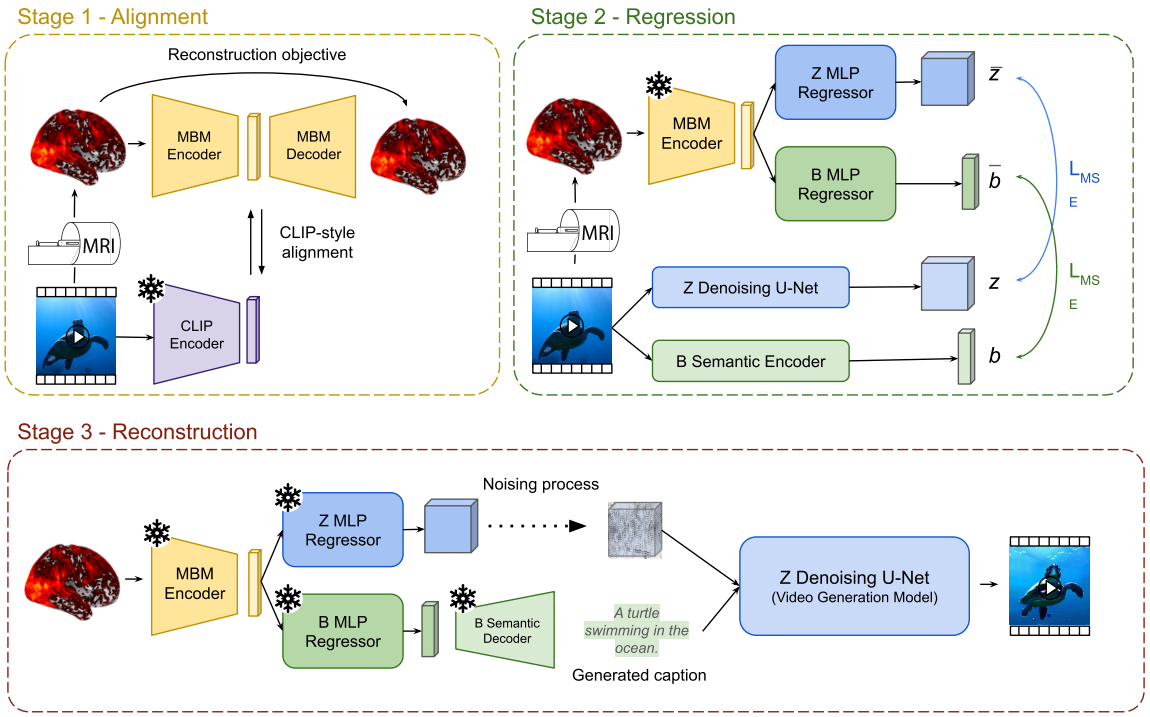

Figure 2. Our proposed reconstruction pipeline. **Stage 1:** We use masked-brain modeling to pretrain an encoder-decoder architecture to reconstruct fMRI responses. The encoder is then aligned with a video's CLIP embeddings. **Stage 2:** The video is used to compute ground truth z and b vectors. These ground truth vectors are used to train a z and b regressor (MLP) from the video's corresponding fMRI signal latent representation from the output of stage 1's MBM encoder. **Stage 3:** Stage 1's frozen MBM encoder is used to generate a latent representation of the fMRI signal. B and z vectors are regressed from the trained regressors in stage 2. A noised z vector and caption generated from the b vector are input into a video generation model to generate the final video.

in stage 2. The z vector undergoes a re-noising process (strength of 0.8, 40 steps following (Takagi & Nishimoto, 2023)) before being denoised in pretrained U-Net. The b vector is decoded into a text caption with repetition penalty of 6, minimum length of 4, and maximum length of 20 to encourage more descriptive captions. The decoded caption is then input into a pretrained (frozen) denoising U-Net as a conditioning vector along with the noised z vector. In this way, the resulting generated video captures visual and semantic meaning derived from the original fMRI signal.

5. Experiments and Results

5.1. Implementation Details.

We train and finetune our models with the fMRI datasets described in Section 3. All videos were downsampled to 15 FPS and resized to 224×224. In the masked-brain modeling reconstruction task in stage 1, we pretrain for 200 epochs with a batch size of 300. We finetune with CLIP-style alignment with BMD and CC2017 for 50 epochs and a batch size 120. For BMD video captions, we use the human annotations available with the dataset supplied by the authors. For CC2017 and HAD, captions were synthetically generated with the EILEV (Yu et al., 2023) video-to-caption model. Stage 1 training utilized 6 V100 GPUs and stage 2 was done on 2 Titan RTX GPUs.

