Generative Decoders of Visual Cortical Representations

Yuchen Teng, Michael Zhou, Qingwen Liu, Yueran Wang,
Chengyan Li

Abstract

This project focuses on developing generative decoders to map brain activity during visual perception to perceived images. By leveraging the THINGS Ventral Stream Spiking Dataset (TVSD), which includes single-neuron recordings from macaques, the project aims to reconstruct high-resolution visual stimuli directly from neural activity. Unlike traditional fMRI-based approaches, this research uses multi-unit activity (MUA) data, providing superior temporal and spatial resolution. The project comprises two primary components: encoding and decoding. The encoding phase extracts image features using deep learning models such as AlexNet. In contrast, the decoding phase employs models like diffusion models and feature inversion to reconstruct images from MUA responses. This work bridges the gap between neural representations and visual perception, contributing to developing more robust brain-computer interfaces and enhancing our understanding of visual processing in the brain.

1 Introduction

1.1 Project Scope and Goal

This Capstone project is done in collaboration between the Data Science Institute and the Visual Inference Lab (advised by Prof. Nikolaus Kriegeskorte) at Columbia University. The goal is to build effective decoders that map from primate brain activity during visual perception to the perceived images. The study leverages single-neuron recordings via the THINGS Ventral stream Spiking Dataset (TVSD) to analyze and reconstruct visual representations in the brain. While prior work in the lab has focused on similar tasks using fMRI data, this project applies the same paradigm to single-neuron recordings for better temporal and spatial resolution.

This project consists of two components:

1. **Encoding** - Extract image features from deep learning models like AlexNet, then map them to multi-unit activity (MUA) responses using linear regression models, and use Pearson Correlation to decide layers selection.
2. **Decoding** - Use MUA responses and the features extracted from the original images to reconstruct images using diffusion models, feature inversion techniques, or learned mappings.

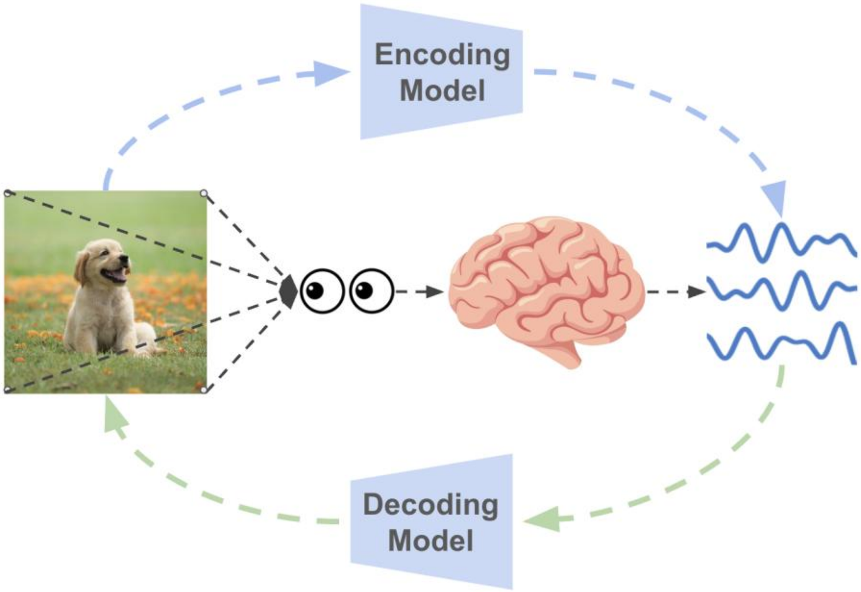

Figure 1. An overall diagram of our project workflow. Image credits: Pinyuan Feng, Columbia Psychology PhD

1.2 Existing Work

Generative vision models and language have experienced considerable growth in the past few years in functionality and user adoption. They also hold great potential for scientific discovery. Prior research has shown that CNNs and Vision Transformers (ViTs) can effectively encode visual features from images [1]. Linear regression models have also effectively mapped brain activity to high-dimensional embeddings [2]. Yet most of the data focuses on using fMRI data [5], and there have yet to be any studies that use this on single-neuron recordings. This project aims to use single-neuron recordings to bridge the gap, which could contribute to a better understanding of the brain, as well as the development of more robust brain-computer interfaces.

