Modeling Brain Activity During Naturalistic Movie Watching Using Video Embeddings and fMRI

Hardik Mittal, Gaurav Bhole

Abstract

This study explores the neural representation of complex naturalistic stimuli by developing encoding and decoding models that link video content with brain activity captured using functional magnetic resonance imaging (fMRI). Participants viewed four diverse short films while undergoing fMRI scanning. Using advanced video embedding models such as ViViT, ViTMAE, and VideoMAE, we extracted high-level visual features from the movie stimuli. We then trained machine learning models to predict voxelwise fMRI activity from video embeddings (encoding) and vice versa (decoding). By evaluating the performance across different brain voxel selections, embedding strategies, and model architectures, we identified trends in brain-video alignment. Additionally, we implemented classification models to infer which movie a participant was watching based on their brain activity. Our work highlights the potential and current limitations of computational models in decoding naturalistic cognition.

1 Introduction

Understanding how the human brain processes dynamic, naturalistic experiences like watching films is a challenging frontier in cognitive neuroscience. Traditional approaches, which often rely on simple stimuli and static tasks, fall short of capturing the complexity of real-world perception. Recent progress in machine learning, particularly in the domain of computer vision, has led to the development of deep video encoders capable of capturing rich spatiotemporal representations of videos. Concurrently, techniques for aligning brain activity across subjects, such as the Shared Response Model (SRM), provide a means to aggregate and analyze group-level neural responses to the same stimulus.

In this project, we explore whether high-dimensional video embeddings can predict patterns of brain activity across time, and conversely, whether brain activity can be used to reconstruct aspects of the viewed content. We also investigate the extent to which these models generalize across individuals, ultimately aiming to classify viewed movies from brain data.

2 Dataset and Stimuli

The experiment involved fMRI data collection from 54 participants while they watched four different short films. The dataset used was a preprocessed version with missing or corrupted participant data excluded, and included SRM-aligned fMRI data ("noncensor_srm-recon") to enable cross-subject analysis.

- **Iteration (12:27)**: Sci-fi narrative involving repeated escape attempts.
- **Defeat (7:57)**: A young girl uses time travel to confront a bullying brother.
- **Growth (8:27)**: A family drama tracking children growing up.
- **Lemonade (7:27)**: A Rube-Goldberg machine edited to remove human appearances.

Each movie was segmented into 1-second intervals to align with the fMRI TR of 1 second.

3 Methods

We began with cleaned, SRM-aligned fMRI data and removed non-informative timepoints as directed by the dataset authors. To focus on the most responsive brain areas, we selected the top 1000, 5000, and 10000 voxels per participant based on their variance across the entire movie duration. High variance across time was assumed to reflect engagement with the stimuli.

To account for the hemodynamic lag in fMRI signals, we convolved the data with a canonical hemodynamic response function (HRF) delayed by 4 seconds. All data were normalized using Z-score normalization unless otherwise specified in experiments.

3.1 Video Embedding Extraction

We used three state-of-the-art video encoding models:

- **ViViT** : A transformer-based model designed for spatiotemporal video representation.
- **ViTMAE** : A masked autoencoder variant adapted for video inputs.
- **VideoMAE** : A masked autoencoder specialized for capturing temporal coherence.

Videos were subsampled at 8 frames per second. For each second of video, a single embedding vector was extracted. These embeddings were then normalized to standardize across feature scales.

3.2 Models

To perform both encoding (video $\rightarrow$ brain) and decoding (brain $\rightarrow$ video), we trained two types of models:

- A deep multilayer perceptron (MLP), consisting of multiple fully connected layers.
- A recurrent LSTM model, which accounts for temporal dependencies.

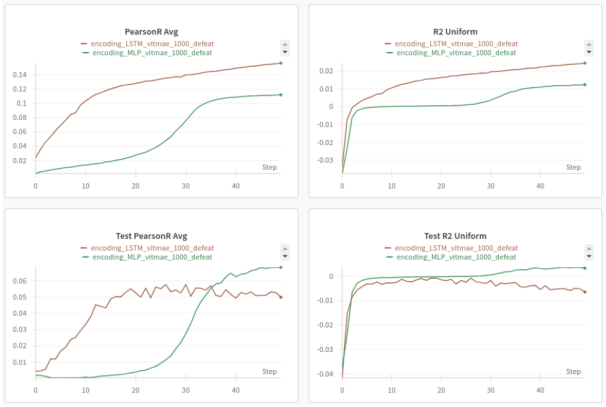
These models were trained using mean squared error loss and evaluated using Pearson correlation and $\mathbf{R^2}$ scores.