5.2. Comparisons to previous work

We compare our pipeline's best reconstructions to previous fMRI-to-video techniques. We report results over BMD and CC2017 in Table 1. Our best results are achieved with a Zeroscope v2 reconstructor, which improves over two previous works (Kupershmidt (Kupershmidt et al., 2022) and Mind-Video (Chen et al., 2023b)) quite comfortably. We showcase visual comparisons in Figure 3, where we observe that our technique improves in visual fidelity, structure and video quality.

5.3. Comparing Different Generative Models.

With the pipeline described above, we seek to understand how different text-to-video generative models handle the fMRI conditioning. Towards that objective, we swap different video generation models in our Stage 3 and qualitatively compare the resulting generations.

Figure 3. We compare to previous works on CC2017 examples. The ground truth videos are shown in the first row, ours in the second (green), and previous approaches below. Our proposed pipeline captures structural similarity better than previous approaches.

Methods	CC2017				BMD			
	SSIM	MSE	2-way	50-way	SSIM	MSE	2-way	50-way
Kupershmidt	0.128	-	-	-	0.031	4.561	0.514	0.004
Mind-Video	0.186	-	0.853	0.202	0.176	0.763	0.711	0.101
Three-stage pipeline, Modelscope (Ours)	0.133	0.881	-	-	0.119	0.885	-	-
Three-stage pipeline, Hotshot-XL (Ours)	0.141	0.701	-	-	0.151	0.790	-	-
Three-stage pipeline, SVD (Ours)	0.140	0.754	-	-	0.133	0.711	-	-
Three-stage pipeline, Zeroscope v2 (Ours)	0.195	0.655	0.888	0.221	0.190	0.671	0.816	0.165

Table 1. We show a quantitative comparison of our reconstruction methodology against previous works. We achieve state-of-the-art results on most metrics. Results from Kupershmidt and Mind-video on BMD are obtained through a reimplementation, as their code is not readily available.

We use the following models in our analysis:

1. Modelscope (Wang et al., 2023): a text-to-video synthesis model constructed from a text-to-image diffusion model. It utilizes spatio-temporal blocks to maintain temporal consistency.
2. Zeroscope v2 (Hysts, 2024): an evolution of Modelscope, this model is also based on a diffusion architecture and trained to output high-quality 16:9 videos. We do not utilize the accompanying high-resolution secondary model to ensure fair comparisons.
3. Stable Video Diffusion (Blattmann et al., 2023a): image-to-video diffusion model trained on a curated set of high-resolution images and videos. Trained with three distinct regimes. Harbors strong video priors.
4. Hotshot-XL (Mullan et al., 2023): a text-to-video model derived from Stable Diffusion XL (Podell et al., 2023) optimized for generating gifs. Makes shorter videos with smaller aspect ratios.

For all the models above, we extract latent vectors and conditioning vectors from their respective variational encoders and text/image encoders. For SVD, wethe conditioning target is an image embedding instead of a text embedding, so we skip the caption generation step for that model.

We showcase results in Figure 4. Our qualitative comparison allows us to make several observations. First, video quality is tied to model complexity and pretraining dataset size. Models trained over larger datasets such as Stable Video Diffusion appear to showcase a more in-depth prior over videos. Second, adherence to the semantic concepts of the observed video is varied accross models. This performance is tied to our pipeline's ability to correctly regress the latent and conditioning vector from fMRI: the latents from some models appear to be an easier target for our regressor than others. Third, some models exhibit frequent failure modes (eg. Hotshot-XL, first example) where both the video quality, and the semantic adherence are incorrect. In those cases, we hypothesize that the regressed latents and

conditioning vectors are too imperfect to generate a meaningful video. This hints at a variation in the difficulty of the regression problem for each generative model: some latents and conditioning vectors (regression targets) appear to be more aligned with the fMRI signals, thus yielding a better regressor. Interestingly, SVD showcases high quality video but a disconnect in terms of semantic: as the conditioning regression target is an image embedding instead of a BLIP embedding, differences in regression difficulty might be the cause for the failure mode in Figure 4.