1.3 Why Visual Neuroscience Matters

The human visual system is highly efficient and adaptive, capable of fast, robust, and context-aware recognition. Unlike artificial vision systems, which struggle with varied lighting conditions, occlusions, and novel objects, biological vision seamlessly adapts to these challenges. The brain achieves this through a hierarchical processing mechanism, allowing for deep feature extraction, generalization, and attention mechanisms that enhance perception.

Figure 2 illustrates a comparison between biological and artificial vision systems. The human visual system processes images through the eye, capturing visual stimuli, and the brain, interpreting and assigning meaning to the scene. This enables a high-level understanding of objects within their context. In contrast, computer vision systems rely on sensing devices (cameras) and interpreting devices (computational models) to recognize objects. While artificial models can achieve object recognition, they lack the flexibility and adaptability of biological vision when faced with complex real-world conditions.

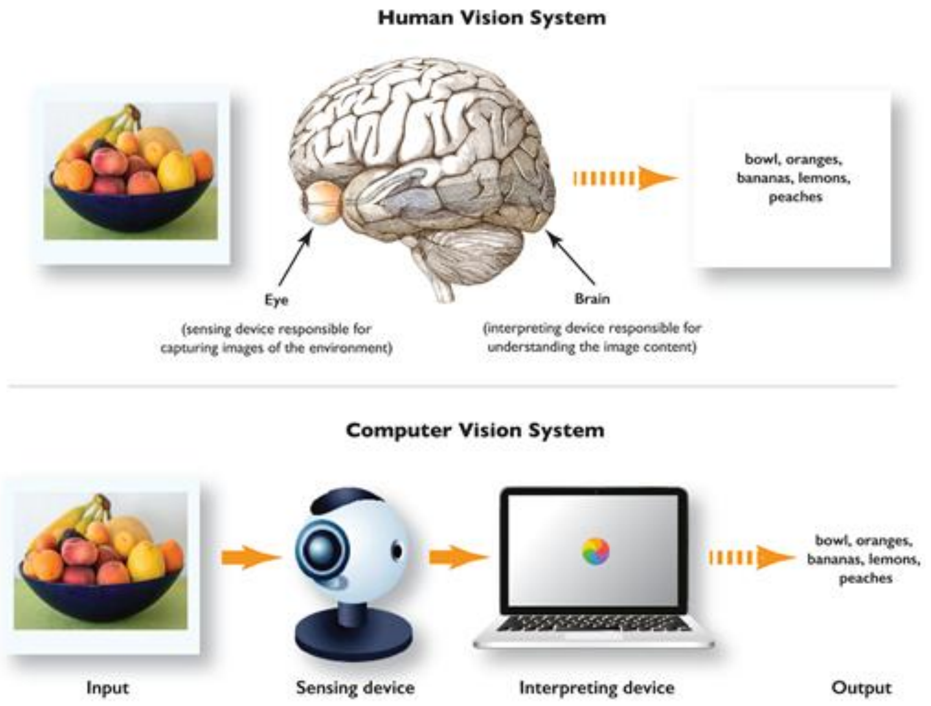
Human Vision System
bowl, oranges, bananas, lemons, peaches
Eye
(sensing device responsible for capturing images of the environment)
Brain
(interpreting device responsible for understanding the image content)
Computer Vision System
bowl, oranges, bananas, lemons, peaches
Input
Sensing device
Interpreting device
Output

Figure 2. A diagram comparing the human biological and computer vision systems.

Understanding how the brain processes visual information is crucial for advancing artificial intelligence (AI), brain-computer interfaces (BCIs), and cognitive models that aim to mimic human perception. By studying neural representations of vision, we can develop more robust and biologically inspired AI systems, bridging the gap between human and machine intelligence.

1.4 Background Knowledge and Terminology

1.4.1 Ventral Stream

Figure 3 shows the **ventral stream**, which is responsible for object recognition and high-level visual processing. Our project data covers three main regions in the ventral stream:

1. **Primary Visual Cortex (V1)** - Handles low-level visual features such as edges and orientation.
2. **Visual Area (V4)** - Handles mid-level visual features such as shape, color, and texture.
3. **Inferotemporal Cortex (IT)** - Handles high-level object recognition such as hands, faces, etc.

