NeuralFlix: A Simple While Effective Framework for Semantic Decoding of Videos from Non-invasive Brain Recordings

Jingyuan Sun[1*], **Mingxiao Li**[2*], **Marie-Francine Moens**[2]
[1]Department of Computer Science, The University of Manchester, UK
[2]KU Leuven, Belgium
jingyuan.sun@manchester.ac.uk
mingxiao.li@kuleuven.be
sien.moens@kuleuven.be

Abstract

In our quest to decode the visual processing of the human brain, we aim to reconstruct dynamic visual experiences from brain activities, a task both challenging and intriguing. Although recent advances have made significant strides in reconstructing static images from non-invasive brain recordings, the translation of continuous brain activities into video formats has not been extensively explored. Our study introduces NeuralFlix, a simple but effective dual-phase framework designed to address the inherent challenges in decoding fMRI data, such as noise, spatial redundancy, and temporal lags. The framework employs spatial and temporal augmentation for contrastive learning of fMRI representations, and a diffusion model enhanced with dependent prior noise for generating videos. Tested on a publicly available fMRI dataset, NeuralFlix demonstrates promising results, significantly outperforming previous state-of-the-art models by margins of 20.97%, 31.00%, and 12.30%, respectively, in decoding the brain activities of three subjects individually, as measured by SSIM.

Code — https://github.com/soinx0629/NeuralFlix

Introduction

The visual world that we encounter in daily life is highly dynamic, marked by fluid and continuously evolving sensory experiences (Varela, Thompson, and Rosch 2017). The human brain, with its remarkable complexity, continuously processes these visual inputs to construct a coherent narrative of perception (Bartels, Zeki, and Logothetis 2008; Nishimoto et al. 2011). Unraveling the layers of this intricate system and deciphering the associated neural processes represent considerable challenges (Chen, Qing, and Zhou 2023). Although invasive neural imaging methods such as ECoG offer certain advantages (Kanth and Ray 2020) over non-invasive techniques in studying human brain's visual processing—particularly in terms of an improved signal-to-noise ratio—their requirement for surgical implantation and associated safety concerns limit their widespread application, especially when compared with non-invasive methods. Consequently, a key objective in this field is to develop systems capable of decoding brain activity using non-invasive techniques to reconstruct human visual perceptions. Achieving such advancements could greatly enhance accessibility for individuals with sensory impairments while deepening our understanding of the neural foundations of visual perception.

Functional Magnetic Resonance Imaging (fMRI) a non-invasive neuro-imaging technology well-regarded for its high spatial resolution, which is crucial for recording detailed activations in the visual cortex and other brain regions (Hénaff et al. 2021). This capability is essential for reconstructing visual content based on brain activity. While reconstructing static images has seen notable progress (Chen et al. 2022; Sun, Li, and Moens 2023), reconstructing videos is still an area of active research. One of the main challenges with fMRI is its reliance on measuring changes in the blood-oxygen-level-dependent (BOLD) signal, which can exhibit spatial redundancy and sometimes lag behind actual neural activity due to the hemodynamic response (Uğurbil et al. 2013; de Zwart et al. 2009). Additionally, the non-invasive nature of fMRI means it can pick up noise from various physiological and scanner-related sources, which complicates the reconstruction of high-quality videos (Parrish et al. 2000).

To address these challenges, we have developed a two-phase framework called NeuralFlix which aims to reconstruct high-resolution videos from fMRI data which are semantically consistent with the videos viewed by human subjects. In the first phase, we use spatial masking and temporal interpolation to enhance fMRI data, while an optimized fMRI encoder is trained to resist disturbances from these augmentations. In the second phase, this trained fMRI encoder guides a video diffusion model in generating videos. We further improve this phase by introducing noise models that compensate for the low signal-to-noise ratio typical of fMRI data. Together, these innovations help our framework to convert complex and noisy fMRI data into precise and meaningful visual reconstructions, showcasing the potential of combining advanced neural imaging with machine learning to decode brain activity.

We tested our method on a publicly available dataset of fMRI-video pairs, involving three individuals watching

*These authors contributed equally.

videos. Our results show significant improvements over previous models in both detailed pixel accuracy and broader semantic understanding. Specifically, our approach improves upon the latest state-of-the-art (SOTA) model (Chen, Qing, and Zhou 2023) by significant margins: 20.97% in decoding brain activities of Subject 1, 31.00% in Subject 2, and 12.30% in Subject 3. Given that our method is straightforward and easy to implement, these findings highlight the potential of our approach to advance the technology of neural decoding and visual reconstruction.

Related Works

Decoding Visual Contents from Brain Activities

Reconstructing Images from Brain Activities Recent advancements in foundational models (Radford et al. 2021a; Wang et al. 2024; Sun et al. 2020) especially deep generative models (Ho, Jain, and Abbeel 2020) have spurred significant interest in the field of reconstructing perceived contents from brain activities (Sun et al. 2019, 2021; Zhao et al. 2024), and visual content such viewed and imagined images (Fang, Qi, and Pan 2021) gained attention in this field. Early research predominantly transformed fMRI signals into image features, which were then processed by fine-tuned Generative Adversarial Networks (GANs) to create images (Mozafari, Reddy, and van Rullen 2020). A notable example is the use of a pre-trained VGG network to extract hierarchical image features from fMRI data, which were then used to synthesize images through a GAN (Shen et al. 2019). More recent efforts in the last couple of years have shifted towards employing Diffusion Models. These models have successfully produced images that are both semantically coherent and visually more accurate (Qian et al. 2023; Lin, Sprague, and Singh 2022; Chen et al. 2022; Sun, Li, and Moens 2023). For instance, (Sun et al. 2023) significantly improved fMRI representation learning through denoising techniques and leveraged pixel-level guidance from image auto-encoders to effectively isolate vision-related neural activities from non-relevant noise.