Evaluation Metrics. Following previous work (Takagi & Nishimoto, 2023; Chen et al., 2023b), we utilize 3 main evaluation metrics that aim to understand different characteristics of performance. To measure pixel-level reconstruction quality, we utilize Structural Similarity (SSIM). For semantic evaluation, we use the N-way top-k classification approach from (Chen et al., 2023b), which measures how often the classification of a simple ImageNet classifier over the reconstruction and ground truth video match, limiting the output of the classifier to N classes. We declare a successful trial if the ground truth class is within the top-k probabilities outputted over the reconstruction, and repeat the test 100 times to report average success rates. We also report MSE results over our target latent embeddings z and semantic embeddings b , to observe how close our regressed embeddings are to ground truth.

Evaluation Datasets We evaluate on two datasets: CC2017 (Wen et al., 2018) and BMD (Lahner et al., 2024) to compare our pipeline against previous work (Chen et al., 2023b)(Kupershmidt et al., 2022)(Wang et al., 2022) and against an additional dataset, BMD, with different subjects and video stimulus type (short video).

Results. We showcase our quantitative results over BMD and CC2017 in Table 1. Results are reported for Subject 1 in both datasets. We observe top results on Zeroscope v2, which exhibited the best semantic consistency across qualitative comparisons. We observe that this model is able to reconstruct examples from BMD and CC2017 with strong structural reliability. We hypothesize that our pipeline's emphasis on regressing an accurate latent, a component that previous approaches lack, enforces accurate structural patterns that can then lead to reliable object positioning across video generation models.

6. Conclusion

We describe a generative pipeline to reconstruct videos from human brain activity and assess the impact of exchanging different pipeline components. Using large text-to-video diffusion models and resting state and video fMRI datasets, we reconstruct the video the viewer previously saw with state-of-the-art quality. Our comparisons show that different generative models perform similarly at a qualitative level, but show clear differences when measuring results quantitatively. The best performing generation model was Zeroscope V2, and we compare our pipeline with this model to previous work, showcasing state of the art results.

Impact Statement

The techniques presented in this paper demonstrate innovative methods for reconstructing visuals from brain signals. Future advancements in these methods could have significant positive impacts. For individuals with communication disabilities, this technology could enable them to express their thoughts directly through images, videos, or words, greatly enhancing their ability to communicate. Additionally, this research contributes to the scientific understanding of the information recovery limits of various brain regions and reveals what can be decoded and simulated using generative models. Continued improvements in this reconstruction process could lead to extensive clinical and therapeutic applications, as well as further advancing the field of human-machine communication.

Acknowledgements

This research was funded by the Multidisciplinary University Research Initiative (MURI) award by the Army Research Office (grant No. W911NF-23-1-0277) to A.O; the MIT EECS MathWorks Fellowship to B.L.; the MIT EECS MathWorks Fellowship to C.F.