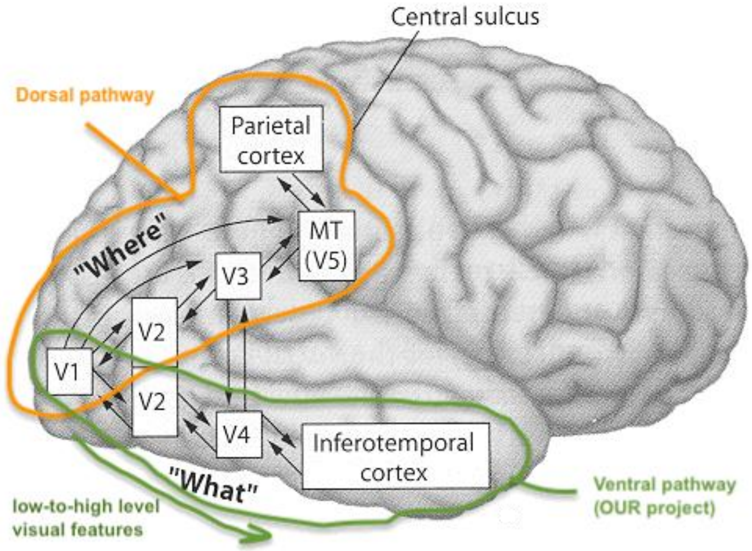

Figure 3. A diagram illustrating the ventral pathway (green) with the V1, V4, and IT regions for image processing.

In general, image processing information flows from V1 to V4 to IT (in order). Understanding this pathway can help researchers and engineers model biological vision and improve AI-based visual perception. Figure 4 shows the progressive transformation of visual features across the ventral stream, from low-level edge and orientation detection in V1 to more complex shape and texture processing in V4, and finally to high-level object recognition in IT.

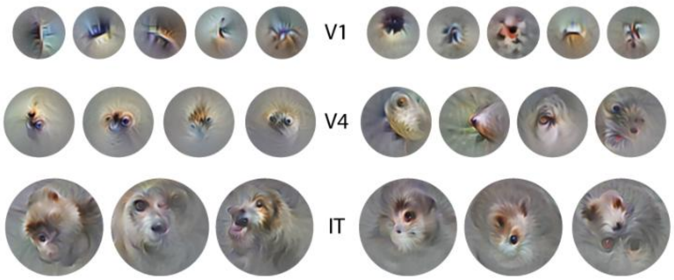

Figure 4. Visualization of neural feature representations in the ventral stream. V1 neurons respond to simple edges and orientations, V4 neurons capture mid-level features like texture and shape, and IT neurons are tuned to complex objects like faces.

1.4.2 Multi-Unit Activity (MUA)

This project makes use of **multi-unit activity (MUA)** data, which is summed spiking activity recorded from multiple neurons via electrodes. MUA data captures population-level neuronal responses, which balance single-unit precision and broader unit activity. It is particularly useful for high-resolution and high-precision recording of neurons, which can improve image generation quality compared to traditional methods like EEG and fMRI.

Figure 5 illustrates the electrode implantation sites and example neural responses recorded from two macaques performing a visual perception task, illustrating the hierarchical organization and functionality of the ventral stream. The left of the arrows shows the spatial organization of Utah electrode arrays implanted in V1 (red), V4 (blue), and IT (green). These electrodes capture spiking activity across different stages of visual processing, from low-level edge detection in V1 to complex object recognition in IT. As shown in panels E and F, the deeper the layer within the ventral stream, the longer the latency between stimulus onset and activation. Specifically, V1 exhibits a sharp, transient increase in activity following stimulus onset, reflecting early-stage feature detection; V4 shows a more gradual and sustained response, corresponding to mid-level processing of shape, color, and texture; IT demonstrates prolonged activation, consistent with high-level object recognition and memory-related processing.

