# High-Fidelity Movie Reconstruction based on the fMRI decoding of Hierarchical Brain Activity

**Myeonggyo Jeong (jmk000817@g.skku.edu)**
Department of Biomedical Engineering, Sungkyunkwan University, Suwon, South Korea

**Seok-Jun Hong (hong.seok.jun@gmail.com)**
Department of Biomedical Engineering, Sungkyunkwan University, Suwon, South Korea
Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, South Korea
Department of Intelligent Precision Healthcare Convergence, Sungkyunkwan University, Suwon, South Korea
Department of MetaBioHealth, Sungkyunkwan University, Suwon, South Korea
Center for the Developing Brain, Child Mind Institute, New York, NY, United States

**Abstract:**
**Recent advances in AI-neuroscience have revolutionized our ability to decode visual images from human brain activities. Yet, reconstructing animated scenes, such as movies, remains a challenging task due to their intricate spatiotemporal dynamics. Here, we introduce a novel fMRI-based movie encoding-decoding framework, using three major self-supervised learning algorithms, that is, VideoMAE, CLIP, and Latent Diffusion Model. These algorithms, along with a simple addition to the Diffusion model, "TEmporally Smoothed LAtent Representation" (TESLAR), enabled to reconstruct the scene with photo-realistic details and enhanced temporal consistency, collectively leading to a semantically richer and natural decoding process. This result was further enhanced by our thorough investigation of brain encoding, which informed the decoding process about which brain areas have the most relevant fMRI signals to reconstruct the set of visual features. Our framework has a high potential to reveal key representational mechanisms underlying complex perceptual processes in the human brain.**
**Keywords: Brain Encoding, Decoding; Generative AI**

## Introduction

Over the last several decades, the field of cognitive neuroscience has made significant strides in revealing the principles of how our brain represents various sensory stimuli of the external world (=encoding). This effort found that the brain encodes each incoming visual scene with distinct neural activities across widely distributed cortices, which collectively form large-scale hierarchical representation (Huth et al. 2016). Notably, recent algorithms in generative AI have been employed to leverage such cortex-wide signals from fMRI to reverse-engineer and reconstruct the visual images presented during the brain scan (=decoding). Indeed, newly proposed decoding approaches based on a "Diffusion" model, a generative algorithm inspired by non-equilibrium thermodynamics, demonstrated ultra-high resolution with great visual details in their image reconstruction (Takagi and Nishimoto 2023; Ozcelik and VanRullen 2023). Yet most of them are still at the level of decoding into a "static" image, not actual dynamics or animated scenes as observed in our real life. This can be suboptimal, since external sensory streams usually contain rich temporal contexts, which may provide an important clue for statistical regularity of the dynamical event occurring in the nature.

To fill this gap, here we propose an fMRI-based movie encoding-decoding framework, which consists of largely 3 self-supervised models (**Fig1**): **A)** VideoMAE (masked autoencoder for video; Tong et al. 2022) for initial coarse-level fMRI encoding-decoding, **B)** CLIP (Contrastive Language-Image Pre-training; (Radford et al. 2021) for the next, higher-order auxiliary encoding-decoding, which jointly embeds image-text pretraining data and finally **C)** Latent Diffusion Model to reconstruct fine-resolution animated scenes using the decoding results from the previous generative models.

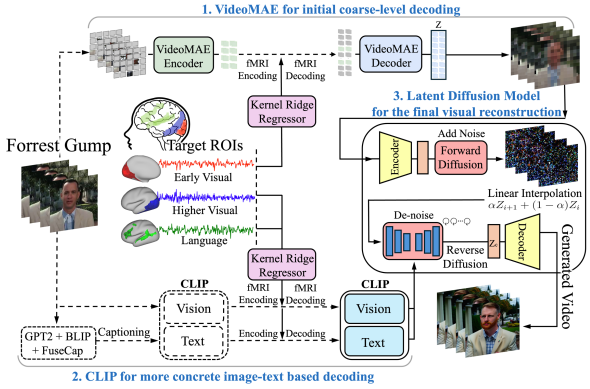


Figure 1. Model overview

Our contributions are two-folds:
1) To preserve temporal consistency across the video frames reconstructed, we incorporated two components into our pipeline: *i)* VideoMAE to efficiently learn (low-level) visual dynamic latents from animated frames and *ii)* TEmporally Smoothed LAtent Representation (TESLAR) in the reverse diffusion process (see **Method** for details), which allows to avoid a sudden change across reconstructed frames.

