Brain Netflix: Scaling Data to Reconstruct Videos from Brain Signals

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

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

Abstract. The field of brain-to-stimuli reconstruction has seen significant progress in the last few years, but techniques continue to be subject-specific and are usually tested on a single dataset. In this work, we present a novel technique to reconstruct videos from functional Magnetic Resonance Imaging (fMRI) signals designed for performance across datasets and across human participants. Our pipeline accurately generates 2 and 3-second video clips from brain activity coming from distinct participants and different datasets by leveraging multi-dataset and multi-subject training. This helps us regress key latent and conditioning vectors for pretrained text-to-video and video-to-video models to reconstruct accurate videos that match the original stimuli observed by the participant. Key to our pipeline is the introduction of a 3-stage approach that first aligns fMRI signals to semantic embeddings, then regresses important vectors, and finally generates videos with those estimations. Our method demonstrates state-of-the-art reconstruction capabilities verified by qualitative and quantitative analyses, including crowd-sourced human evaluation. We showcase performance improvements across two datasets, as well as in multi-subject setups. Our ablation studies shed light on how different alignment strategies and data scaling decisions impact reconstruction performance, and we hint at a future for zero-shot reconstruction by analyzing how performance evolves as more subject data is leveraged.

1 Introduction

The human brain processes several million bits of information per second and compresses most of this information in ways that are still, in great part, unknown. What if we could partially recover this information by reading out brain signals? Reconstructing visual stimuli from brain activity is a promising way to investigate what information content is encoded in the human brain and how it may differ between participants. Reconstruction efforts typically require large amounts of neural-stimuli pairs to train high-parameter computational models, and inter-subject analyses additionally require repeated experiments across subjects. However, these neural-stimuli pairs are often not abundantly available due to their expensive and time-consuming collection process. Reconstructing dynamic videos, as opposed to static images, poses the additional challenges of temporal continuity.

Previous fMRI (Functional Magnetic Resonance Imaging) work has shown that a greater extent of cortex responds to videos than still images [53] [38] [2], that some cortical regions are selective for motion features [2] [17] [27], and that inter-subject correlations to movie viewings are high [16] [13], all suggesting that fMRI reliably encodes dynamic content. Humans do not all perceive videos in the same way, however; in-the-wild longform videos, such as an unedited video of the happenings in a park, show markedly lower inter-subject correlations than edited cinematic films [12] [13]. These higher inter-subject correlations in cinematic films are largely due to production effects, such as camera angle, editing, and lighting, that structure everyone's attention and interpretation of the content in a similar way [42]. The everyday human visual experience includes events sampled at both ends of this continuum, from unstructured meanderings down a busy street to a structured viewing of a work presentation. Thus, while videos can be reconstructed from brain activity [4], cross-subject reconstructions and the structure of the video stimulus set have yet to be systematically studied in the video reconstruction domain.

Here, we leverage two recent computational developments (latent diffusion models (LDMs) for video and masked-brain modeling) and several large-scale fMRI datasets to achieve high fidelity reconstructions of previously viewed 2- and 3-second videos. Latent diffusion models [34] are a class of generative models that generate high-fidelity images or videos, conditioned on inputs such as text, through a denoising diffusion process over a latent representation of the pixels. Masked brain modeling (MBM), an extension of masked training methods from natural language processing and computer vision to brain data, is a technique to learn robust self-supervised representations of brain signals by masking the signal and training a encoder-decoder network to reconstruct it. It has been shown to be a powerful way for a model to learn compressed representations of fMRI brain activity [4] [3].

We leverage these two concepts to propose a novel 3-stage pipeline that a) learns robust fMRI representations through masked brain modeling and alignment, b) regresses conditioning vectors from these representations, and c) reconstructs target videos with high fidelity, leveraging the estimated conditioning vectors. Importantly, our alignment and training strategy combines information from multiple subjects, as opposed to previous works that train one model per subject. This allows our model to not only improve its general reconstruction quality, but also to be able to generalize from one subject to another with much higher performance. Our contributions are as follows:

1. We introduce a 3-stage pipeline for high-quality video reconstruction that leverages multiple fMRI datasets and aligns inter-subject data, outperforming existing approaches.
2. We showcase a detailed ablation study that illustrates how the model performs in different conditions, including modeling and data variants.

3. We perform multi-subject and zero-shot reconstruction experiments to explore the possibility of a subject-agnostic visual reconstruction technology.

Together, this work offers both modeling and neuroscience insights for video reconstruction from brain activity.