Reconstructing Videos from Brain Activities While static image reconstruction has made strides, decoding videos from fMRI data remains challenging. Traditional methods treat video reconstruction as a sequence of individual image reconstructions, often resulting in lower frame rates and inconsistent quality (Wen et al. 2018). Improvements include using a linear layer to encode fMRI data and a conditional video GAN to enhance frame quality (Wang et al. 2022), though dataset limitations restrict their effectiveness. Further advancements came from (Kupershmidt et al. 2022) with a separable autoencoder for self-supervised learning, and from (Chen, Qing, and Zhou 2023), who used contrastive learning and spatial-temporal attention to enhance fMRI representation accuracy. However, these methods still face challenges, particularly in handling disruptions in space and time and in dealing with the noisy nature of fMRI data. This paper compares our enhanced approach to these baseline methods (Kupershmidt et al. 2022; Chen, Qing, and Zhou 2023; Wang et al. 2022).

Image and Video Generation with Diffusion Models

Diffusion Models, inspired by nonequilibrium thermodynamics, are probabilistic models that convert data into Gaussian noise and then reconstruct the original data, demonstrating excellent performance in content generation tasks such as text-to-image (Rombach et al. 2022) and 3D object creation (Poole et al. 2022). The typically iterative process, requiring hundreds of steps, has been streamlined by the Denoising Diffusion Implicit Model (DDIM) (Song, Meng, and Ermon 2020), which decreases the steps needed for high-quality output. Enhancements such as the integration of ordinary differential equation solvers (Lu et al. 2022a,b; Liu et al. 2022), variance optimization (Bao et al. 2022), reduction of exposure bias (Li et al. 2024a; Ning et al. 2024, 2023), and improved noise schedulers (Nichol and Dhariwal 2021) have further accelerated inference and enhanced generative quality. Early uses of Diffusion Models for video generation were led by the introduction of the 3D diffusion UNet by VDM (Ho et al. 2022b). Later, ImageN (Ho et al. 2022a) developed a cascaded sampling framework with super-resolution techniques for producing high-resolution videos. Additional advances include temporal attention mechanisms by Make-A-Video (Singer et al. 2022), integration within latent diffusion models by MagicVideo (Zhou et al. 2022) and LVDM (He et al. 2022), and the incorporation of additional guiding signals (Li et al. 2024b) to improve video generation. In this work, we modify the image diffusion model to support video by adding a temporal layer after each spatial layer and incorporating dependent noises.

Method

Our method consists of fMRI feature leaning and video decoding two-phase framework for reconstructing videos from fMRI-recorded brain activities. Phase 1 involves tuning a pre-trained fMRI encoder with spatial and temporal augmented contrastive learning to align fMRI data with CLIP's text and image features, enhancing the extraction of semantic information from fMRI signals. Phase 2 uses the trained fMRI encoder to guide a video diffusion model, incorporating dependent prior noise to compensate for fMRI's low signal-to-noise ratio.

In the first subsection, we delve into learning fMRI representation with spatial and temporal augmented contrastive learning. In the second subsection, we elaborate on the design of prior noise for the video diffusion model to decode more coherent videos from brain activity. In the third one, we describe our experimental approach for analyzing our results, aiming to clarify the contribution of each brain region throughout different stages of learning.

FMRI Feature Learning

Pre-training and FMRI Input Format Given the limited availability of fMRI-video pair data, we leverage a pre-trained fMRI representation space using a model proven effective in image reconstruction from fMRI, as introduced by

