Cross-Subject Brain Decoding for Naturalistic Movie Reconstruction

Myeonggyo Jeong (jmkkorea0817@gmail.com)
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, NY, United States

Abstract:
Understanding how the brain processes dynamic visual stimuli remains a central challenge in neuroscience. Although recent AI-based methods have succeeded in reconstructing static images from fMRI data, decoding continuous movie scenes entails another complexity layer to navigate spatiotemporal brain activities that are distinctly represented across different individuals. Here, we propose a multi-subject fMRI decoding framework to combine inter-subject functional alignment—which uncovers shared neural representations among participants—with subject-specific tokens to tag idiosyncrasy of an individual's functional dynamics in learning fMRI representation. By integrating these two complementary techniques, our method simultaneously achieves robust cross-subject generalization and person-optimized modeling, requiring only minimal fine-tuning. Moreover, we also employed a whole-brain Transformer to link fMRI signals to the CLIP image-text embeddings, preparing enriched brain-video mapping input representation for a subsequent video generation. Finally, we employed AnimateDiff and FreeInit, the two up-to-date algorithms to maximize temporal coherency across reconstructed frames. Advancing fMRI movie decoding techniques holds a promise to develop a quantitative mean to scrutinize brain dynamics underlying naturalistic visual experiences.
Keywords: fMRI decoding; movie reconstruction; shared response model; cross-subject generalization

Introduction

The human brain processes visual information through distributed neural activity across multiple cortical areas, forming hierarchical representations of both low-level visual features and high-level semantic content (Huth et al., 2016). Decoding these complex neural signals to reconstruct the original stimuli has recently become a more realistic goal in cognitive neuroscience, thanks to the rapidly advanced AI fields. Indeed, recent self-supervised learning techniques such as latent Diffusion models has enabled significant progress in fMRI-based visual reconstruction, generating high-quality images purely from the brain activity (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023).

Yet, the current approaches face two key issues to address: i) so far the test has been conducted only for 'static' image reconstruction, while our biological brain processes nearly all the time spatiotemporally active stimuli (e.g., animated scenes) and ii) most of these algorithms require extensive subject-specific training data. This (second) issue is particularly critical in terms of generalizability, as the methods trained only on one person typically perform poorly when applied to others—necessitating costly and time-consuming data collection for each new subject.

Here, we provide a cross-subject fMRI-based movie decoding framework to tackle inter-subject variability through FastSRM (Richard et al., 2019), a technique to align individual fMRI responses across participants into a common representational space while preserving stimulus-driven neural dynamics. We also implemented a whole-brain Transformer to boost a reconstruction accuracy, adding a subject-specific token to tag a remaining individual idiosyncrasy during the algorithm training. By extracting both group-common and personalized neural representation, our method aimed at a high accuracy of individualized brain decoding, while making it more practical for real-life application.

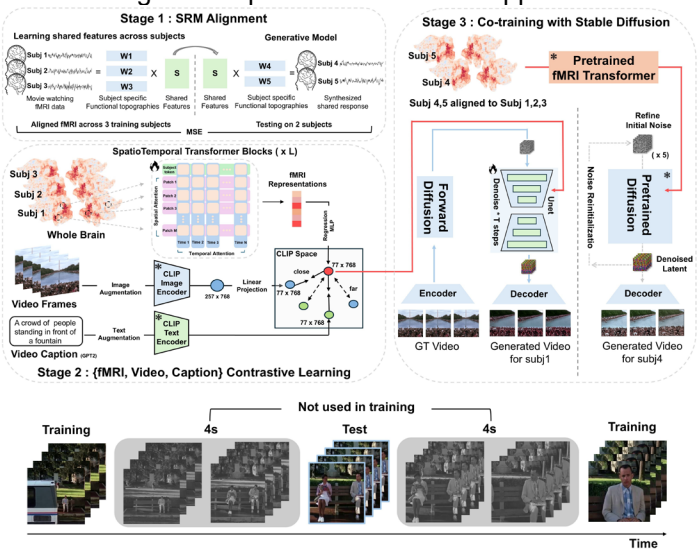

Figure 1. Model overview. Three-stage framework: SRM alignment, contrastive learning with Transformer, and video generation. Bottom: train/test data splitting strategy.