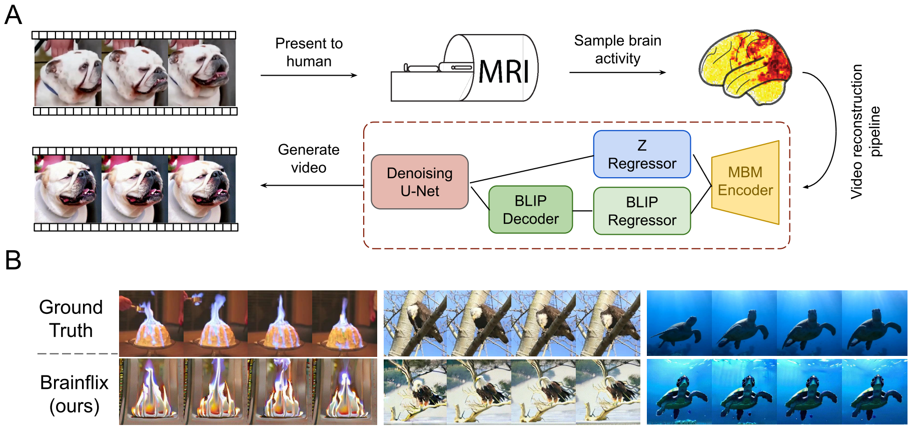

Fig. 1: Reconstruction of short videos from fMRI activity A) Brain responses of subjects viewing video clips was measured with fMRI and then used to reconstruct the seen video. B) We show three examples of the ground truth video (top row) and our reconstruction using our approach (bottom row).

2 Related Work

Diffusion models: Latent diffusion models (LDMs) [34] are a class of generative model that achieves photorealistic generations through a process of denoising latent representations. This denoising process can be conditioned on other inputs, such as text, in order to guide the model toward generating outputs with specific semantic meaning of visual appearance. These models have been successful in reconstruction tasks, including image generation, super-resolution and recoloring [6, 35–37, 43], audio generation [23], among others. Recent text-to-video latent diffusion models have generated high fidelity reconstructions of short videos faithful to both content and temporal dynamics [1]. Here, we leverage these advances in video generation to form the backbone of our framework: we use Zeroscope V2, which creates 3 second videos based on text conditioning.

Masked Brain modeling: Masked Brain Modeling [3] (MBM) is a recent method for pre-training fMRI reconstruction models. It is based on Masked Signal Modeling (MSM), a self-supervised learning task commonly used for pre-training in natural language processing and vision applications [5, 14, 50, 52]. In

MSM, some amount of the signal is masked during training, and an autoencoder is trained to predict the masked content, which can be fine tuned with additional training. In MBM, a percentage of the brain activity in masked and the model is tasked with reconstructing the complete input, which pushes the MBM model to learn compressed representations of fMRI brain activity.

Video reconstruction from brain activity: An early naturalistic video reconstruction work used a voxelwise motion-energy model to recover videos with similar spatiotemporal energy, showing that the BOLD signal captures dynamic visual information [29]. Other approaches used convolutional neural networks (CNNs) to reconstruct videos from fMRI data [51] [21], with one even leveraging a convolutional encoder-decoder framework to synthesize high framerate fMRI-video frame pairs [18]. A work using GANs employed dual spatial and temporal discriminators to achieve spatiotemporally faithful reconstructions [47], and variational autoencoders have been used to reconstruct videos frame-by-frame from a hidden latent space [11]. Most similar to our approach, [4] uses masked-brain modeling to train a vision transformer encoder that then feeds latent representations into a diffusion model to reconstruct the video.

3 Video fMRI Datasets

We train and evaluate our reconstruction method on a compilation of four large-scale fMRI datasets [19] [56] [45] [51] to obtain a large quantity and diversity of brain responses. In order to effectively perform inter-subject and inter-dataset analyses we analyze all fMRI responses in Cifti fsLR32k space (see data preprocessing details in the supplement.) [10] [7]. This preprocessing pipeline accurately registers gray matter voxels to a shared cortical surface mesh, thereby allowing the model to learn informative spatial activation patterns across datasets. From the cortical surface, we select a subset of 41 regions of interest (ROIs) from the Glasser Atlas (see supplement for details) [9]. This subset of cortex is isolates brain regions that might respond to dynamic stimuli [46] [24] [20] [8] [33] [40] [30] [31] [49] as well as reduce the computational load of model training. All datasets combined, our task-based video fMRI data draws from 43 subjects, over 28,100 short video segments, and over 123,000 fMRI response trials, and our resting state fMRI data totals over one thousand hours across 1084 subjects.

Human Connectome Project Dataset (HCP). The 1200-subject release of the Human Connectome Project [45] includes nearly an hour of resting state scans on 1084 subjects. During acquisition of resting state data, subjects were instructed to remain awake and fixate on a cross-hair but completed no task nor viewed any other visual stimulus. Resting state fMRI brain activity captures temporally correlated functional networks that can be useful for understanding cortical organization [41]. We utilize this time series data to pre-train our Masked Brain Modeling encoder (Figure 2).

BOLD Moments Dataset (BMD). The BOLD Moments Dataset (BMD) [19] consists of fMRI responses to 1,102 3-second videos for ten human subjects collected in a rapid event-related design. Beta values were estimated at each cortical vertex for each video using a general linear model (GLM). The stimuli were divided into a non-overlapping 1,000 video training set and a 102 video testing set. Each subject viewed the training set videos 3 times and the testing set videos 10 times to facilitate both inter-subject and intra-subject analyses. The 1,102 videos were sampled from the 1 million video Moments in Time dataset [25] (overlapping with the Multi-Moments in Time [26] and Memento10k datasets [28]) and capture naturalistic, in-the-wild content like home videos. We use BMD's five sentence text description metadata provided for each video to generate ground truth text embeddings that our pipeline learns to estimate. To the best of our knowledge, this work is the first time BMD has been used for video reconstruction.