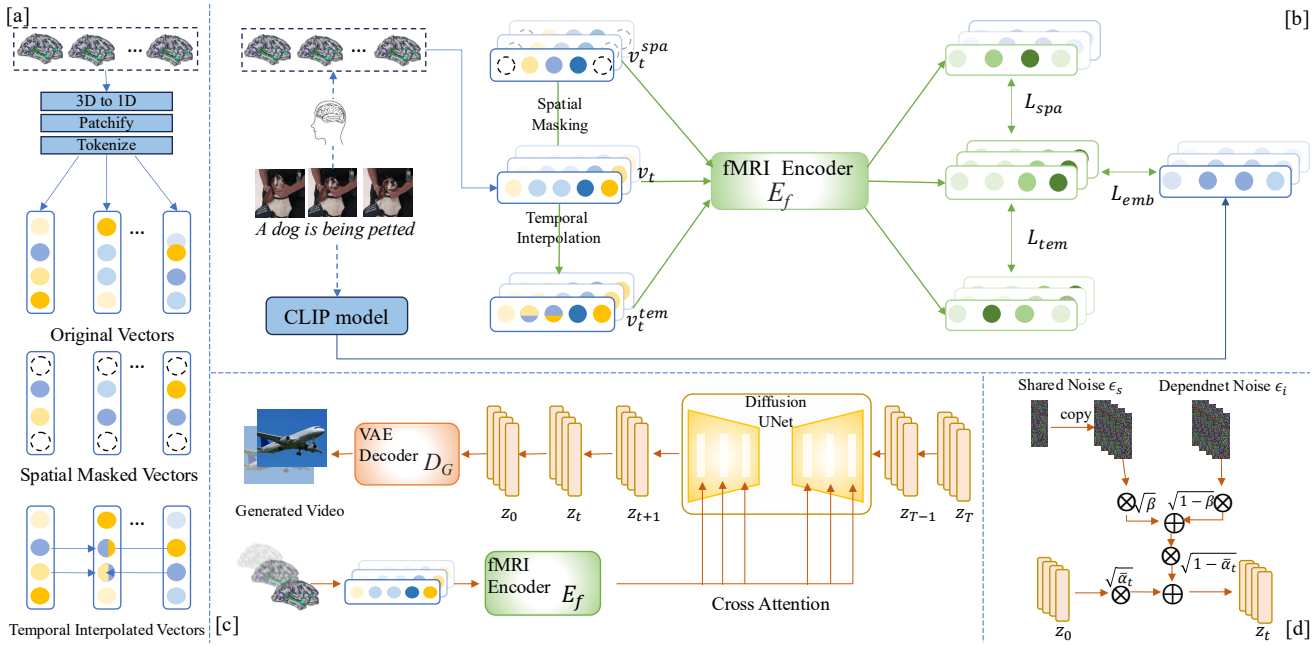

Figure 1: The proposed double-phase framework to reconstruct seen videos from fMRI. [a]: Using spatial masking and temporal interpolation to produce augmented samples [b]: Phase 1 trains an fMRI encoder to map from fMRI to CLIP text and image embeddings with contrastive learning. [c]: Phase 2 conditions generation of a diffusion model with Phase 1's fMRI encoder incorporating dependent noises. [d]: Dependent noise generation.

(Sun et al. 2023). This model uses a Vision Transformer-based encoder to process masked fMRI signals and a decoder for restoring unmasked signals, utilizing a double-contrastive masked autoencoding (DC-MAE) technique, optimized on the HCP dataset (Van Essen et al. 2013). The "Double-Contrastive" model optimizes contrastive losses through two contrasting operations during fMRI representation learning (Sun et al. 2023; Chen, Qing, and Zhou 2023; Chen et al. 2022; Sun, Li, and Moens 2023). We use this fMRI encoder to establish a robust pre-trained representation space. Since the pre-trained fMRI encoder was trained on the HCP dataset without video labels, there was no concern of possible data leakage with fMRI-Video dataset we are using in this paper.

In our approach to video decoding, we treat fMRI data as a sequence of 3D tensors. However, due to the scarcity of fMRI-video pairs, we opt for a simpler model structure that does not require extensive parameterization to capture the 3D structure of fMRI data. Following successful precedents (Sun et al. 2023; Chen et al. 2022), we convert the 3D fMRI data from the visual cortex into 1D, align it with the visual processing hierarchy, segment it into uniform patches, and then tokenize it. Traditionally, neural decoding methods have simplified the conversion of fMRI data to video frames by using a fixed "fMRI frame window." This approach, however, does not effectively handle the inherent temporal delay in fMRI data. To address this, we use a sliding window technique, defined as $\boldsymbol{v_t} = \{v_t, v_{t+1}, \ldots, v_{t+w-1}\}$, where $v_t \in \mathbb{R}^{n \times p \times b}$ represents the token embeddings at time t, and n, p, and b denote the batch size, patch size, and embedding dimension, respectively. This results in $\boldsymbol{v_t} \in \mathbb{R}^{n \times w \times p \times b}$, where w is the window size. Our decoding algorithms apply to these windows, taking into account the specific spatiotemporal characteristics of fMRI data as we will discuss in subsequent sections.

Spatial and Temporal Augmentation To tackle the dual challenges of limited fMRI-video pair data and the inherently low signal-to-noise ratio in fMRI, we propose a novel method for training a noise-robust fMRI encoder. Central to our approach is the cognitive plausible augmentation of samples for contrastive learning, tailored to the unique spatial and temporal characteristics of fMRI data. Conventional computer vision augmentations such as cropping and rotation are not suitable for fMRI images, which require maintaining the spatial integrity that correlates with neurological function. After extensive review (Glover 2011; Buxton 2009) and consultations with experts, we employ two primary augmentation techniques: spatial masking and temporal interpolation.

For spatial masking, we randomly select a portion of the tokens in $\boldsymbol{v_t} \in \mathbb{R}^{n \times w \times p \times b}$ and set them to zero. Specifically, $\gamma_{spa} b$ values are zeroed out in the fourth dimension, b, with γ_{spa} as a tunable hyperparameter. The positions to be masked are consistent within the same window but vary across different batches.

Temporal interpolation involves replacing randomly selected frames within a window with interpolations of other frames based on their temporal proximity—the farther away a frame, the less it contributes. This method uses weighted interpolations for fMRI sequences, unlike static image aug-

mentation methods like CutMix which involve cropping and pasting. Mathematically, for a window of w fMRI frames $\boldsymbol{v_t}$, and a selected i^{th} frame v_{t_i}, the interpolated frame $\hat{v_{t_i}}$ is calculated as:

$$\hat{v_{t_i}} = \sum_{j=1, j\neq i}^{n} \left(1 - \frac{|i-j|}{n}\right) v_{t_j} \tag{1}$$

The original frame v_{t_i} is then replaced by $\hat{v}_{t_i}$.