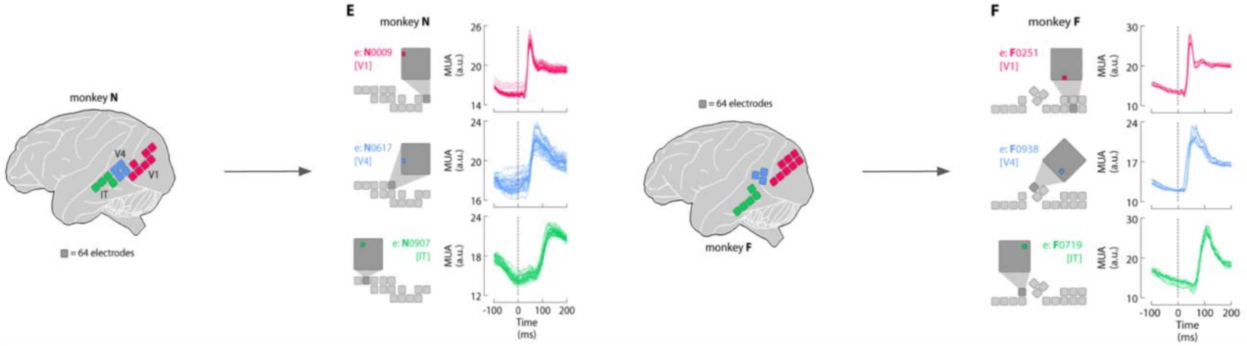

Figure 5. Electrode placements and multi-unit activity (MUA) responses from the ventral visual stream. The brain map shows electrode placements in monkeys N and F in three visual areas: V1 (red), V4 (blue), and IT (green). Each Utah array covers 64 electrodes. Panels E and F show MUA responses from representative electrodes in V1, V4, and IT, averaged across trials. The temporal profiles highlight differences in response dynamics, with V1 showing rapid, transient activity, V4 exhibiting more sustained responses, and IT displaying prolonged activity patterns.

Figures 4 and 5 can be found in reference [3].

2 Encoding

2.1 Dataset

We use the THINGS Ventral stream Spiking Dataset (TVSD) for single-neuron recording data. This dataset provides large-scale single-neuron recording data from two male macaques (monkeys N and F) during visual perception tasks [3]. The dataset captures spiking activity from the V1, V4, and IT regions in the ventral visual stream of the two primates in response to 22,248 images from the THINGS image database. This allows for a detailed investigation of how the visual system encodes and decodes information.

TVSD was chosen primarily because it is the only large-scale, publicly available dataset with single-neuron spiking activity recording during natural image perception. Its high temporal and spatial resolution and its broad coverage of ventral stream regions make it ideal for investigating entations of visual features. The dataset can be downloaded at: https://gin.g-node.org/doi/TVSD.

The dataset contains two main components:

1. Raw Multi-Unit Activity (MUA) Data
2. Normalized MUA Data

The raw MUA data consists of neural recordings collected over multiple days, structured in 20-minute trial blocks. 4 sessions were recorded for each monkey, with Monkey F spanning 7 days and 6,952 trials, and Monkey N spanning 4 days and 6,771 trials. On the other hand, the normalized MUA data contains preprocessed that have been aligned, averaged, and normalized for analysis. It includes train and test partitions sorted by stimulus order and maps brain responses to visual features. A more detailed data collection process is described in [3].

The dataset is split into 22,248 training images and 100 test images from the THINGS database, ensuring sufficient diversity evaluation.

2.2 Preprocessing

The dataset used in this study includes normalized multi-unit activity (MUA) signals recorded from two monkeys (Monkey F and Monkey N). We began by loading both the neural data and the corresponding image information. The MUA signals were stored in .mat files and separated into training and test sets. The image paths were encoded in HDF5 format, requiring special decoding using UTF-16 and manual conversion from Windows-style backslashes to Unix-style forward slashes. Once decoded, the image filenames were joined with their base directory to generate full paths, and the existence of each file was confirmed to ensure complete alignment between image stimuli and neural data.

Each image was then preprocessed to fit the input requirements of AlexNet, the convolutional neural network model used in this analysis. This included resizing the images to 224×224 pixels, normalizing them using ImageNet statistics, and converting them into PyTorch tensors. To ensure reproducibility across runs, we set a global random seed and configured PyTorch for deterministic computation, eliminating randomness in model behavior.

2.3 Analytical Methods

We considered AlexNet, ResNet, and DINOv2 to extract image features. AlexNet, a relatively shallow convolutional neural network, captures low-level visual features such as edges and textures, which closely resemble the early visual representations in the brain. ResNet, with its deeper architecture and residual connections, extracts more abstract and semantic features that may go beyond the processing capabilities of the early visual cortex. DINOv2, a self-supervised transformer-based model, encodes even higher-level scene understanding and semantic relationships, which might not align well with the immediate neural responses to raw visual input. Based on these features, we chose AlexNet as our first encoding model, likely because its representations are more biologically plausible and better matched to the kind of information encoded in MUA signals.