Human Actions Dataset (HAD). The Human Actions Dataset (HAD) [56] is composed of 30 human subjects each viewing 720 two second videos sampled from the Human Action Clips and Segments (HACS) dataset [55]. Similar to BMD, data were collected in a rapid event-related design and beta values were estimated across cortex for each video. Each video is annotated with one of 180 human-centric actions, and each subject viewed four (different) videos from each of the 180 action categories. Video presentations were not repeated within or across subjects. For each subject, we define a testing/training split by randomly selecting one of the four videos per action category as a testing video. We use this dataset primarily to train Stage 1 in our pipeline. To the best of our knowledge, this work is the first time HAD has been used in video reconstruction.

CC2017 Dataset (CC2017). In the CC2017 dataset [51], three subjects viewed continuous training and testing movies composed of shorter 1̃0-15 second movie clips under a longform viewing paradigm. A training movie was composed from 374 video clips and randomly split into 18 eight minute segments. A testing movie was composed of 598 video clips and randomly split into five eight minute segments. Thus, each eight minute segment was composed of a concatenation of shorter video clips varying in duration. The clips are sourced from professionally shot videos, such as cinematic movies or stock advertising footage. The training and testing segments were repeated two and ten times per subject, respectively. Here, we divide the 8 minute segments into 2 second clips for reconstruction. We compose fMRI-video pairs by sampling the fMRI time series at an offset of 4 seconds (to account for the hemodynamic lag in the BOLD response) to capture peak BOLD activation corresponding to short snippets of the movie (2 second snippets, same as CC2017's acquisition TR). As this dataset is one of the main benchmarks used in previous work [4], we utilize this data extensively to train our regressors and provide reconstructions and evaluation metrics.

4 Reconstruction Pipeline

Approach Overview. Our reconstruction pipeline aims to regress key latent and conditioning vectors from brain signals to guide a pretrained video generation model towards reconstructing the original stimulus. This pipeline consists of

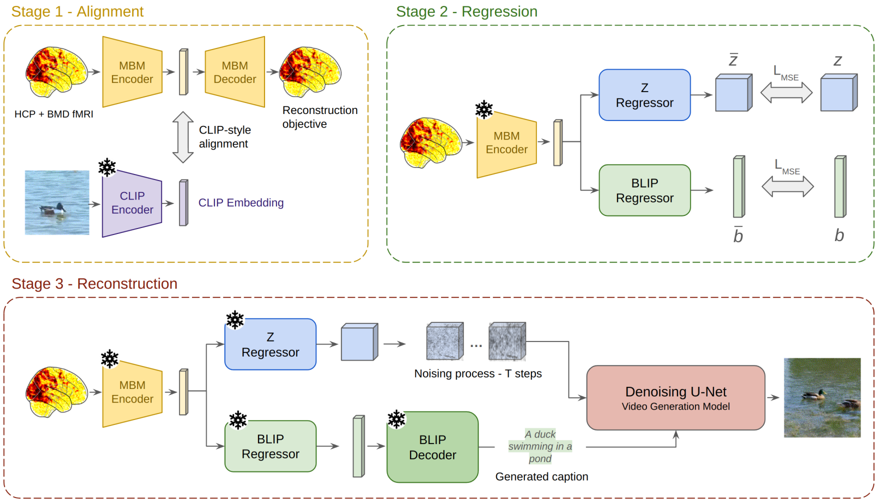

Fig. 2: Our proposed reconstruction framework. **Stage 1:** We train an encoder-decoder structure to reconstruct fMRI signals from a compressed representation. We then finetune the MBM encoder to align with a CLIP embedding of the video. **Stage 2:** we regress the latent vector (z) and blip embedding (b) of the target video from stage 1's aligned embedding. **Stage 3:** we use the regressed conditioning vectors to generate the video. The estimated vector z is artificially noised and then denoised with a pre-trained denoising U-Net, following a video-to-video procedure. The estimated vector b is sent to a pretrained BLIP model to generate a caption, which is used as a conditioning input to the U-Net.

3 stages. First, in an initial **alignment stage**, an encoder-decoder architecture is trained to reconstruct masked fMRI embeddings while simultaneously following a CLIP-like contrastive learning objective. This setup forces the encoder to map fMRI inputs into an embedding that captures both knowledge about the raw signal and knowledge about the semantic properties of the original stimuli. Second, the output of stage 1's encoder is used in a **regression stage**: we regress the necessary inputs required by our generative model with two MLPs that esstimate an initial latent vector (z) and BLIP embeddings (b) [22]. The BLIP embeddings are used to infer a conditioning caption. Third, our **reconstruction stage** uses the predicted latent vectors and BLIP embeddings to generate a video: Given a new fMRI input, we pass them through the regressors to predict the latent and conditioning vectors (z and b), re-noise z and use b to generate a caption. We feed the noised z and caption to a pretrained denoising U-Net (Zeroscope v2 [48]) to estimate the final reconstruction. We describe each stage in detail below and show a pseudocode implementation in the supplement Algorithm 1.