The degree of interpolation is governed by the temporal interpolation ratio γ_{tem}, another adjustable hyperparameter. Details on the settings and impacts of γ_{tem} and γ_{spa} will be further explored in the Section 5.2 Ablation Study.

Contrastive Mapping In our method, we employ a vision-Transformer-based fMRI encoder to process fMRI token vectors, aligning them with the CLIP model's text and image embeddings. This process is enhanced by incorporating contrastive learning using augmented examples as described in the previous subsection.

Initially, we utilize the pre-trained CLIP model (Radford et al. 2021b) to encode the stimuli used in the collection of fMRI data. For each video, captions are generated using the pre-trained BLIP model (Li et al. 2022) and then encoded with the CLIP model to produce text embeddings. Similarly, we process each video frame to generate corresponding image embeddings. The fMRI encoder is fed with the token vectors and trained to map from fMRI to CLIP embeddings. Additionally, the fMRI encoder is fed with both the spatially and temporally augmented examples and is optimized with contrastive losses to learn fMRI features robust to the spatial and temporal disturbances.

The formal representation of our loss functions, considering the original fMRI token vectors $\boldsymbol{v_t}$, their spatially augmented version $\boldsymbol{v_t^{spa}}$, and temporally augmented version $\boldsymbol{v_t^{tem}}$, is as follows:

$$\begin{aligned} L_{tem} =& L_{IfNCE}[E_f(\boldsymbol{v_t}), E_f(\boldsymbol{v_t^{tem}})] \\ L_{spa} =& L_{IfNCE}[E_f(\boldsymbol{v_t}),\ E_f(\boldsymbol{v_t^{spa}})] \\ L_{emb} =& L_{IfNCE}[E_f(\boldsymbol{v_t}),\ e_t^{txt}] + L_{IfNCE}[E_f(\boldsymbol{v_t}),\ e_t^{img}], \end{aligned} \tag{2}$$

where $L_{InfoNCE}$ is the InfoNCE loss and E_f denotes the fMRI encoder. e_t^{txt} and e_t^{img} mean the CLIP text and image embeddings. We aim to optimize these losses jointly, with the overall loss function being defined as:

$$L_{E_f} = \mu_{spa} L_{spa} + \mu_{tem} L_{tem} + L_{emb} \tag{3}$$

In this equation, μ_{spa} and μ_{tem} are hyperparameters that adjust the weight of the corresponding losses. The setting and effects of μ_{spa} and μ_{tem} will be detailed in Section 5.2.

Generation with Diffusion Model

Prelimiaries Diffusion Models (Sohl-Dickstein et al. 2015) show significant potential in generating both images and videos. In this work, we adopt the widely used Stable Diffusion (SD) (Rombach et al. 2022) as the baseline model, known for its efficient denoising capabilities in the image's latent space, which requires considerably fewer computational resources. During training, the SD begins by using a KL-VAE (Kingma and Welling 2013) encoder to convert image x_0 to latent space: $z_0 = \mathcal{E}(x_0)$. It then progressively transforms this latent representation into a Gaussian noise, following the equation:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \tag{4}$$

Here, the ϵ represents a noise sampled from a normal distribution: $\epsilon \sim \mathcal{N}(0, 1)$. $\bar{\alpha}_t$ is the predefined noise schedule. The model is trained to predict the added noise at each step, and the loss function could be formulated as :

$$\mathcal{L}_t^{simple} = E_{t, x_0, \epsilon_t \sim \mathcal{N}(0,1)}[\|\epsilon_t - \epsilon_\theta(z_t, t, c)\|_2^2] \tag{5}$$

t is the diffusion time step, and c is the text prompt condition. During inference, the SD gradually reconstructs an image aligned with the provided text prompt from Gaussian noise. The denoised results are then processed through the decoder of the KL-VAE to reconstruct the colored images from their latent representation: $x_0 = \mathcal{D}(z_0)$.

Dependent Prior Video Diffusion. Following previous studies (Wu et al. 2023; Chen, Qing, and Zhou 2023), we utilize a pre-trained text-to-image Stable Diffusion (SD) model as our foundational video generator. While adept at creating high-quality individual frames, the original SD model lacks temporal coherence for video generation. To address this, we modify it by converting 2D convolutions to pseudo 3D and adding a temporal attention layer after each spatial self-attention layer. This modification introduces temporal awareness, allowing each visual token to attend to tokens from the previous two frames. The temporal attention layer operates as:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

with Q, K, V being the query, key, and value matrices, and W^Q, W^K, W^V as learnable parameters.

For decoding brain activity into video, we start by sampling m latent codes from Gaussian noise and progressively refine them using the fMRI representation. Given the low signal-to-noise ratio in fMRI signals, enhancing video quality is challenging. Previous research (Ge et al. 2023) demonstrates that employing a deterministic ODE solver in the generative process of the SD model results in a high correlation of initial noise in frames from the same video. Similarly, it has been observed that fMRI signals from similar visual stimuli exhibit a high degree of correlation. Based on these observations, we utilize correlated noise as a form of prior knowledge within the generative model and the fMRI decoding process. To create a sequence of dependent noise, where each noise is sampled from Gaussian Distribution with a mean of zero and a variance of one, we divide each noise into two components: ϵ_s and ϵ_i^j, and the dependent noise is obtained by following formula:

$$\epsilon^j = \sqrt{\beta} \cdot \epsilon_s + \sqrt{1 - \beta} \cdot \epsilon_i^j \tag{7}$$

where $\epsilon_s \sim \mathcal{N}(0, 1)$ and $\epsilon_i \sim \mathcal{N}(0, 1)$. $\sqrt{\beta}$ is the hyperparameter conditioning the noise ratio, whose setting and effects are discussed in Section 5.2's Ablation Study. A visualization of generating dependent noise is presented in Figure 1 [d]. During training, we substitute the original noise

in SD model with our customized dependent noise to generate noisy latent codes at each time step. Conversely, in the generative phase, the process begins with the introduction of our dependent noise.