2) Before decoding, we first examined the whole brain encoding for various visual scenes in the training movie segment. This allowed to unbiasedly select the ROIs with high encoding accuracy, from which we extracted fMRI timeseries for the input to the next decoding process. This encoding-based ROI selection enhanced interpretability of our framework to better understand a representational mechanism of the human brain.

## Method

**Dataset.** We analyzed 'StudyForrest' (Hanke et al. 2016), 2 hours of opensource fMRI data acquired during movie watching. (3T, 8 runs, each 15 mins). From this dataset, we randomly chose 5 subjects (sub1-5) and used their 7 runs for training (3150 paired movie-fMRI data) and a remaining run for test (450 paired data).

**Multi-Level Encoding-Decoding Process.** Our framework was inspired by hierarchical visual-semantic processing in the brain. In the 1st VideoMAE step, we aimed at learning representation of a low-level visual latent of consecutive video frames to reconstruct a coarse-resolution but still informative initial input to the next, more fine-scale decoding process. In the 2nd CLIP step, we utilized higher-order visual-language features in the movie, to further enhance the decoding result with enriched semantic contexts. Reflecting this motivation, our encoding test based on the training data showed

the highest encoding accuracy (Fig2) *i)* in the low-level visual cortices (*e.g.* V3A, V4) for VideoMAE, *ii)* in higher-order extrastriate visual areas (*e.g.* V3 c/d, V8) for CLIP-vision and *iii)* in visual-language association areas (*e.g.* PGp, TPOJ) for CLIP-text modules. We took these areas as input ROIs for which fMRI data were fed into the following decoding pipeline.

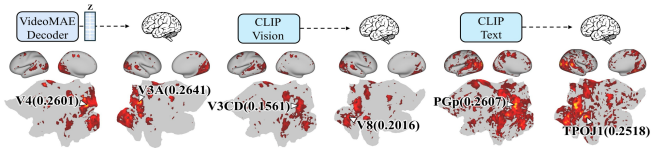


Figure 2. Encoding performance for each generative model

**1) Low-level encoding-decoding.** VideoMAE, a self-supervised algorithm based on masked autoencoder, is designed to capture contextual representation of video frames. To this end, VideoMAE masks out 95% of areas in the scene and learns to predict these masked aeras based on the 5% remaining, unmasked regions (**Fig1A**). More specifically, the encoder of VideoMAE (Vision Transformer; ViT) first extracted the latent features from the masked video segments. We then trained a kernel-ridge regression between the ViT latents and fMRI from early visual-cortex ROIs (=6384 voxels). Using this trained model, we predicted the ViT latents for the test movie. These were in turn fed into the decoder part to reconstruct the scene in a coarse-level resolution (16 frames per 1TR [2 secs]), which captured abstract layouts and color features of the scene.

**2) High-Level encoding-decoding.** Next, to learn higher-level semantic features, we employed CLIP, a contrastive learning algorithm to jointly learn a shared information between image and text data. We trained two different kernel-ridge regressions for decoding (**Fig1B**): the one for between CLIP-vision features and fMRI from high-order extrastriate visual cortices (=8151 voxels), and another for between CLIP-text and fMRI from the visual-language areas (=5024 voxels).

**3) Diffusion-based video reconstruction.** We fed those previous decoding results to a final latent diffusion model (as conditioning) to shape the reconstruction of the video frames. We employed Versatile Diffusion, a recently developed multimodal Diffusion algorithm, and modified its reverse diffusion such that it temporally smooths the latent representation ('TESLAR'; a linear interpolation of Gaussian noises between the previous and next frames (**Fig1C**). This resulted in a smooth transition across the reconstructed frames, providing both visual naturalness and semantic richness.