Method

fMRI Preprocessing We analyzed the StudyForrest data (Hanke et al., 2016), which includes the 2-hours of fMRI from subjects watching natural movies ("Forrest Gump", 3T scanner, 8 runs of ~15 minutes each). From the available 15 participants, we selected five (sub1–5) for analysis due to computational constraints, with plans to extend this approach to the full dataset in future work.

To address inter-individual variability, we implemented Fast Shared Response Modeling (FastSRM) to align subject-specific fMRI time series into a shared representational space. This alignment identifies common stimulus-driven neural representations while preserving subject-specific idiosyncrasies that are later captured by our subject specific tokens. We trained FastSRM on three participants (sub 1-3) and applied this to two new cases (sub 4-5). Post-alignment analysis showed significant increase of inter-subject correlation (ISC; particularly in the visual and attention networks; **Figure 2**), demonstrating its validity.

For evaluation, we divided each data into four segments (Run12, Run34, Run56, Run78), each comprising approximately 30 minutes. We selected

every 10th clip for testing and excluded two clips before and after that (testing) clip to avoid an overfitting issue. This setup ensures fair performance evaluation while preserving the richness of the original movie stimulus.

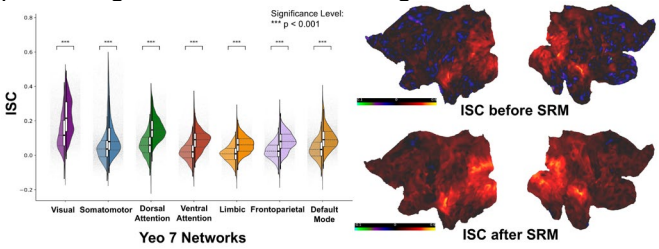

Figure 2. FastSRM Results: (Left) ISC distributions across Yeo 7 networks before and after SRM between sub 4 and 5 (Right) Brain maps showing increased ISC after SRM

Movie Decoding Pipeline

Our decoding pipeline integrates three components to transform fMRI signals into video reconstructions:

1. fMRI Transformer Architecture: We implemented a Transformer model that processes whole-brain activity (59,412 voxels) divided into 928 patches. With 1024-dimensional embeddings, 8 attention heads, and 12 layers, this whole-brain approach learns spatiotemporal fMRI signals without relying on predefined brain regions (see the next paragraph for the training details). Our approach maximize the data utility by balancing the learning of both shared and individual neural patterns: FastSRM provides cross-subject alignment of stimulus-driven responses, while subject-specific tokens capture unique neural signatures that persist after alignment.

2. Contrastive Learning Framework: We trained the Transformer through a contrastive learning to establish a relationship between fMRI signals and CLIP's visual-semantic representations. Specifically, the objective function minimizes distances between brain activity patterns and their corresponding CLIP representations and pulled relevant embeddings closer, while pushing unrelated ones apart. This approach enables the model to capture both perceptual features and semantic content from brain activity. For generalizability, we trained this framework on subjects 1-3, and applied lightweight fine-tuning to subjects 4-5.

3. Video Reconstruction: We combined two recent video-processing algorithms, namely AnimateDiff (Guo et al., 2023), a Stable Diffusion-based model, and FreeInit (Wu et al., 2023) to enhance a temporal stability. Notably, FreeInit's noise initialization ensures coherent transitions between frames and reduces flickering artifacts. We applied five FreeInit iterations during inference to produce 2-second videos at 6fps, successfully preserving both structural details and high-level semantics from the original movie scenes. This approach allows us to generate videos that maintain consistent visual flow and temporal coherence throughout the movie sequences.

Results

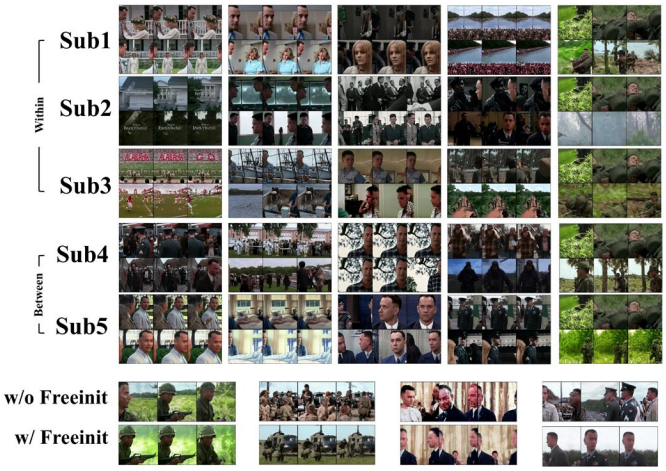

Figure 3. Decoding results and the effect of FreeInit.