Experimental Setup

Evaluation Metrics and Baselines

We evaluate our approach using both pixel-level and semantic-level metrics. For pixel-level assessment, we use the Structural Similarity Index Measure (SSIM) (Wang et al. 2004). For semantic-level evaluation, we perform a 50-way top-1 accuracy classification test, where SSIM and classification accuracy are calculated for each frame against its ground truth. The classification test involves an ImageNet classifier, and success is defined when the ground truth class is within the top-K probabilities of the predicted frame's classification from 50 randomly selected classes, including the ground truth. This test is repeated 100 times to calculate an average success rate. For video-based metrics, a similar classification approach is used with a video classifier based on VideoMAE (Tong et al. 2022), trained on the Kinetics-400 dataset (Kay et al. 2017) for action classification across 400 categories, covering various motions and human interactions. We compare our methodology with prior works using these metrics, including studies from Chen, Qing, and Zhou (2023), Wen et al. (2018), Wang et al. (2022), and Kupershmidt et al. (2022), citing their results directly from the SOTA report by Chen, Qing, and Zhou (2023) for fair evaluation.

Dataset

We utilize a publicly accessible fMRI-video dataset (Wen et al. 2018) that contains fMRI recordings from three participants alongside corresponding video stimuli. These data were acquired using a 3T MRI scanner, with a repetition time of 2 seconds. The training set consists of 18 distinct video clips, each lasting 8 minutes, totaling approximately 2.4 hours and producing 4,320 paired training instances. The test set includes five 8-minute video clips, collectively lasting 40 minutes and producing 1,200 test fMRI samples. Each video is presented at 30 FPS and spans a broad range of subjects, including animals, humans, and natural landscapes. The semantic categories do not fully match between the training and test sets, with an overlap of 0.56 assessed using the intersection_set/union_set metric. We allocate 20% of the training data for validation. For consistency with the reported state-of-the-art approach (Chen, Qing, and Zhou 2023), we reduce the original frame rate to 3 FPS and select a window size of 2, resulting in each fMRI frame corresponding to six video frames. With this configuration, each fMRI sample can be used to reconstruct a 2-second video segment, and this approach can be extended to longer sequences depending on available GPU memory.

Implementation Details

In the first phase, we use a Vision Transformer (ViT)-based fMRI encoder pre-trained on large-scale fMRI data (Sun et al. 2023). The encoder is initially pre-trained with a mask ratio of 0.75 and a patch size of 16, across 24 layers, and embedding dimension of 1024. It also includes a projection head transforming token embeddings into a 77×768 dimensionality. The subject embedding is in dimension 512 and the subject masker is a 1-layer Vit encoder with the nearly same setup as in the fMRI encoder. In the second phase, we employ Stable Diffusion V1-5 (Rombach et al. 2022), fine-tuned to process videos at 256x256 resolution and 3 FPS frame rate. This involves modifying spatial attention, cross-attention, and introducing a temporal layer. Training involves 1000 steps with a learning rate of 2×10^{-5} and batch size of 14. Post-training, the text encoder of the video diffusion model is replaced with our fMRI encoder. Further fine-tuning is carried out for spatial self-attention, visual-fMRI cross-attention, and temporal attention, using a learning rate of 2×10^{-5} and a batch size of 24 on a single A100 GPU.

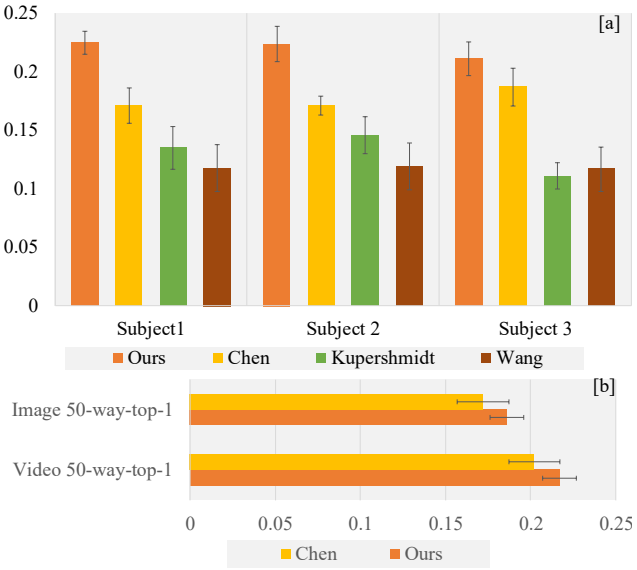

Figure 2: Comparisons of Structural Similarity Index Measure (SSIM) Scores and 50-way-top-1 Image/Video Classification Accuracy. [a] Comparing the SSIM scores of our method with other three benchmarks on Subject 1, 2 and 3. [b] Comparing our method's 50-way Image and Video Classification Accuracy with previous SOTA model on Subject 1.