We previously experimented with using Signal-to-Noise Ratio (SNR) and oracle correlation，that SNR quantifies the reliability of neural responses by measuring how strong the stimulus-evoked signal is relative to background noise, and oracle correlation, represents the theoretical upper bound of prediction accuracy, computed by comparing repeated neural responses to the same stimulus. These two metrics were initially considered as potential tools to help us identify more reliable electrodes or to balance and improve the quality of MUA data. However, SNR tends to be stimulus-dependent—some electrodes may respond more strongly to specific items—and oracle correlation did not effectively serve as a criterion for electrode selection. Therefore, we decided to discard both metrics from further analysis.

2.4 Evaluation

AlexNet’s feature extractor consists of 13 sequential layers that progressively transform input images into increasingly abstract feature representations. Layer 0 is a convolutional layer that takes in RGB images and extracts low-level features such as edges and color blobs. Layer 1 applies a ReLU activation to introduce non-linearity. Layer 2 performs max pooling to reduce spatial resolution and retain dominant features. Layer 3 is another convolutional layer that builds upon the initial features to detect more complex patterns. Layer 4 is a ReLU activation, followed by Layer 5, another pooling layer that further compresses the spatial information. Layer 6 is a deeper convolutional layer that captures more localized structures. Layer 7 applies ReLU again, and Layer 8 is another convolutional layer that increases feature complexity. Layer 9 is the corresponding ReLU, followed by Layer 10, another convolutional layer that refines high-level part-based features. Layer 11 applies a final ReLU, and Layer 12 concludes the feature extractor with a max pooling layer, summarizing the spatially distributed information into compact feature maps. These 13 layers form the backbone for visual representation in AlexNet before the features are passed to the fully connected classifier.

In our analysis, we select all 13 layers from AlexNet and evaluate predictive performance by progressively aggregating features up to each layer. This approach allows us to assess how visual representations evolve through the network and to identify which cumulative feature set most effectively models neural responses. Since each convolutional layer builds upon the outputs of preceding layers, the upper layers inherently influence the subsequent ones. Convolutional layers extract increasingly complex spatial patterns—from edges and textures in early layers to object parts and structures in deeper ones—which are highly relevant for modeling the visual responses in neural data such as MUA. Although layers like ReLu and MaxPooling do not generate new features themselves, they shape the transformation and propagating of feature maps, and may still contribute to the formation of biologically meaningful representations. By evaluating all layers - including both convolutional and intermediate operations - we aim to develop a comprehensive understanding of how different levels of visual processing relate to neural activity.

To reduce the dimensionality of the high-dimensional image features extracted from each AlexNet layer, we applied Principal Component Analysis (PCA). High-dimensional features can be computationally expensive to work with and may contain redundant or noisy information, which can negatively affect model performance and lead to overfitting. PCA helps to compress the data while retaining the most important variance components, making the training of regression models more efficient and robust. We used Incremental PCA to handle large datasets by processing them in mini-batches. For each layer, we first determined the optimal number of principal components needed to preserve 95% of the variance. Then, we transformed both the training and testing feature sets using the learned PCA model and saved the results for downstream linear regression analysis.

To evaluate the model performance across electrodes, we use the **Pearson correlation** coefficient to quantify the similarity between predicted and observed neural responses. Pearson correlation measures the strength of the linear relationship between the model's predictions and the actual neural activity, providing an interpretable metric for assessing prediction accuracy. A higher correlation indicates better alignment and suggests that the model captures meaningful structure in the neural data. By comparing Pearson correlation values across layers and electrodes, we can identify which feature representations are most predictive of neural responses and assess how well the model generalizes across different neural recording sites.

2.5 Results

To evaluate which layer of AlexNet most effectively predicts multi-unit activity (MUA) responses recorded from monkeys, we trained separate linear regression models using PCA - reduced features from each of the 13 layers of the neural network. For every layer, we trained 1,024 models - one per electrode - on the training set and evaluated their performance on the test set. Prediction accuracy was quantified using the Pearson correlation between predicted and actual MUA responses. These correlations were then mapped to their corresponding electrodes and visualized as a heatmap to highlight spatial patterns in modal performance.