We visualized the decoded movie frames across five participants (**Figure 3**). Sub1-3 showcase within-subject reconstructions, while Sub4-5 reveal to some extend a generalization to the new brains after fine-tuning. The reconstructions capture cinematic elements, highlighting its ability to extract visual narratives from diverse neural patterns. Notably, FreeInit enhanced a temporal flow, eliminating jarring transitions for more naturally flowing visual sequences.

average scores across all runs	Video-based		Frame-based			
	Semantic-level : 50-way		Semantic-level : 50-way		Structure-level (SSIM)	
	Shuffled	Non-Shuffled	Shuffled	Non-Shuffled	Shuffled	Non-Shuffled
Subject 1	0.0697	0.1544	0.1278	0.2449	0.1954	0.2017
Subject 2	0.0782	0.2096	0.1267	0.2292	0.1858	0.1996
Subject 3	0.0671	0.1896	0.1306	0.2477	0.1945	0.2095
Subject 4	0.0561	0.2668	0.1340	0.2459	0.2072	0.2163
Subject 5	0.0769	0.2042	0.1292	0.2408	0.2182	0.2288

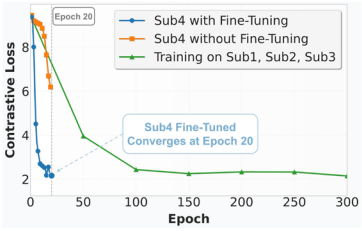

Figure 4. Quantitative evaluation (values averaged across all runs) (left) and Training Efficiency (right)

To quantitatively evaluate our method, we compared normal versus shuffled brain-movie pairings (left). Non-shuffled conditions consistently outperformed shuffled baselines, confirming our meaningful reconstruction of brain-stimulus relationships. Fine-tuned subjects (4-5) performed comparably to training subjects (1-3). The **Figure 4** (right) shows efficiency of our framework: fine-tuning on sub4 converges in just 20 epochs (blue line), while training from scratch requires significantly more iterations (~300 epochs, green line), and no fine-tuning on sub 4 yields poor results (orange line). These results confirm that our method enables effective cross-subject generalization with minimal additional training.

Conclusion

Here we propose a multi-subject fMRI movie decoding framework that enables semantically and temporally consistent video reconstruction. By combining FastSRM alignment, Transformers, and diffusion models, our method showcase a potential to generalize for new subjects, with minimal fine-tuning.

References

Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., & Dai, B. (2023). AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. arXiv preprint arXiv:2307.04725. https://doi.org/10.48550/arXiv.2307.04725

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F.R., Nigbur, R., Pohlack, S.T., Waite, F., Bitzer, S., & Stadler, J. (2016). A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. Scientific Data, 3(1), 1-15. https://doi.org/10.1038/sdata.2016.92

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., & Gallant, J.L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453-458. https://doi.org/10.1038/nature17637

Ozcelik, K., & VanRullen, R. (2023). Brain decoding with a transformer-based diffusion model. Nature Human Behaviour, 7, 1877-1891. https://doi.org/10.1038/s41562-023-01766-w

Richard, H., Martin, L., Pinho, A.L., Pillow, J., & Thirion, B. (2019). Fast shared response model for fMRI data. arXiv preprint arXiv:1909.12537. https://doi.org/10.48550/arXiv.1909.12537

Takagi, Y., & Nishimoto, S. (2023). 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, 12854-12864. https://doi.org/10.1109/CVPR52729.2023.01248

Wu, T., Si, C., Jiang, Y., Huang, Z., & Liu, Z. (2023). FreeInit: Bridging Initialization Gap in Video Diffusion Models. arXiv preprint arXiv:2312.07537. https://doi.org/10.48550/arXiv.2312.07537