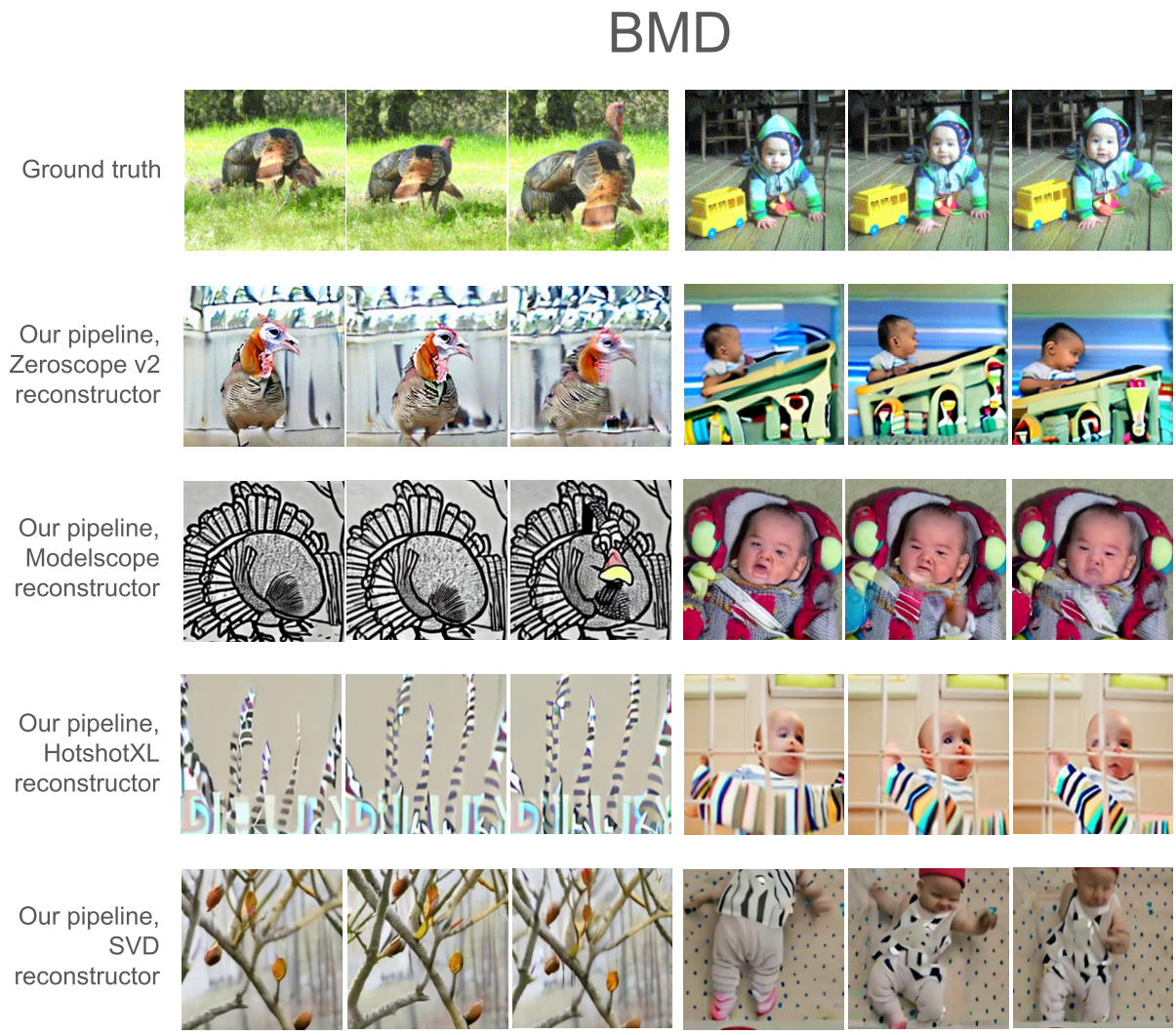

Figure 4. Reconstructions from different video generation models over BMD. We compare 4 different models: Zeroscope v2, Modelscope, Stable Video Diffusion and Hotshot-XL. We observe that reconstruction quality and semantic similarity is varied. Models with weaker image priors tend to generate distorted imagery with artifacts (e.g. modelscope), while models with higher complexity tend to generate coherent video, but don't reconstruct the semantic concept adequately (e.g. SVD).

References

Adelson, E. H. and Bergen, J. R. Spatiotemporal energy models for the perception of motion. *Josa a*, 2(2):284–299, 1985.

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.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, June 2023b.

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., and Aminoff, E. M. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1): 49, 2019.

Chen, Z., Qing, J., Xiang, T., Yue, W. L., and 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, 2023a.

Chen, Z., Qing, J., and Zhou, J. H. Cinematic mindscapes: High-quality video reconstruction from brain activity. *arXiv preprint arXiv:2305.11675*, 2023b.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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., and Voineskos, A. N. Ciftify: A framework for surface-based analysis of legacy mr acquisitions. *Neuroimage*, 197:818–826, 2019.

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.

Etzel, J. A., Zacks, J. M., and Braver, T. S. Searchlight analysis: promise, pitfalls, and potential. *Neuroimage*, 78:261–269, 2013.

Felleman, D. J. and Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., and Kanwisher, N. Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, 113(34):E5072–E5081, 2016.

Fosco, C., Lahner, B., Pan, B., Andonian, A., Josephs, E., Lascelles, A., and Oliva, A. Brain netflix: Scaling data to reconstruct videos from brain signals. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Gazzola, V. and 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.

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.

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.

Grill-Spector, K. and Malach, R. The human visual cortex. *Annu. Rev. Neurosci.*, 27:649–677, 2004.

Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., Fu, D., and Liu, Z. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550, 2008.

Haxby, J. V. Multivariate pattern analysis of fmri: the early beginnings. *Neuroimage*, 62(2):852–855, 2012.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and 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.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., and 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.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Hysts. Zeroscope v2. https://huggingface.co/spaces/hysts/zeroscope-v2, 2024. Accessed: 2024-06-05.

Konen, C. S. and Kastner, S. Representation of eye movements and stimulus motion in topographically organized areas of human posterior parietal cortex. *Journal of Neuroscience*, 28(33):8361–8375, 2008.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12 (5):535–540, 2009.

Kupershmidt, G., Beliy, R., Gaziv, G., and Irani, M. A penny for your (visual) thoughts: Self-supervised reconstruction of natural movies from brain activity. *arXiv preprint arXiv:2206.03544*, 2022.

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., and Cichy, R. Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature Communications*, July 2024. Received: 14 August 2023; Accepted: 4 July 2024.

Le, A., Vesia, M., Yan, X., Crawford, J. D., and Niemeier, M. Parietal area ba7 integrates motor programs for reaching, grasping, and bimanual coordination. *Journal of Neurophysiology*, 2017.

Le, L., Ambrogioni, L., Seeliger, K., Güçlütürk, Y., van Gerven, M., and Güçlü, U. Brain2pix: Fully convolutional naturalistic video frame reconstruction from brain activity. *Frontiers in Neuroscience*, 16:940972, 2022.

Liu, Z., Guo, Y., and Yu, K. Diffvoice: Text-to-speech with latent diffusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Logothetis, N. K. and Sheinberg, D. L. Visual object recognition. *Annual review of neuroscience*, 19(1):577–621, 1996.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

Mullan, J., Crawbuck, D., and Sastry, A. Hotshot-XL, October 2023. URL https://github.com/hotshotco/hotshot-xl.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

Peeters, R., Simone, L., Nelissen, K., Fabbri-Destro, M., Vanduffel, W., Rizzolatti, G., and 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.

Peeters, R. R., Rizzolatti, G., and Orban, G. A. Functional properties of the left parietal tool use region. *Neuroimage*, 78:83–93, 2013.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., and Kay, K. N. Improving the accuracy of single-trial fmri response estimates using glmsingle. *Elife*, 11: e77599, 2022.

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.

Rizzolatti, G. and Sinigaglia, C. The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature reviews neuroscience*, 11(4):264–274, 2010.

Robinson, E. C., Garcia, K., Glasser, M. F., Chen, Z., Coalson, T. S., Makropoulos, A., Bozek, J., Wright, R., Schuh, A., Webster, M., et al. Multimodal surface matching with higher-order smoothness constraints. *Neuroimage*, 167: 453–465, 2018.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Rust, N. C., Mante, V., Simoncelli, E. P., and Movshon, J. A. How mt cells analyze the motion of visual patterns. *Nature neuroscience*, 9(11):1421–1431, 2006.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022a.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022b.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022c.

Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K., et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36, 2024.

Silver, M. A. and Kastner, S. Topographic maps in human frontal and parietal cortex. *Trends in cognitive sciences*, 13(11):488–495, 2009.

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.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Takagi, Y. and 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.

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.

VanRullen, R. and Thorpe, S. J. The time course of visual processing: from early perception to decision-making. *Journal of cognitive neuroscience*, 13(4):454–461, 2001.

Wang, C., Yan, H., Huang, W., Li, J., Wang, Y., Fan, Y.-S., Sheng, W., Liu, T., Li, R., and Chen, H. Reconstructing rapid natural vision with fmri-conditional video generative adversarial network. *Cerebral Cortex*, 32(20): 4502–4511, 2022.

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., and Zhang, S. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Wang, L., Mruczek, R. E., Arcaro, M. J., and Kastner, S. Probabilistic maps of visual topography in human cortex. *Cerebral cortex*, 25(10):3911–3931, 2015.

Wei, C., Fan, H., Xie, S., Wu, C.-Y., Yuille, A., and 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.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12):4136–4160, 2018.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Yu, K. P., Zhang, Z., Hu, F., and Chai, J. Efficient in-context learning in vision-language models for egocentric videos. *arXiv preprint arXiv:2311.17041*, 2023.