Stage 1: MBM and Alignment. Our model employs a straightforward encoder-decoder framework aimed at reconstructing the input fMRI signal by compressing it into a 1024-dimensional latent vector. During this phase, the input fMRI signal is divided into several patches, and certain patches are masked.

The encoder processes the signal with masked patches to produce a sequence of 1024-dimensional latent vectors. The decoder then attempts to transform these vectors back into the fMRI signal. Throughout the training process, the encoder-decoder structure is supervised through its ability to accurately reconstruct the masked patches. This approach, which we call Masked Brain Modeling (MBM), draws inspiration from the masked image modeling task [14] typical in the self-supervision literature. Its primary purpose is to acquaint the model with the spatial structure of the fMRI signal, rather than to imbue it with any semantic understanding. We first train our encoder-decoder pipeline on the masked modeling task alone with the large-scale HCP dataset [45] and HAD dataset [56], which ensures our embeddings learn basic fMRI structure. We utilize a simple MSE loss over the masked patches.

After this step, we further finetune the MBM encoder through a contrastive learning task. Contrastive learning is a powerful self-supervised technique to learn representations across modalities. We extract CLIP embeddings from captions describing each video, and we apply a contrastive loss over batches of encoder embeddings and CLIP embeddings. This contrastive loss follows [32]: for each batch, the loss enforces high cosine similarity between an fMRI embedding and its matching CLIP embedding (positive pair), while minimizing the similarity between that same embedding and all other CLIP embeddings in the batch (negative pairs):

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

Here, 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 train this model by following the training details in [3] for MBM and [39] for the alignment procedure, where we use a encoder-decoder model with patch size of 16, hidden dimension of 1024 and 24 layers in the encoder. We use a temperature of 0.9 and a 75% mask ratio following previous work [3].

Stage 2: Regression. To map from embeddings generated by the encoding model to latent and conditioning vectors suitable for the video generation pipeline, we utilize a multi-target regression model with regularization. Our regression models, instantiated as both a Ridge regression and a regularized MLP (see section 5.2 for comparisons), map outputs from Stage 1 into a latent vector z of shape (4, 15, 33, 33) and a BLIP [22] embedding b of shape (226, 768). The decision to map to BLIP vectors is rooted in how video generators handle noise in their conditioning inputs: similar to [44], we found that feeding a raw caption, even if imperfect, to our denoising U-Net performed better than feeding a regressed (and thus noisy) conditioning vector (comparisons in section 5.2). We hypothesize that the performance of video generators is susceptible to noise in the conditioning vector, and thus encapsulating the regression's imperfections at the level of BLIP's inputs allows the generator to receive clean (but imprecise) text tokens, which improves performance.

Our MLPs are composed of an initial linear layer, followed by 3 residual blocks and a final output layer. The initial layer and the residual blocks contain 2048 units, dropout regularization with $p = 0.3$, a GELU [15] activation and Batch Normalization. They are trained with an MSE loss and predict flattened vectors, which are then reshaped and used in Stage 3. Perfect regression at this stage equates to obtaining exact latent and conditioning vectors, which would result in a near-perfect reconstruction when fed to Stage 3.

Stage 3: Reconstruction. During video reconstruction, we extract the latent vector z and BLIP embedding b using the pretrained MBM encoder followed by the z and b regressors from stages 1 and 2. z undergoes a re-noising process, followed by a denoising procedure using a pretrained U-Net that follows [48]. We renoise the latents with a strength of 0.8 and apply 40 denoising steps, following insights from previous work [44]. Concurrently, b is decoded using BLIP's decoder to produce a semantically relevant caption that conditions the U-Net during the denoising process. We observed during experiments that enforcing more descriptive and less repetitve captions tended to improve the quality of BLIP outputs, so we implement decoding with a high repetition penalty of 6, a minimum length of 4 and a maximum length of 20. The integration of z and b helps facilitate video generation that is both visually and contextually aligned with the video represented in the original fMRI signal. Our video generation model is frozen during this process.

5 Experiments and Results

5.1 Comparisons to previous work

To showcase the quality of Brainflix, we obtain reconstructions from our full pipeline and compare to previous fMRI-to-video techniques. We report results over BMD and CC2017.

Implementation Details. We use the large-scale fMRI datasets introduced in section 3 for model training and finetuning. For all datasets, videos are downsampled to 15 FPS and resized to 224 × 224. We first train our alignment stage over the masked brain reconstruction task leveraging the HCP, HAD and BMD datasets for 200 epochs with a batch size of 300. We then finetune this model with a CLIP-like alignment objective over BMD and CC2017 simultaneously for 50 epochs with a batch size of 120. This aligns the output of the MBM encoder with embeddings of the captions describing each video. The captions are obtained through human labelers for BMD and synthetically for the rest of the datasets, leveraging the EILEV [54] video-to-caption model. Training for Stage 1 is done over 6 V100 GPUs, while training for Stage 2, a smaller scale endeavor given the smaller size of the MLPs, is performed over 2 Titan RTX GPUs.