Results

Video Reconstruction Performance

In this section, we first compare our methodology with baselines focusing on Structural Similarity Index Measure (SSIM) scores and classification accuracy. We first discuss the results of subject-wise decoding to compare with previous work. Our SSIM results, shown in Figure 2 [a], indicate our method sets a new state-of-the-art (SOTA) with scores of 0.225, 0.224, and 0.211 for Subject 1, 2, and 3. Specifically, our method surpasses the previous (Chen, Qing, and

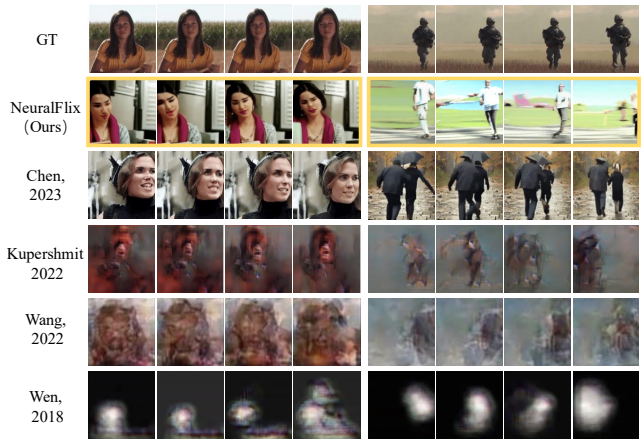

Figure 3: Comparison of the decoded results considering our framework NeuralFlix and baselines.

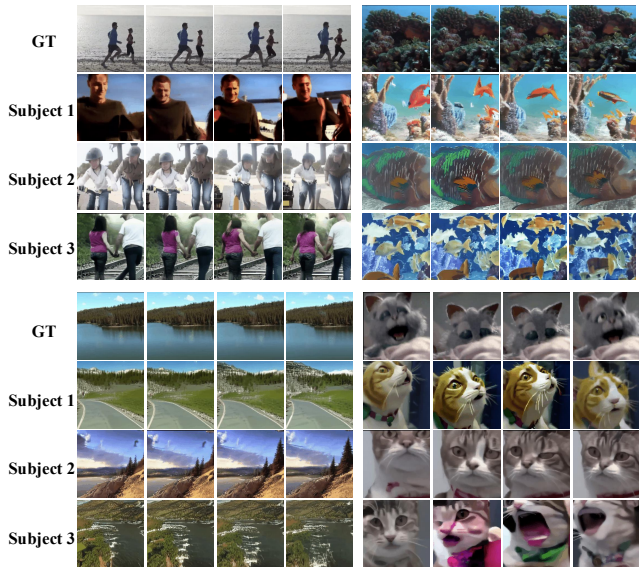

Figure 4: Decoded results of our framework NeuralFlix from all three subjects in the dataset.

Zhou (2023)) by a significant margin of $\mathbf{20.97}\%$ in Subject 1, $\mathbf{31.00}\%$ in Subject 2 and $\mathbf{12.30}\%$ in Subject 3. (You may refer to hyperparamter settings in Table 1's experiment 10 for reproduction of Subject 1's decoding performance.) Figure 2 [b] further shows that our method achieves superior accuracy in both 50-way Image and Video Classification Accuracy.

Qualitative comparisons in Figure 3 show that, unlike other models that yield blurry or unrecognizable outputs, our method and Chen, Qing, and Zhou (2023)'s approach produce high-quality, semantically accurate videos. A detailed comparison with the previous SOTA method (Chen, Qing, and Zhou (2023)) in Figure 5 reveals our superior semantic alignment with ground truth videos. For instance, where our method accurately generates a video of a swimming turtle, Chen, Qing, and Zhou (2023)'s output shows a group of fish. The versatility of our model is further demonstrated in Figure 4 which shows our model's reconstructed results on all subjects, where it consistently decodes high-quality, semantically accurate videos from different subjects. We present more video frames generated by our model as compared with ground-truth in the Figure 6.

Ablation Study

In this subsection, we conduct an ablation study on the validation set to evaluate the impact of each model component and the significance of hyperparameters on video decoding performance.

Spatial and Temporal Augmentation: The ratios for spatial masking and temporal interpolation modify fMRI token vectors to create augmented samples. As shown in experiments [0-3] in Table 1, spatial augmentation notably improves decoding, with a spatial mask ratio of 0.2 yielding the best results. However, excessive masking reduces the SSIM of reconstructed samples, indicating the need for a balanced approach to leverage fMRI's spatial redundancy effectively. Similarly, a temporal interpolation ratio of 1/3 enhances performance without the negative effects seen with higher ratios.

Augmentation Loss Weight: The augmentation loss weights influence their contribution to optimizing the fMRI encoder. Experiments [1, 5-7] in Table 1 show that balancing augmentation loss between spatial and temporal aspects improves performance, with both weights set to 1 achieving the highest SSIM. This underscores the importance of equally capturing spatial and temporal features of fMRI data for accurate video decoding.

Dependent Noise Ratio: Introducing dependent noise addresses inherent fMRI noise, improving video decoding as demonstrated in experiments [1, 8-10] in Table 1. Our hypothesis that dependent noise as a prior enhances video reconstruction with Diffusion Models is supported. However, increasing the noise ratio ($\sqrt{\beta}$) too much negatively impacts performance, as it causes frames to become overly similar, leading to static-like videos with insufficient variation.