Zhao, H., Torralba, A., Torresani, L., and Yan, Z. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8668–8678, 2019.

Zhou, M., Gong, Z., Dai, Y., Wen, Y., Liu, Y., and Zhen, Z. A large-scale fmri dataset for human action recognition. *Scientific Data*, 10(1):415, 2023.

A. fMRI Preprocessing and Data Preparation

Our brain-to-video reconstruction pipeline relies on large amounts of spatially aligned fMRI-video pairs so the models can learn robust spatial patterns in the brain data. To achieve this scale, we register all brain responses across four large datasets (the Human Connectome Project (HCP) dataset (Van Essen et al., 2013), BOLD Moments Dataset (BMD) (Lahner et al., 2024), Human Actions Dataset (HAD) (Zhou et al., 2023), and CC2017 dataset (Wen et al., 2018)) to a common space (fsLR32k) and form discrete fMRI-video pairs for the task-based data.

We use the versions of the HCP and CC2017 datasets that were previously aligned to the fsLR32k space with the HCP preprocessing pipeline (Glasser et al., 2013). For BMD and HAD, we register the data into fsLR32k space using Ciftify (Dickie et al., 2019) after preprocessing the data with fMRIPrep (Esteban et al., 2019). We achieve fMRI-video pairs in BMD and HAD by using a general linear model (GLM) to estimate single-trial beta responses to each 3 and 2 second video stimulus, respectively. For CC2017 we trim the longform video stimulus into 2s chunks (the fMRI acquisition rate of CC2017) and pair each chunk with the fMRI activity 4 seconds after the chunk's onset to account for the hemodynamic lag, as done in (Wang et al., 2022). Note that fMRI-video pairs were not made for HCP's resting state data because resting state brain activity has no associated stimulus. For all datasets, we use brain activity from 41 ROIs pre-defined in the Glasser Atlas (Glasser et al., 2016) (see 2 for details). We describe the preprocessing steps that we performed for each dataset below in detail. Please refer to each dataset's original manuscript for detailed acquisition protocols.

A.1. HCP Preprocessing

We train the masked-brain model with the resting state fMRI data from 1,084 subjects available in HCP's 1200-subject release. Specifically, we use the "rfMRI_RESTX_LR_Atlas_MSMAll_hp2000_clean.dtseries.nii" files for each subject, where the run number X is 1, 2, 3, or 4. At each voxel, we normalize the data across time and average the responses over a 10 second window prior to model input.

A.2. BOLD Moments Dataset Preprocessing and Preparation

The authors of BMD (Lahner et al., 2024) gave us permission to use their data for this work. The data was first preprocessed using fMRIPrep (Esteban et al., 2019), then registered to fsLR32k space using Ciftify (Dickie et al., 2019). The fMRI activity was first temporally interpolated from the acquisition TR of 1.75s to 1s to timelock the fMRI timeseries to the stimulus onset (1.75 does not evenly divide into the trial's 4s duration). The stimulus duration (modeled as a 0s impulse response), stimulus onsets, and interpolated fMRI timeseries was input into a GLM (GLMsingle (Prince et al., 2022)) to estimate single-trial beta value at each vertex for each session separately. The beta values within each session were then z-scored across conditions for the train and test conditions separately. In this way, 40,200 fMRI-video pairs were obtained across BMD's ten subjects.

A.3. Human Actions Dataset Preprocessing and Preparation

The HAD fMRI data (Zhou et al., 2023) was preprocessed using fMRIPrep (Esteban et al., 2019) and registered to fsLR32k space using Ciftify (Dickie et al., 2019) by the authors. Similar to BMD preprocessing, GLMsingle (Prince et al., 2022) was used to estimate single-trial beta values. However, since no stimuli were repeated, we used GLMsingle's type-B estimates. The responses were modeled with a 0s impulse stimulus duration. For each subject separately, the beta estimates were z-scored across stimuli conditions. This process resulted in a total of 21,600 fMRI-video pairs across all thirty subjects.