Evaluation Metrics. Following previous work [4,44], 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

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
Brainflix (Ours)	0.189	0.681	0.871	0.203	0.183	0.691	0.767	0.145
Brainflix multi-subject (Ours)	**0.195**	**0.655**	**0.888**	**0.221**	**0.190**	**0.671**	**0.816**	**0.165**

Table 1: Quantitative comparison of our reconstruction performance against previous video reconstruction methods. Our methodology achieves state of the art on most metrics. Results from Kupershmidt and Mind-video on BMD are obtained through a reimplementation, as their code is not readily available.

from [4], 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 BLIP embeddings b , to observe how close our regressed embeddings are to ground truth.

Evaluation Datasets While we train on all datasets introduced in Section 3, we evaluate our reconstruction results on two large-scale video datasets, CC2017 [51] and BMD [19], to compare our method with previous work [4] [18] [47], as well as to introduce an additional dataset, BMD, with different subjects and video stimuli. These two datasets exhibit reasonable variety, complexity and semantic diversity, which makes them strong options for evaluation exercises.

Results. We showcase our quantitative results over BMD and CC2017 in Table 1. Results are reported for Subject 1 in both datasets. Both our single-subject and our multi-subject approaches outperform previous work. Our multi-subject model, trained on all subjects from a given dataset before predicting on one, showcases larger boosts in performance which provides evidence towards the advantages of multi-subject training. We show qualitative results in Figure 3 and 4: we observe that our model is able to reconstruct examples from BMD and CC2017 with strong structural reliability. We hypothesize that our model'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.

Human Evaluation. To further measure how preferable our reconstructions are, we collect qualitative evaluations of semantic fidelity using human judgments. We designed a simple web experiment where participants indicated which of 6 reconstructed videos was the best semantic match to a real video. One of the 6 videos was a reconstruction of the reference video. We calculated the percent of trials where the selected video corresponded to the reconstruction of the reference video. Table 2 reports Brainflix results for two datasets. Our best performing method reached a semantic fidelity score of 86% on BMD and 84% on CC2017. This substantially exceeds chance (16.66%), but remains below the

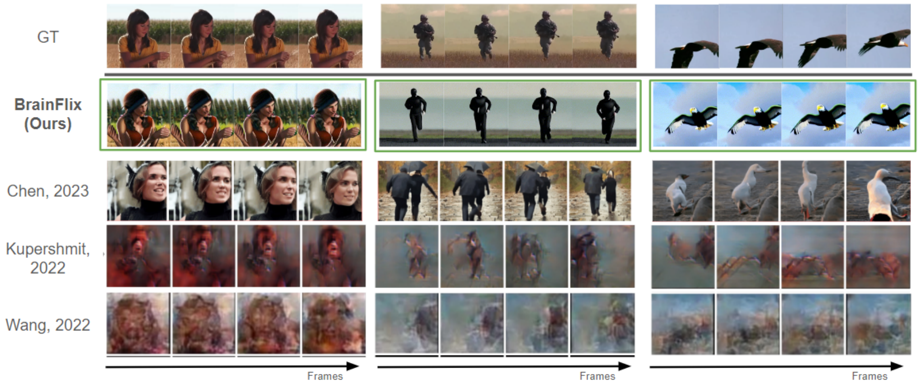

Fig. 3: Comparison to previous works on examples from CC2017. We show the ground truth videos in the first row, our results in the second row, and previous approaches in the following rows. Our model manages to capture structural similarity in a better way than previous approaches.

Method	BMD	CC2017
Baseline (Oracle)	94.1	92.8
MindVideo [4]	80.3	83.9
Brainflix (Ours)	**86.1**	**84.7**

Table 2: Human evaluation results. Accuracy on a 6-way retrieval task, where subjects select the ground truth video that most closely matches a reconstructed reference.

Training data	BMD		CC2017	
	SSIM	MSE	SSIM	MSE
Subject 1 only	0.123	0.721	0.699	**0.127**
Subject 1 + all subs	**0.129**	**0.702**	**0.710**	0.131
Subject 2 only	0.120	0.735	0.680	**0.122**
Subject 2 + all subs	**0.128**	**0.722**	**0.708**	0.126

Table 3: Multi-subject vs single-subject results. We observe that training on all subjects is value-adding for both datasets and both metrics. This effect is measurable across subjects.

best possible performance expected from oracle reconstructions given our backbone model (Zeroscope V2 [48]), which ranged from 87.9% to 98.0% depending on caption and seed (mean=94.1%).

5.2 Ablation studies.

To test the validity of our proposed pipeline, we perform a thorough set of ablation studies to showcase how each modeling decision impacts our final method. Our ablations cover two main dimensions: Modeling strategies and data management. Modeling strategies include details related to the inner working of our pipeline's stages, model components, and supervision paradigms. Data management relates to the inclusion of data from additional datasets and different subjects. We report results over BMD, but utilize CC2017 and HAD as pretraining datasets in some experiments as described below. We analyze the impact of the following features:

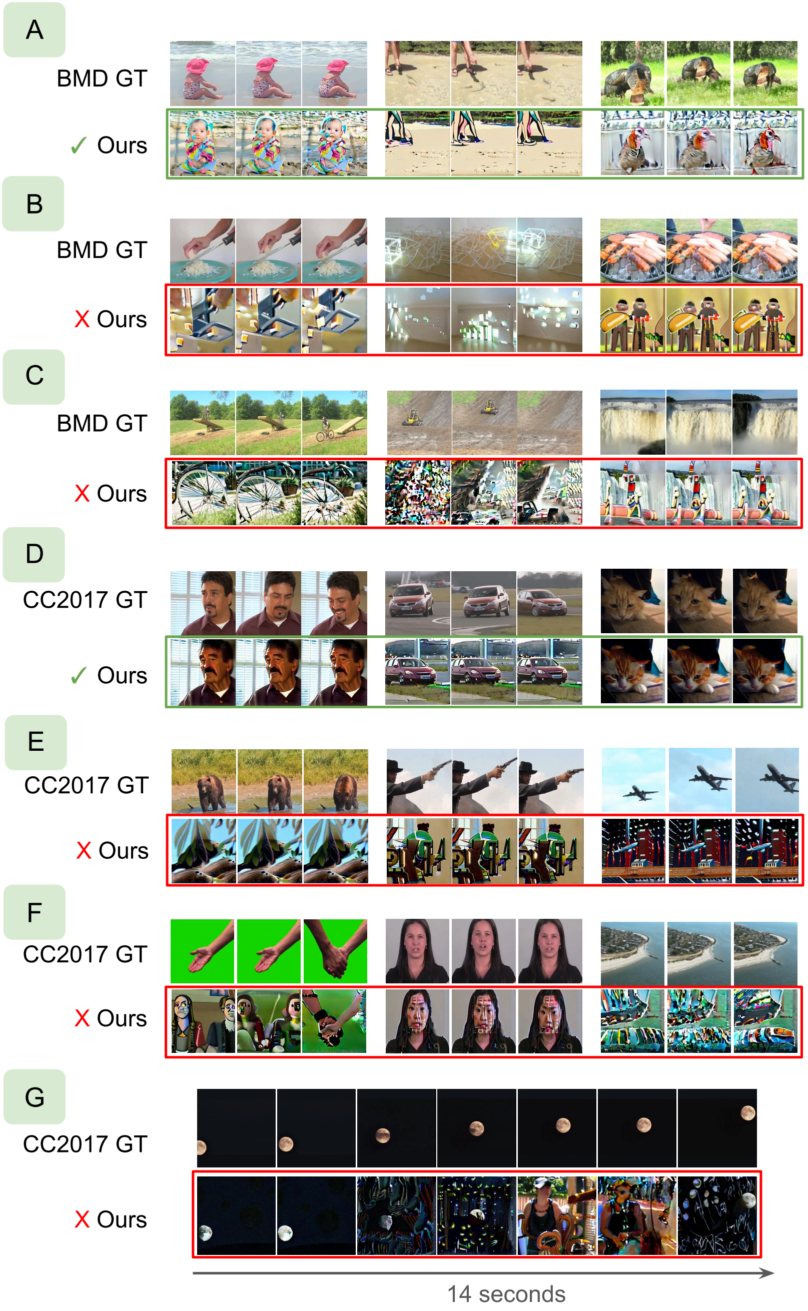

Fig. 4: Success and failure cases of reconstructions. A) Additional success cases from BMD. B) Additional failure cases from BMD. C) Specific failure cases from BMD exemplifying semantic dominance (left), high frequency noise (middle), and superposition of indiscernible shapes onto correct reconstruction (right). D) Additional success cases of CC2017. E) Additional CC2017 failure cases. F) Failure cases on CC2017 exemplifying semantic dominance (left), poor facial reconstruction (middle), high frequency noise (right). G) CC2017 failure case: Sudden failure to reconstruct video despite showing previous success in the segment.

- **Impact of masked brain modeling:** we measure how performance varies when we skip stage 1 and train a regressor from scratch to estimate the latent and conditioning vectors.
- **Impact of semantic alignment:** we observe how the contrastive learning objective affects the regression performance.
- **Impact of regression backbone:** we compare a linear regressor and a modern MLP.
- **Impact of conditioning target:** we analyze the difference between a direct regression to a 257x1024 text conditioning vector compared to regressing a BLIP vector used to build an estimated caption.
- **Impact of data scaling:** we observe how adding an additional pretraining dataset and multiple subjects affect the results.

We report results in Table 4. Our simple pipeline without alignment is the worst performer: the simple regression model struggles to correctly estimate conditioning vectors, showcasing a large MSE. When switching to an MLP, we observe improved MSE, but the reconstructed vectors are still not good enough to correctly retrieve the video. The boost introduced by adding Stage 1, even if it consists only of MBM training, is quite significant and hints at the importance of a structural pretraining stage.

Interestingly, regressing BLIP embeddings to then produce captions performs better than regressing raw conditioning vectors. The conditioning vectors accepted by zeroscope are hidden states from a CLIP-text model, of shape (257x1024). The BLIP embeddings, although of similar complexity, give a performance boost: we hypothesize that feeding the U-Net a clean caption, which BLIP produces even in the presence of a noisy input, allows the generation model to work in a familiar part of its latent space. This enables less deteriorated generations which translates to better metrics. Finally, scaling data showcases large improvements as well: both including HAD in Stage 1's pretraining and adding training data from other subjects gives boosts across metrics.