Reliance on Diffusion Model Priors: To determine whether our model decodes video dynamics from fMRI sequences rather than relying solely on diffusion model priors, we re-ran the pipeline using time-averaged brain signals (results not shown in Table 1). This led to a significant performance drop, with 50-way-top-1 video classification accuracy on Subject 2 decreasing from 0.187 ± 0.015 to 0.157 ± 0.013, and on Subject 3 from 0.195 ± 0.014 to 0.158 ± 0.012. Time-averaging inhibits the model's ability to predict motion by using temporal changes in brain signals, confirming that our model's dynamic predictions rely on brain signals rather than diffusion model priors. This also validates the effectiveness of our metrics in assessing video dynamics.

Conclusion

In conclusion, this study introduces a novel dual-phase framework for decoding high-quality videos from fMRI data, effectively tackling challenges like spatial redundancy

Ablated Parameter	ID	Spatial Mask Ratio (γ_{spa})	Temporal Interpolation Ratio (γ_{tem})	Spatial Loss Weight (μ_{spa})	Temporal Loss Weight (μ_{tem})	Dependent Noise Use	Dependent Noise Ratio (β)	SSIM
Spatial Mask Ratio	0	\	\	\	\	\	\	0.171
	1	0.2	1/3	1	1	No	\	0.204
	2	0.4	1/3	1	1	No	\	0.184
	3	0.6	1/3	1	1	No	\	0.184
Temporal Jittering Ratio	1	0.2	1/3	1	1	No	\	0.204
	4	0.2	1/2	1	1	No	\	0.186
Spatial and Temporal Loss Weight	1	0.2	1/3	1	1	No	\	0.204
	5	0.2	1/3	0.5	0.5	No	\	0.191
	6	0.2	1/3	0.25	0.75	No	\	0.186
	7	0.2	1/3	0.75	0.25	No	\	0.185
Dependent Noise Ratio	1	0.2	1/3	1	1	No	\	0.204
	8	0.2	1/3	1	1	Yes	0.25	0.242
	9	0.2	1/3	1	1	Yes	0.5	0.230
	10	0.2	1/3	1	1	Yes	0.75	0.225

Table 1: Ablation study about NeuralFlix's important hyperparameters' effects on final video decoding performance measured by SSIM, including spatial mask ratio, temporal interpolation ratio, spatial loss weight, temporal loss weight, using of dependent noise and dependent noise ratio.

Figure 5: Additional comparisons of decoded results between our method NeuralFlix and the previous state-of-the-art (SOTA) model.

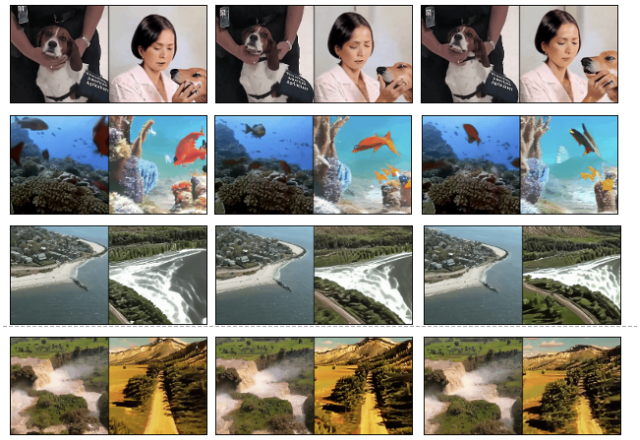

Figure 6: Additional comparisons of decoded results of our method NeuralFlix (right) to the ground-truth (left).

and temporal lags in fMRI signals. Our method, which combines spatial-temporal contrastive learning with an enhanced video diffusion model, shows notable improvements over existing models in both SSIM and semantic accuracy.The empirical results, benchmarked against prior works, demonstrate the efficacy of our approach. This research not only advances the the technique of neural decoding but also explores new avenues for exploration in neural imaging and cognitive neuroscience, with potential applications in understanding human cognition and developing assistive technologies for people with disabilities.

Ethical Statement

The fMRI data employed in our training have undergone processing to ensure that they do not contain any information that could be directly traced back to individual participants. Furthermore, the original collection of this fMRI data adhered to strict ethical guidelines and reviews, as detailed in the respective source publications.

Acknowledgements

This work is funded by the European Research Council under the European Union's Horizon 2020 research and innovation program (projects CALCULUS, H2020-ERC-2017-ADG 788506).

References

Bao, F.; Li, C.; Zhu, J.; and Zhang, B. 2022. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models.

Bartels, A.; Zeki, S.; and Logothetis, N. K. 2008. Natural vision reveals regional specialization to local motion and to contrast-invariant, global flow in the human brain. *Cerebral cortex*, 18(3): 705–717.

Buxton, R. B. 2009. *Introduction to functional magnetic resonance imaging: principles and techniques*. Cambridge university press.

Chen, Z.; Qing, J.; Xiang, T.; Yue, W. L.; and Zhou, J. H. 2022. Seeing Beyond the Brain: Masked Modeling Conditioned Diffusion Model for Human Vision Decoding. In *arXiv*.

Chen, Z.; Qing, J.; and Zhou, J. H. 2023. Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity. *arXiv preprint arXiv:2305.11675*.

de Zwart, J. A.; van Gelderen, P.; Jansma, J. M.; Fukunaga, M.; Bianciardi, M.; and Duyn, J. H. 2009. Hemodynamic nonlinearities affect BOLD fMRI response timing and amplitude. *Neuroimage*, 47(4): 1649–1658.