A.4. CC2017 Dataset Preprocessing and Preparation

In this work, we use a publicly available version of CC2017's (Wen et al., 2018) fMRI data preprocessed and registered to fsLR32k space with the HCP preprocessing pipeline (Glasser et al., 2013). Each of the 8 minute test and train segments were divided into 2 second non-overlapping but continuous chunks to create 5,497 short video clips (after accommodating scanner errors described in their manuscript). We obtain fMRI-video pairs by pairing each of the short video clips with the fMRI response that occured 4s after the clip's onset to account for the hemodynamic lag. This offset corresponds to approximately the BOLD signal's peak evoked by each clip. The values for the train and test stimuli were separately z-scored across conditions. We note these time series estimates used in CC2017 are fundamentally different measures of brain activity than the beta estimates used in BMD and HAD, but they capture similar spatial patterns across the brain and are z-scored to be within the same range.

B. Regions of interest definition

We define 41 regions of interest (ROIs) from the Glasser atlas (Glasser et al., 2016) to mask the fMRI data for stimulus reconstruction. The Glasser atlas (Glasser et al., 2016) uses structural and functional neural information from 210 healthy adults to divide the whole brain into 180 non-overlapping ROIs, which further compose 22 super groupings. Our unbiased selection (Kriegeskorte et al., 2009) of 41 ROIs effectively balances a broad sampling of the whole brain with computational efficiency. The 41 ROIs sample from 9 of the 22 super groups (see the 'Group Number' column in 2) while focusing on regions that have previously shown to respond to dynamic stimuli (VanRullen & Thorpe, 2001) (Logothetis & Sheinberg, 1996) (Le et al., 2017) (Gazzola & Keysers, 2009) (Rizzolatti & Sinigaglia, 2010) (Silver & Kastner, 2009) (Peeters et al., 2009) (Peeters et al., 2013) (Wang et al., 2015). These 41 ROIs recognize the brain's complex interconnected networks but also recognize that most networks that contribute to visual perception reside in visual and visual-adjacent cortices (Etzel et al., 2013)(Felleman & Van Essen, 1991)(Grill-Spector & Malach, 2004). This concentration of perceptually relevant vertices in and around the visual cortex can be seen in the within-subject correlations presented in Appendix Figure 5 and is replicated across many fMRI datasets (Allen et al., 2022)(Hebart et al., 2023)(Chang et al., 2019)(Zhou et al., 2023). To this end, a whole brain ROI selection would have introduced primarily noisy vertices in largely inconsequential regions at a large computational expense.

We demonstrate this accuracy to computational resource tradeoff by measuring MSE from the regressions of four different ROI groupings: our 41 ROI selection (13,156 vertices) $\rightarrow$ 0.721, core vision (6,549 vertices from super-groups 1-4) $\rightarrow$ 0.755, the average of 10 randomly selected sets of 41 ROIs (average of 12,390 vertices) $\rightarrow$ 0.983, and the whole brain (59,412 vertices) $\rightarrow$ 0.717. The whole brain's slight improvement over Group41 comes at large computational cost (5x to approximately 20x depending on the layer). Our ROI selection samples informative ROIs beyond the core visual regions and performs significantly better than a random ROI selection of similar size. We list the 41 ROI names, ID, and Group Number in Table 2.

C. Within Subject fMRI Correlations

We compute the similarity of brain responses within each subject in the CC2017 (Figure 5A) and BMD (Figure 5B) datasets to provide intuition about the quality and response pattern of our model's inputs. These similarities highlight highly reliable responses in the visual and visual adjacent cortical regions and are similar in magnitude and pattern to other fMRI datasets (Allen et al., 2022)(Hebart et al., 2023)(Chang et al., 2019).

In CC2017, we first extract each subject's brain responses to each of the trimmed 2s clips (as explained above in Appendix section A.4). We then correlate (Pearson's R) the vector of brain responses corresponding to the first and second repetitions of each of the 18 training set segments. The correlations were then averaged over the segments and visualized on a flattened brain (Figure 5A). Note that this procedure of correlating the vector of brain responses is identical to correlating the fMRI timeseries itself but trimmed at the beginning and end using the 4s offset of the first and last 2s segment.

In BMD, we estimate single trial beta values to each video presentation for each subject (as explained above in Appendix section A.2). Since each training video was repeated three times, we correlate (Pearson's R) the vector of beta estimates corresponding to the 1,000 training videos between all three unique repetition pairs. The average of the three correlations are visualized on a flattened brain (Figure 5B).

The correlation values should only be compared within datasets, not across datasets, because the fMRI timeseries values used in CC2017 are a fundamentally different measure of brain activity than the beta estimates used in BMD.