Multi-subject vs. single subject training. Most previous work showcases reconstruction performance on a per-subject basis, as brain responses even to the same stimulus differs between observers [16] [13]. These discrepancies arise mostly from individual differences in brain composition, thought patterns, and the sensitivity of fMRI data collection. As indicated in Section 4, we explicitly train with multiple subjects in our pretraining and regression stages, attempting to build a model that can leverage training data not only from subject s_0, but also $s_1 ... s_n$ to improve test set reconstruction performance for s_0. To shed light on the effect of training on multiple subjects, we report results over subject 1 and subject 2, comparing performance between a pipeline trained solely on subject N's data and a pipeline trained on all subjects. We show results in Table 3.

5.3 Zero-shot reconstruction

In this section, we ask to what extent a model could reconstruct one's visual experience without training on their brain activity. This is a very challenging

Method	BMD			
	SSIM	MSE	2-way	50-way
Oracle	0.891	-	1.0	0.980
Random	0.051	1.677	0.501	0.041
Image-only reconstruction	0.041	0.721	0.687	0.125
Linear regression	0.111	0.828	0.671	0.120
MLP regressor	0.123	0.721	0.699	0.127
+ MBM, no CLIP alignment	0.138	0.702	0.701	0.131
+ CLIP alignment	0.178	0.761	0.715	0.134
+ BLIP instead of raw cond vector	0.180	0.718	0.721	0.140
+ HAD pretraining	0.183	0.700	0.767	0.145
Full pipeline (Brainflix multi-subject)	**0.190**	**0.671**	**0.816**	**0.165**

Table 4: Ablation results and additional baselines on BMD's test set. We show results using: an Oracle baseline, where we feed the original video's ground truth z and b vectors to stage 3; a "random" baseline, where we instantiate random fMRI vectors and attempt to map those to z and b vectors, an image-only baseline, where, instead of using a video generator, we generate an image with an image generator (Stable Diffusion XL) from the fMRI, and duplicate the frame to match our videos' frame count; a pipeline without stage 1, using a linear regression for stage 2; a pipeline without stage 1, but with our final MLP backbone; a pipeline with a stage 1 comprised of masked-brain-modeling only; a pipeline with CLIP alignment, but still regressing a raw conditioning vector; a pipeline where we replace that regression target (raw conditioning vector) with a BLIP embedding; a pipeline with HAD pretraining, and our full pipeline, trained on multiple subjects. We observe that all design decisions tend to improve performance, and we see a significant boost from the addition of CLIP alignment.

task, further complicated by the structure of the stimulus to reconstruct (see [13] [42] [12]). Current work, including this paper, reports results by training and testing on the same subject, but future applications of this technology would greatly benefit if they could make *zero-shot predictions*. We analyze the impact of scaling data in this regime: we first train a model on data from subject 2, and attempt to reconstruct test data from Subject 1. As expected, performance deteriorates. However, we find that adding data from more subjects and datasets improves regression performance on subject 1 test data, without ever seeing training examples from that subject. We showcase these results in Figure 5. To further understand this effect, we compute between-subject correlations on the fMRI signals, observing that there are significant clusters of highly correlated responses even over brains from different subjects. We expand on these results in Section 4 of the supplement.

5.4 Limitations.

Figure 4 shows instances where our reconstruction method fails. Reconstructing complex, high spatial frequency videos (e.g., a crowd of people) remains challenging. In some cases, semantic reconstruction fails: we hypothesize that this is due

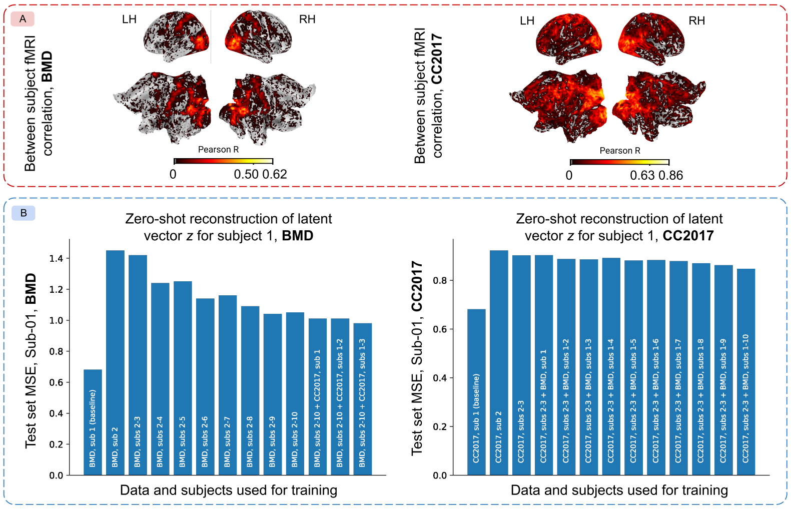

Fig. 5: fMRI Correlations and Zero-shot regression performance over latent vector z. A) Between subject pairwise correlations for subject 01 are plotted in an inflated brain and flattened brain, showing left hemisphere (LH) and right hemisphere (RH). Both BMD and CC2017 show areas of high correlation across subjects, mostly clustered around visual areas. We hypothesize this correlation might enable zero-shot strategies. B) Zero-shot performance as data from more subjects are added. The leftmost bar corresponds to MSE when training on subject 01's train set and evaluating on subject 01's test set. Subsequent bars train on different combinations of subjects except 01, and test on subject 01's test set. We observe that as we add more subjects and datasets, MSE decreases in both examples, hinting at the possibility of developing accurate zero-shot reconstruction with the right scale.

to issues in the fMRI data, as our system is vulnerable to situations where the subject's mind wanders and loses focus on the stimulus while data is recorded. Temporally faithful reconstructions are also limited by the temporal capabilities of the available video generation models.