3.3 Evaluation Metrics

Performance was assessed using the average Pearson correlation coefficient across voxels or embedding features. Additionally, R^2 scores were computed in both uniform average and variance-weighted formats. A detailed Pearson correlation function ensured robustness to low-variance features.

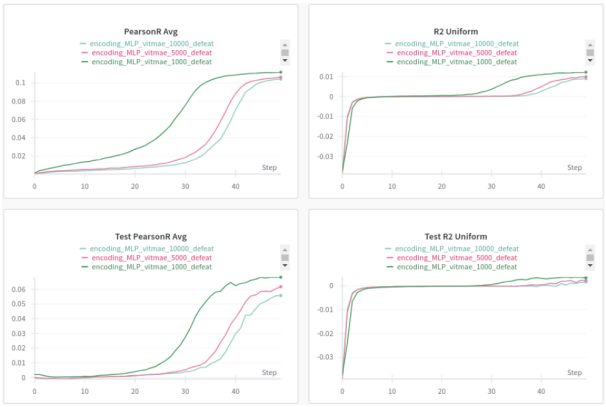
4 Encoding Experiments

4.1 Model Architectures



The comparison between MLP and LSTM architectures revealed consistent patterns. The LSTM showed steeper improvements in both Pearson correlation and R^2 during training; however, this came at the cost of overfitting. The training curves for LSTM showed a continuous rise, while the test performance plateaued and even declined slightly, especially in terms of R^2. Conversely, the MLP showed more gradual improvement but achieved higher test set performance, confirming its robustness and generalizability.

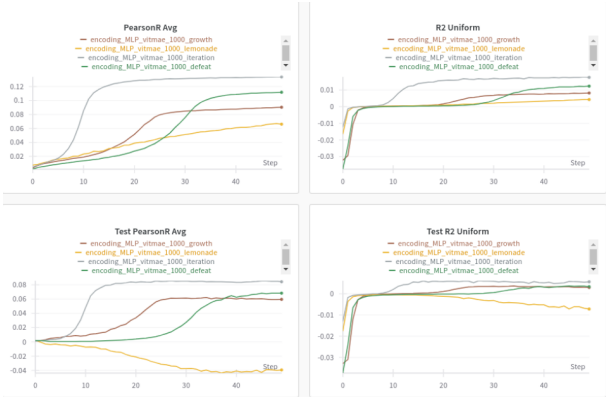
4.2 Voxel Selection



We tested voxel subsets of 1000, 5000, and 10000 voxels. Results showed that the MLP performed best using the 1000 most variant voxels, as both Pearson and R^2 values reached their peak with this subset. Interestingly, while training metrics continued to improve with larger voxel counts, test metrics slightly decreased, highlighting potential

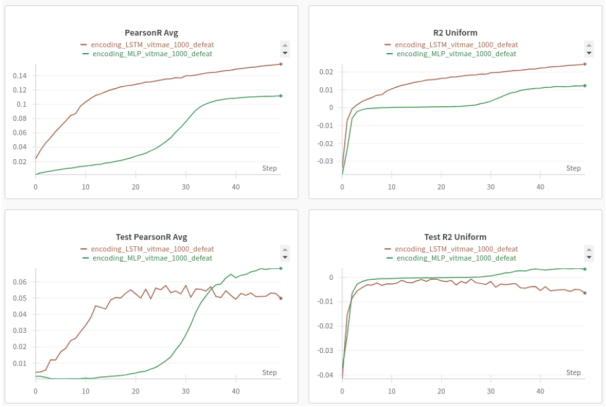
overfitting or inclusion of noisy voxels. The LSTM was less sensitive to voxel count but showed diminishing returns beyond 5000 voxels.

4.3 Movie-wise Performance



Movie-specific analysis showed that *Iteration* performed the best across all metrics, likely due to its longer runtime and rich temporal structure. *Growth* and *Defeat* followed, while Lemonade performed the worst, even achieving negative R^2 scores on the test set. This suggests that human-centric and narratively rich content elicits more consistent, predictable brain responses compared to mechanical, low-variance visual stimuli.