## Results

**Fig3** shows examples of our decoding result, *i)* from one subject across different scenes (left) and *ii)* from five subjects for a randomly-chosen, single scene (right).


| **Bold**: significance (p<0.0001) | Frame-based | | | | Video-based | |
|---|---|---|---|---|---|---|
| | Pixel level | | Semantic (N-way, top-1classification) | | Semantic (N-way, top-1 classification) | |
| | SSIM↑ | MSE↓ | 2-way↑ | 50-way↑ | 2-way↑ | 50-way↑ |
| Full Model | 0.2701 | 98.97 | **0.8612**±0.02 | **0.1981**±0.04 | **0.8376**±0.01 | **0.2002**±0.03 |
| VideoMAE | **0.6356** | **66.97** | 0.7519±0.05 | 0.0815±0.02 | 0.8157±0.02 | 0.0994±0.03 |

Figure 3. fMRI-based brain decoding result for the subject 1

To evaluate the accuracy, we employed two different levels of metrics, one for pixel-level (Structural similarity index measure [SSIM] and Mean Squared Error [MSE]) and another for semantic-level (using ViT-based N-way, top-1 classification with 100 trials). At a pixel level, VideoMAE outperformed the reconstruction of Versatile Diffusion (with TESLAR; full model), underscoring its utility of capturing low-level primary visual features. At a semantic level, Versatile Diffusion excels VideoMAE in both frame- and video-based accuracies, indicating the advantage of CLIP-vision and -text components in identifying visuo-semantic features from the video data. Collectively, these accuracies are suggestive of neural correlates underlying hierarchical information process in the human brain.

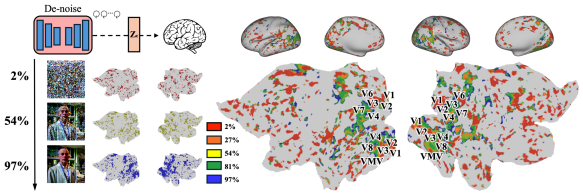


Figure 4. Encoding of the Diffusion denoising process

Finally, we were also interested if there was any brain correspondence to the details of generative process in the Diffusion model. To answer this, we performed an encoding analysis along the denoising steps. As in **Fig4**, discernible brain activation patterns were observed as the step increased, showing that they gradually diffuse from early visual cortices up to high-order areas. This finding suggests that in order to generate (or imagine) a dynamical scene with greater visual details, our brain requires increasingly widespread cortical allocation, especially for multiple transmodal systems.

## Conclusion

The proposed movie encoding-decoding framework has a high potential in developing a more naturalistic mind-reading technique, which may be helpful to unveil atypical perceptual process in psychiatric disorders.

## References


Hanke, Michael, Nico Adelhöfer, Daniel Kottke, Vittorio Iacovella, Ayan Sengupta, Falko R. Kaule, Roland Nigbur, Alexander Q. Waite, Florian J. Baumgartner, and Jörg Stadler. 2016. "Simultaneous FMRI and Eye Gaze Recordings during Prolonged Natural Stimulation - a Studyforrest Extension." *BioRxiv*. bioRxiv. https://doi.org/10.1101/046581.

Huth, Alexander G., Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. "Natural Speech Reveals the Semantic Maps That Tile Human Cerebral Cortex." *Nature* 532 (7600): 453–58.

Ozcelik, Furkan, and Rufin VanRullen. 2023. "Natural Scene Reconstruction from FMRI Signals Using Generative Latent Diffusion." *Scientific Reports* 13 (1): 15666.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. "Learning Transferable Visual Models from Natural Language Supervision." *ArXiv [Cs.CV]*. arXiv. http://arxiv.org/abs/2103.00020.

Takagi, Yu, and Shinji Nishimoto. 2023. "High-Resolution Image Reconstruction with Latent Diffusion Models from Human Brain Activity." In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr52729.2023.01389.

Tong, Zhan, Yibing Song, Jue Wang, and Limin Wang. 2022. "VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training." *ArXiv [Cs.CV]*. arXiv. http://arxiv.org/abs/2203.12602.