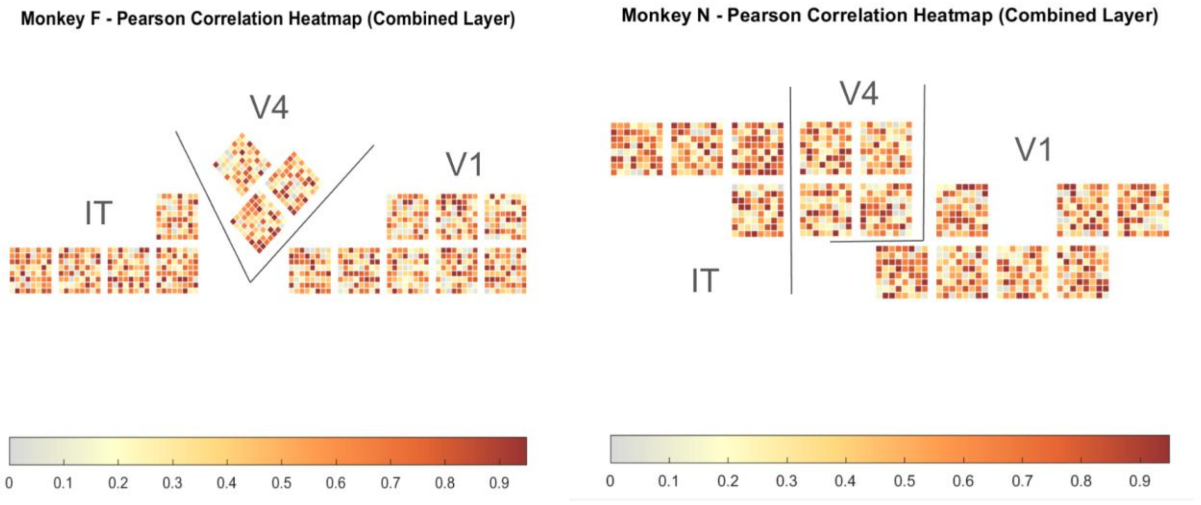

Figure 6. Electrode-wise prediction accuracy of MUA responses using features from all 13 layers of AlexNet (combined)

The heatmaps show Pearson correlation coefficients between predicted and actual MUA responses across all 1,024 electrodes for Monkey F (left) and Monkey N (right). Each value represents the performance of a linear regression model trained on PCA-reduced features aggregated from all 13 AlexNet layers. Warmer colors indicate higher predictive accuracy. The spatial distribution of correlation values highlights differences in model performance across visual areas (V1, V4, IT). Notably, the majority of electrodes exhibit moderate to high prediction accuracy in both monkeys, indicating that combined AlexNet features effectively capture neural response patterns.

To further explore the spatial distribution of prediction accuracy, we visualized electrode-wise Pearson correlations as heatmap, using features from the combined layers that showed the highest overall performance. Additionally, we performed a layer-wise and range-wise analysis of mean Pearson correlation. The line plot (Figure 7) shows the average Pearson correlation between predicted and actual MUA responses across all electrodes, for each of the 13 layers of AlexNet, separately for Monkey F and Monkey N. Prediction accuracy increases sharply from early layers and peaks between layer 5 and 8, indicating that intermediate-level features are most predictive of neural activity. Layer 7 yields the highest mean Pearson correlation in both monkeys. Specifically, for Monkey F, layer 7 yielded a mean correlation of approximately 0.62, while for Monkey N, the same layer reached 0.64. These values were higher than those obtained from any other layer, indicating that the features extracted at this stage of the network are the most predictive of the neural responses. So, mid-level representations in AlexNet, particularly those captured in layer 8, are most aligned with the way the brain encodes visual stimuli, likely corresponding functionally to mid or high-level visual areas such as V4 or IT in the primate visual system.