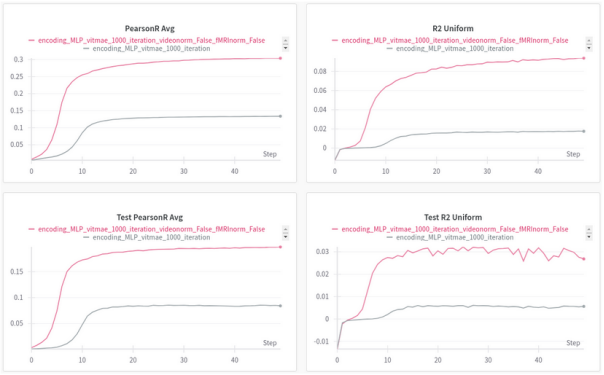
4.4 Embedding Comparison



Although training performance was nearly identical across ViViT, ViTMAE, and VideoMAE, test set evaluations revealed slight variations. ViTMAE tended to generalize slightly better for certain movies (notably Defeat), but overall, all three embedding types performed comparably, supporting the robustness of deep video features for fMRI modeling.

4.5 Effect of Normalization

Normalization experiments yielded the most surprising insight. Removing Z-score normalization for both the video embeddings and fMRI voxels led to a dramatic improvement in encoding performance. For example, in Defeat and Iteration, test Pearson correlation rose from 0.06 to 0.15, and test R^2 improved by more than 0.03. This indicates that

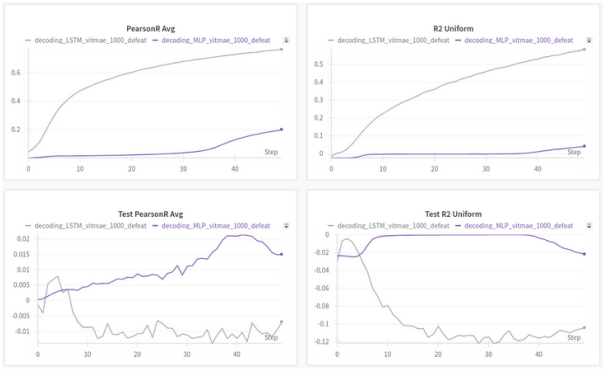


absolute signal dynamics, not just relative trends, may hold key information relevant for fMRI encoding tasks.

These findings validate our selection of the MLP model and top 1000 voxels for subsequent experiments and suggest that simpler preprocessing steps may better preserve task-relevant signal variance.

5 Decoding Experiments

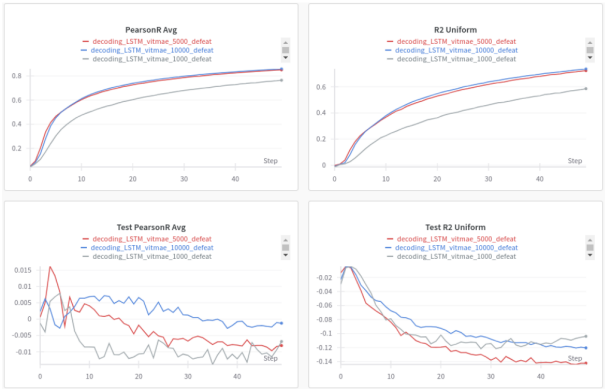
5.1 Model Architectures



The comparison between LSTM and MLP models for decoding yielded clear evidence of severe overfitting in the LSTM model. While the LSTM achieved extremely high training Pearson correlations and R^2 values (approaching 0.6–0.8 in some cases), its test performance was near or below zero. This discrepancy highlights a failure to generalize, making the LSTM unsuitable for decoding tasks in this dataset. In contrast, the MLP, although modest in training performance, consistently outperformed the LSTM on test data. This behavior reflects its more stable and regularized learning dynamic, making it more effective for decoding under data-limited conditions.