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

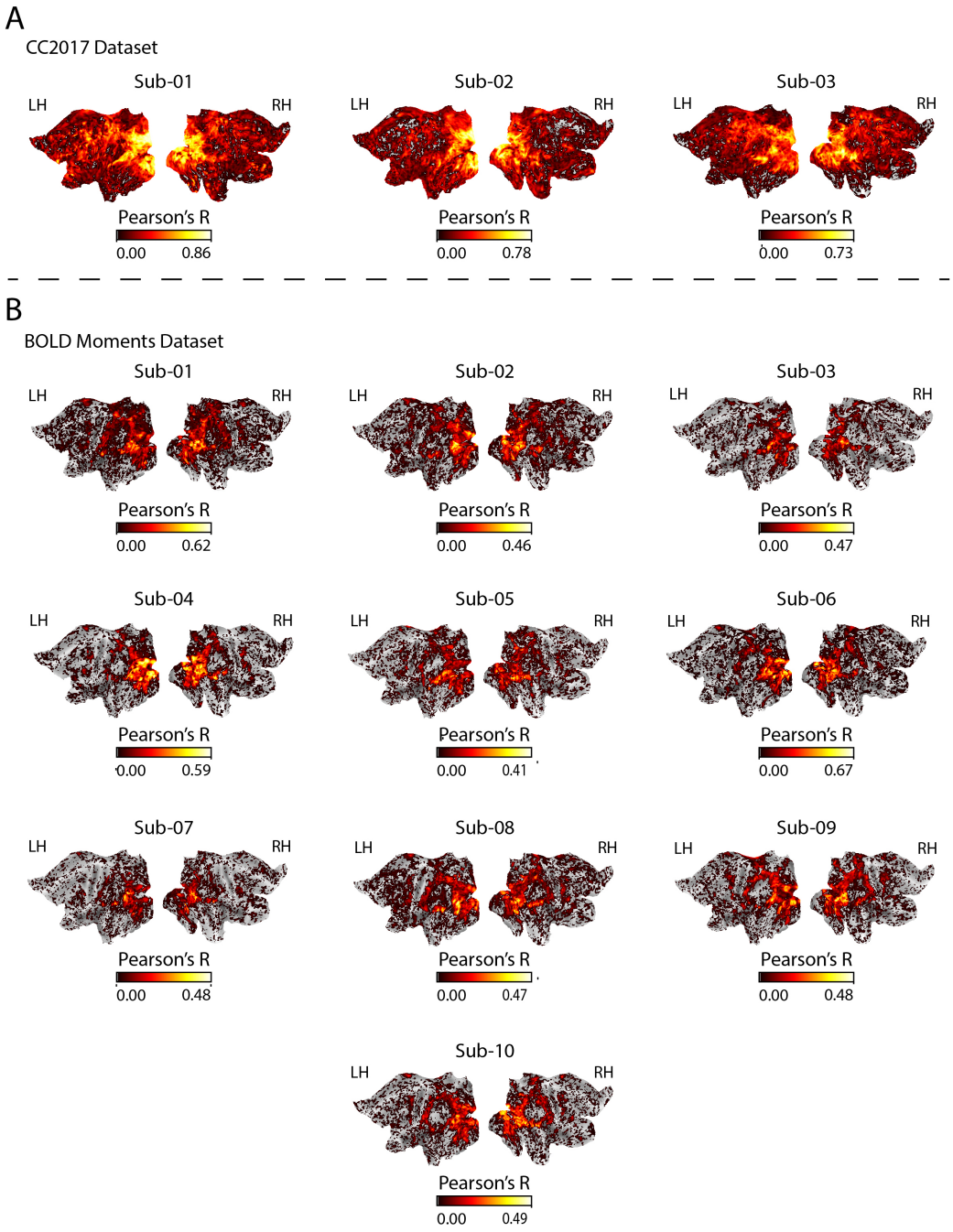

Figure 5. A. Each CC2017 subject's fMRI timeseries response to the two training set movie repetitions are correlated (Pearson's R) together and shown on a flattened brain. B. Each BOLD Moments Dataset (BMD) subject's vector fMRI beta estimates to the three training video repetitions are pairwise correlated (Pearson's R). The average of the correlation pairs are shown on a flattened brain. All correlations are clipped to a threshold of 0.01.