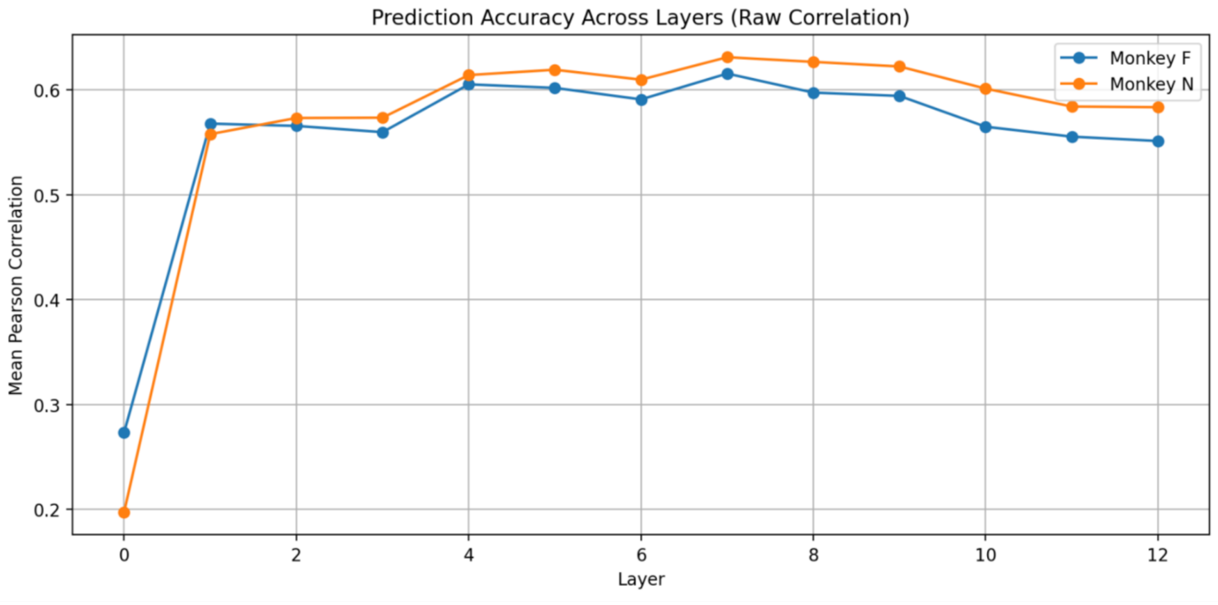

Figure 7. Mean Pearson correlation between predicted and actual MUA responses across 13 layers of AlexNet for Monkey F (blue) and Monkey N (orange)

To complement the layer-wise analysis, we assess how different stages of AlexNet contribute to prediction accuracy across cortical regions (V1, V4, IT). Each line in the plots represents the mean Pearson correlation between predicted and actual MUA responses for a specific region, computed separately for each AlexNet layer. In Monkey F, V1 consistently shows the highest correlation, particularly in early and mid layers. This pattern might be influenced by the electrode distribution, as Monkey F has fewer electrodes implanted in the V4 region and more in IT, potentially contributing to lower V4 prediction accuracy. In contrast, Monkey N shows the highest prediction accuracy in V4 across most layers, followed by V1 and IT. These region-specific trends suggest that different stages of the visual hierarchy align differently with the hierarchical feature representations learned by AlexNet. For each region in both monkeys, we can also identify the specific layer that yields the highest prediction accuracy.

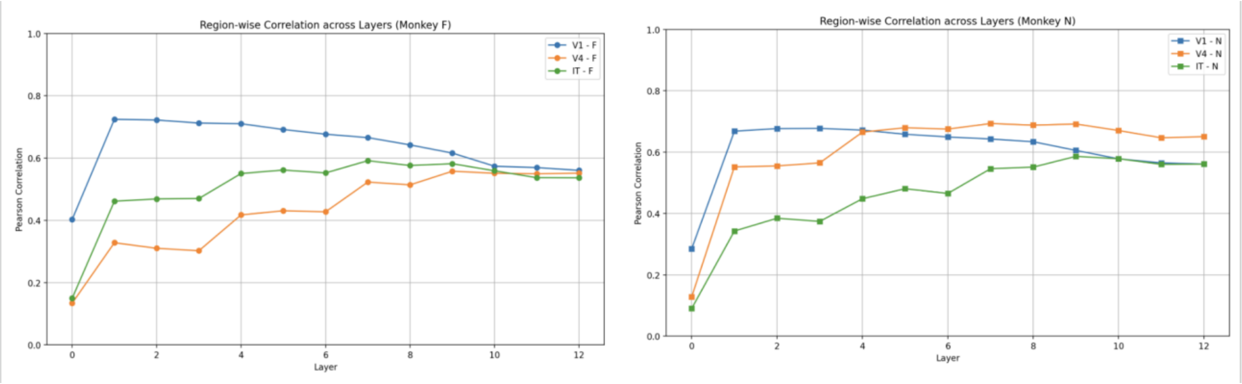

Figure 8. Region-wise Pearson correlation across AlexNet layers for Monkey F (left) and Monkey N (right)

3 Decoding

3.1 Introduction

The goal of the decoding model is to reconstruct visual stimuli from neural activity, specifically multi-unit activity (MUA) recorded from monkeys. The approach aims to bridge the gap between low-level neural signals and high-level visual content. Inspired by the Brain-Diffuser framework introduced by Ozcelik & vanRullen (2023) [5], which demonstrated successful image reconstruction from human fMRI data, we adapt and extend this two-stage decoding pipeline to work with MUA signals. By leveraging powerful pre-trained generative models and feature embedding spaces, this framework enables high-fidelity image reconstructions guided entirely by neural data.