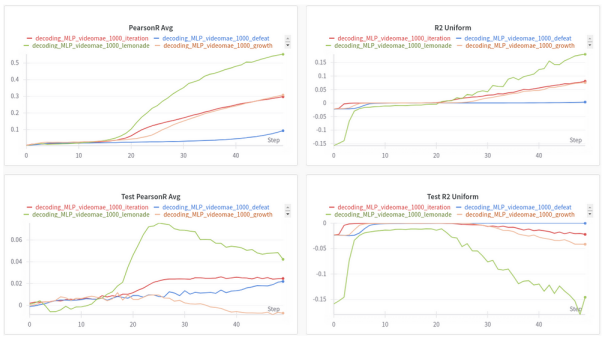
5.2 Voxel Selection

Decoding performance strongly depended on the number of voxels used. Increasing voxel counts led to better training performance across both models. The MLP saw a steady improvement in Pearson correlation and R^2 as the voxel count increased from 1000 to 10000. Interestingly, test set performance also improved slightly with more voxels, but the gains diminished quickly, and overfitting became evident beyond 5000–10000 voxels. The



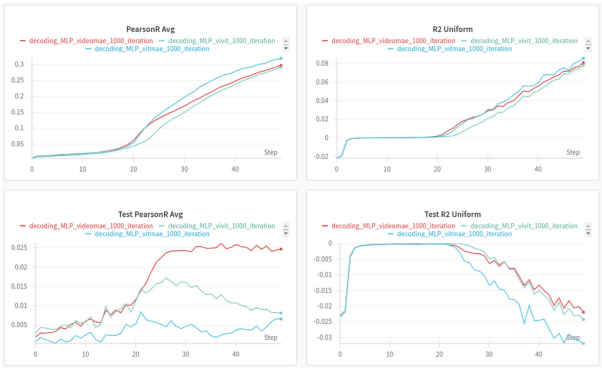
LSTM exhibited chaotic trends on the test set across all voxel settings, again highlighting its lack of generalization.

5.3 Movie-wise Performance



Contrary to the encoding findings, decoding models performed best on *Lemonade*. This movie, devoid of complex narrative and human interaction, may produce more predictable patterns in the visual cortex, hence being easier to decode from fMRI. Iteration and *Defeat* also performed reasonably, while Growth struggled, showing consistently low and even negative R^2 scores. These results suggest that decoding is more successful for simpler, less variable visual content.

5.4 Embedding Comparison



Decoding models built on **ViViT** and **VideoMAE** embeddings provided more stable test performance than those using **ViTMAE**, particularly on Iteration and Defeat.

Although training metrics were similar across all models, **VideoMAE** showed a slight edge in generalization. This pattern supports the idea that temporal continuity and spatial structure captured by **VideoMAE** might better align with neural representations, especially for simpler stimulus content.

5.5 Effect of Normalization

Removing normalization had mixed effects in decoding tasks. For *Defeat*, removing normalization led to a drastic improvement in both training and test Pearson scores (from 0.06 to 0.15), but test R^2 remained near zero. In contrast, for Iteration, normalization seemed beneficial or at least not harmful. This inconsistency across movies suggests that the impact of normalization may depend on the specific temporal and semantic dynamics of the stimulus, and that a more flexible, movie-specific preprocessing pipeline might yield better decoding results overall.

Taken together, these decoding experiments reinforce the conclusion that simple, robust architectures and careful voxel selection are critical. They also highlight the need for better regularization and possibly more advanced domain adaptation techniques to bridge the encoding-decoding generalization gap.

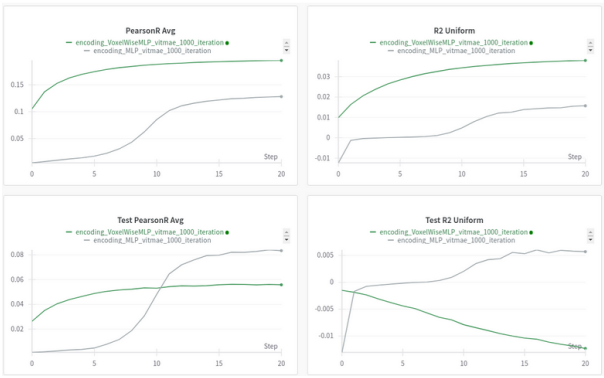
6 Voxelwise Encoding Model

To better capture voxel-specific relationships between visual stimuli and brain activity, we experimented with a voxelwise encoding model. Unlike the standard MLP that predicts all selected fMRI voxels jointly from video embeddings, this approach trains a separate MLP for each voxel, allowing each region to develop its own specialized representation of the stimulus features.

6.1 Model Architecture

Each voxel is assigned a small neural network with 2 hidden layers of 128 units, ReLU activations, and dropout (0.1). The model receives the video embedding for a given time point and outputs a single scalar — the predicted activity of one voxel. For a full fMRI prediction, all individual voxel models are applied in parallel and their outputs stacked.