6 Conclusion

We present a generative framework to reconstruct short clips from human brain responses. We leverage large pre-trained text-to-video models and two fMRI datasets to train a multi-stage pipeline to reconstruct a video observed by a given subject. Future generative neurotechnology in this vein could help determining the information recovery limits of a brain region, an uncharted terrain of neuroscience. Such technology could foster the development of therapeutic treatments to restore or compensate for sensory processing deprivation, as well as allow for more efficient 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.

References

1. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., 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 2023)
2. Buccino, G., Binkofski, F., Fink, G.R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R.J., Zilles, K., Rizzolatti, G., Freund, H.J.: Action observation activates premotor and parietal areas in a somatotopic manner: an fmri study. European journal of neuroscience **13**(2), 400–404 (2001)
3. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22710–22720 (2023)
4. Chen, Z., Qing, J., Zhou, J.H.: Cinematic mindscapes: High-quality video reconstruction from brain activity. arXiv preprint arXiv:2305.11675 (2023)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. 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)
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. Han, K., Wen, H., Shi, J., Lu, K.H., Zhang, Y., Fu, D., Liu, Z.: Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. NeuroImage **198**, 125–136 (2019)
12. Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., Heeger, D.J.: Neurocinematics: The neuroscience of film. Projections **2**(1), 1–26 (2008)
13. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R.: Intersubject synchronization of cortical activity during natural vision. science **303**(5664), 1634–1640 (2004)

14. 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)
15. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
16. 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)
17. Konen, C.S., 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)
18. Kupershmidt, G., Beliy, R., Gaziv, G., Irani, M.: A penny for your (visual) thoughts: Self-supervised reconstruction of natural movies from brain activity. arXiv preprint arXiv:2206.03544 (2022)
19. 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
20. 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)
21. Le, L., Ambrogioni, L., Seeliger, K., Güçlütürk, Y., van Gerven, M., Güçlü, U.: Brain2pix: Fully convolutional naturalistic video frame reconstruction from brain activity. Frontiers in Neuroscience **16**, 940972 (2022)
22. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
23. Liu, Z., Guo, Y., 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)
24. Logothetis, N.K., Sheinberg, D.L.: Visual object recognition. Annual review of neuroscience **19**(1), 577–621 (1996)
25. 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)
26. Monfort, M., Pan, B., Ramakrishnan, K., Andonian, A., McNamara, B.A., Lascelles, A., Fan, Q., Gutfreund, D., Feris, R.S., Oliva, A.: Multi-moments in time: Learning and interpreting models for multi-action video understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9434–9445 (2021)
27. Morrone, M.C., Tosetti, M., Montanaro, D., Fiorentini, A., Cioni, G., Burr, D.: A cortical area that responds specifically to optic flow, revealed by fmri. Nature neuroscience **3**(12), 1322–1328 (2000)
28. Newman, A., Fosco, C., Casser, V., Lee, A., McNamara, B., Oliva, A.: Multimodal memorability: Modeling effects of semantics and decay on video memorability. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 223–240. Springer (2020)

29. Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L.: Reconstructing visual experiences from brain activity evoked by natural movies. Current biology **21**(19), 1641–1646 (2011)
30. 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)
31. Peeters, R.R., Rizzolatti, G., Orban, G.A.: Functional properties of the left parietal tool use region. Neuroimage **78**, 83–93 (2013)
32. 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)
33. Rizzolatti, G., Sinigaglia, C.: The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nature reviews neuroscience **11**(4), 264–274 (2010)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., 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)
35. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
36. 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 (2022)
37. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
38. Schultz, J., Pilz, K.S.: Natural facial motion enhances cortical responses to faces. Experimental brain research **194**, 465–475 (2009)
39. 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)
40. Silver, M.A., Kastner, S.: Topographic maps in human frontal and parietal cortex. Trends in cognitive sciences **13**(11), 488–495 (2009)
41. 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)
42. Smith, T.J., Levin, D., Cutting, J.E.: A window on reality: Perceiving edited moving images. Current Directions in Psychological Science **21**(2), 107–113 (2012)
43. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
44. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14453–14463 (2023)
45. 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)

46. 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)
47. 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)
48. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
49. 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)
50. 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)
51. 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)
52. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., 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)
53. Yildirim, I., Wu, J., Kanwisher, N., Tenenbaum, J.: An integrative computational architecture for object-driven cortex. Current opinion in neurobiology **55**, 73–81 (2019)
54. Yu, K.P., Zhang, Z., Hu, F., Chai, J.: Efficient in-context learning in vision-language models for egocentric videos. arXiv preprint arXiv:2311.17041 (2023)
55. Zhao, H., Torralba, A., Torresani, L., 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)
56. 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)