Fang, T.; Qi, Y.; and Pan, G. 2021. Reconstructing Perceptive Images from Brain Activity by Shape-Semantic GAN. *ArXiv*, abs/2101.12083.

Ge, S.; Nah, S.; Liu, G.; Poon, T.; Tao, A.; Catanzaro, B.; Jacobs, D.; Huang, J.-B.; Liu, M.-Y.; and Balaji, Y. 2023. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22930–22941.

Glover, G. H. 2011. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics*, 22(2): 133–139.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*.

Hénaff, O. J.; Bai, Y.; Charlton, J. A.; Nauhaus, I.; Simoncelli, E. P.; and Goris, R. L. 2021. Primary visual cortex straightens natural video trajectories. *Nature communications*, 12(1): 5982.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video Diffusion Models. arXiv:2204.03458.

Kanth, S. T.; and Ray, S. 2020. Electrocorticogram (ECoG) is highly informative in primate visual cortex. *Journal of Neuroscience*, 40(12): 2430–2444.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kupershmidt, G.; Beliy, R.; Gaziv, G.; and Irani, M. 2022. A Penny for Your (visual) Thoughts: Self-Supervised Reconstruction of Natural Movies from Brain Activity. *arXiv preprint arXiv:2206.03544*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, M.; Qu, T.; Yao, R.; Sun, W.; and Moens, M.-F. 2024a. Alleviating Exposure Bias in Diffusion Models through Sampling with Shifted Time Steps. *International Conference on Learning Representations*.

Li, M.; Wan, B.; Moens, M.-F.; and Tuytelaars, T. 2024b. Animate Your Motion: Turning Still Images into Dynamic Videos. *European Conference on Computer Vision*.

Lin, S.; Sprague, T.; and Singh, A. K. 2022. Mind Reader: Reconstructing complex images from brain activities. *arXiv preprint arXiv:2210.01769*.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2022. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022a. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022b. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

Mozafari, M.; Reddy, L.; and van Rullen, R. 2020. Reconstructing Natural Scenes from fMRI Patterns using BigBiGAN. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Ning, M.; Li, M.; Su, J.; Salah, A. A.; and Ertugrul, I. O. 2024. Elucidating the Exposure Bias in Diffusion Models. *International Conference on Learning Representations*.

Ning, M.; Sangineto, E.; Porrello, A.; Calderara, S.; and Cucchiara, R. 2023. Input Perturbation Reduces Exposure Bias in Diffusion Models. *arXiv preprint arXiv:2301.11706*.

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

Parrish, T. B.; Gitelman, D. R.; LaBar, K. S.; and Mesulam, M.-M. 2000. Impact of signal-to-noise on functional MRI. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 44(6): 925–932.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Qian, X.; Wang, Y.; Fu, Y.; Xue, X.; and Feng, J. 2023. Semantic Neural Decoding via Cross-Modal Generation. *arXiv preprint arXiv:2303.14730*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021a. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Shen, G.; Dwivedi, K.; Majima, K.; Horikawa, T.; and Kamitani, Y. 2019. End-to-end deep image reconstruction from human brain activity. *Frontiers in computational neuroscience*, 13: 21.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *arXiv:2010.02502*.

Sun, J.; Li, M.; and Moens, M.-F. 2023. Decoding Realistic Images from Brain Activity with Contrastive Self-supervision and Latent Diffusion. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023), 2023*.

Sun, J.; Li, M.; Zhang, Y.; Moens, M.-F.; Chen, Z.; and Wang, S. 2023. Contrast, Attend and Diffuse to Decode High-Resolution Images from Brain Activities. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sun, J.; Wang, S.; Zhang, J.; and Zong, C. 2019. Towards Sentence-Level Brain Decoding with Distributed Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 7047–7054.

Sun, J.; Wang, S.; Zhang, J.; and Zong, C. 2020. Distill and Replay for Continual Language Learning. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 3569–3579. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Sun, J.; Wang, S.; Zhang, J.; and Zong, C. 2021. Neural Encoding and Decoding With Distributed Sentence Representations. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2): 589–603.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093.

Uğurbil, K.; Xu, J.; Auerbach, E. J.; Moeller, S.; Vu, A. T.; Duarte-Carvajalino, J. M.; Lenglet, C.; Wu, X.; Schmitter, S.; Van de Moortele, P. F.; et al. 2013. Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *Neuroimage*, 80: 80–104.

Van Essen, D. C.; Smith, S. M.; Barch, D. M.; Behrens, T. E.; Yacoub, E.; Ugurbil, K.; Consortium, W.-M. H.; et al. 2013. The WU-Minn human connectome project: an overview. *Neuroimage*, 80: 62–79.

Varela, F. J.; Thompson, E.; and Rosch, E. 2017. *The embodied mind, revised edition: Cognitive science and human experience*. MIT press.

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

Wang, S.; Sun, J.; Zhang, Y.; Lin, N.; Moens, M.-F.; and Zong, C. 2024. Computational Models to Study Language Processing in the Human Brain: A Survey. arXiv:2403.13368.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

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

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7623–7633.

Zhao, X.; Sun, J.; Wang, S.; Ye, J.; Zhang, X.; and Zong, C. 2024. MapGuide: A Simple yet Effective Method to Reconstruct Continuous Language from Brain Activities. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3822–3832. Mexico City, Mexico: Association for Computational Linguistics.

Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; and Feng, J. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.