This architecture enhances voxel-level specificity and allows us to investigate how different brain regions vary in their encoding of visual stimuli.



6.2 Experimental Setup

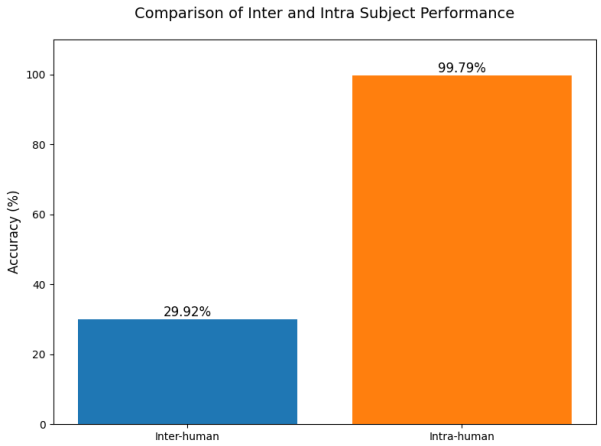
Iteration Movie. For the movie *Iteration*, the voxelwise encoding model achieved superior performance in training, outperforming the shared MLP across both Pearson correlation and R^2 metrics. On the test set, however, the shared MLP showed better generalization, with a peak Pearson correlation of approximately 0.075 vs. 0.055 for the voxelwise model, and a slight positive R^2 compared to a declining trend for the voxelwise model. This suggests the voxelwise MLP is more prone to overfitting, despite its per-voxel specialization.

Defeat Movie. In contrast, for the *Defeat* movie, the voxelwise model performed consistently better than the shared MLP across training and testing metrics. The test Pearson correlation reached over 0.04 compared to just 0.01 for the shared model, and R^2 remained positive throughout training. This implies that for stimuli with less structural repetition or noisier voxel response patterns, individual voxel models may offer better adaptation and robustness.

6.3 Discussion

These findings demonstrate that voxelwise modeling can outperform traditional global encoding methods under certain conditions, particularly when the signal-to-noise ratio varies significantly across voxels. However, the added model complexity and training overhead must be carefully managed. Additionally, generalization performance is not always improved, highlighting the need for stronger regularization or voxel selection heuristics in per-voxel modeling.

7 Classification



In addition to the encoding and decoding tasks, we also explored the ability to classify which movie a participant was watching based solely on their fMRI activity. This task was formulated as a four-way classification problem, corresponding to the four stimulus videos used in the study: *Iteration*, *Defeat*, *Growth*, and *Lemonade*.

7.1 Inter-subject

In this setting, the model was trained on data from a subset of participants and tested on data from unseen participants. This task yielded an accuracy of 29.92%, slightly above the 25% chance level for four classes. This low accuracy highlights the significant inter-subject variability in fMRI data and underscores the challenges of generalizing across individuals.

7.2 Intra-subject

In contrast, the intra-subject model was trained and tested on different temporal segments of data from the same participants. The model achieved an outstanding 99.79% accuracy. This indicates that individual neural signatures are highly consistent over time and can be effectively captured with the given architecture.

This stark contrast is meaningful. The intra-subject success reflects how brain responses are highly predictive of ongoing stimulus when tailored to an individual, while the low inter-subject accuracy underscores the extent of variability between participants' neural representations, even after SRM alignment. This result supports the idea that shared-response alignment (SRM) helps but does not fully normalize inter-subject differences for fine-grained temporal decoding.

8 Limitations

Our most significant limitation was computational. The training of advanced video models or large-scale transformer-based fMRI models was restricted due to lack of GPU resources. This constraint limited the depth and scale of our experiments, especially for more data-hungry models.

Additionally, while our use of SRM-aligned data helped with inter-subject consistency, variability across participants remained a major bottleneck for classification. The absence of statistical tests means that differences in performance between models should be interpreted cautiously.

9 Future Work

Future directions include using transformer-based models for decoding brain data and incorporating contrastive or self-supervised objectives to improve brain-embedding alignment. Exploring semantic features (e.g., emotion, objects, or scenes) rather than raw embeddings may enhance decoding fidelity. Improved subject alignment techniques such as hyperalignment may further help with inter-subject generalization.

With better computational resources, experiments involving longer videos, higher-resolution embeddings, and more participants would allow for deeper analysis of naturalistic brain responses.