3.2 Overview of Two-Stage Framework

Our decoding pipeline follows a two-stage framework adapted from the Brain-Diffuser model. In the first stage, we perform low-level visual reconstruction by predicting the latent variables of a Very Deep Variational Autoencoder (VDVAE) from MUA signals. These predicted latents are then decoded by pre-trained VDVAE to produce an initial, coarse image that captures the global layout and structure of the stimulus. In the second stage, we refine this image using Versatile Diffusion, a powerful image-to-image generative mode. This stage incorporates semantic conditioning by guiding the diffusion process with CLIP features - both vision and text - predicted from the same MUA signals. Importantly, only the regression models that map neural activity to feature spaces are trained, and all generative components - VDVAE, CLIP, and Versatile Diffusion - allow us to leverage their pre-trained representational power without requiring large-scale model training.

3.3 Data and Preprocessing

We used multi-unit activity (MUA) recordings from two monkeys (Monkey F and Monkey N) viewing natural images as visual stimuli. Neural activity was recorded via Utah arrays implanted in visual areas V1, V4, and IT, yielding responses from 1,024 electrodes per monkey. Each image was presented multiple times, and MUA signals were extracted in a defined time window to capture evoked responses. The MUA data were preprocessed by z-scoring across trials for each electrode to ensure normalization and comparability across sessions.

For each stimulus image, we extracted both low-level and high-level features. Low-level features were obtained by encoding images through a pre-trained VDVAE, producing hierarchical latent variables used in the first decoding stage. High-level semantic features were extracted using the CLIP model: CLIP-Vision embeddings captured abstract visual representation, while CLIP-Text embeddings were derived from image captions generated using a pre-trained BLIP model. These multimodal features served as targets for regression from MUA, enabling both structural and semantic reconstruction in the stages of the decoding pipeline.

3.4 Stage1: Low-Level Reconstruction with VDVAE

In the first stage of the decoding pipeline, we focused on reconstructing the coarse visual layout or the stimuli using a VDVAE (Figure 9). Each training image was passed through a pre-trained VDVAE encoder to obtain a set of hierarchical latent variables that capture the image's low-level structure. These latent variables served as the decoding targets for a set of ridge regression models trained to map preprocessed MUA signals to VDVAE latent space. Importantly, each regression model was trained independently for each latent variable dimension, using the MUA responses as input features.

Onced trained, these models were used to predict the VDVAE latents from neural activity in the test set. The predicted latents were then passed through the VDVAE decoder to generate low-resolution image representations. These initial reconstructions retained key aspects of the visual stimuli such as object layout, shape contours, and scene structure, despite lacking fine semantic detail. This stage provided a structurally grounded foundation for subsequent refinement in the second stage of the pipeline.

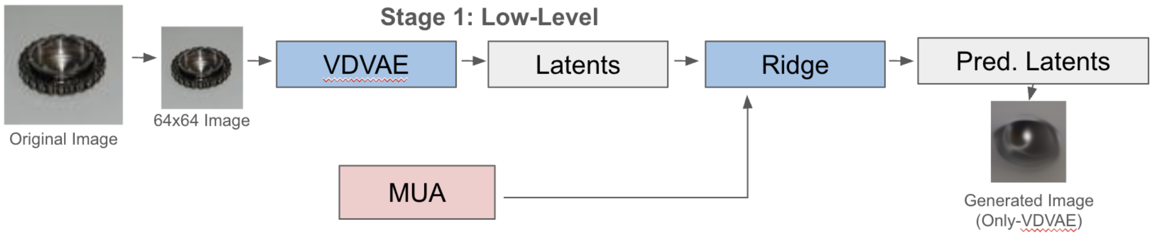

Figure 9. Stage 1- Low - Level Flow

3.5 Stage2: High-Level Reconstruction with Versatile Diffusion

In the second stage of the decoding framework, we refined the coarse VDVAE generated images using high-level semantic features extracted directly from the original stimulus images. For each image, we extracted two types of CLIP features: (1) CLIP-Vision embeddings representing abstract visual semantics, and (2) CLIP-Text embeddings derived from captions generated using a pre-trained BLIP model. These embeddings served as regression targets for two separate ridge regression models trained to map MUA signals to the CLIP-Vision and CLIP-Text spaces (Figure 10).

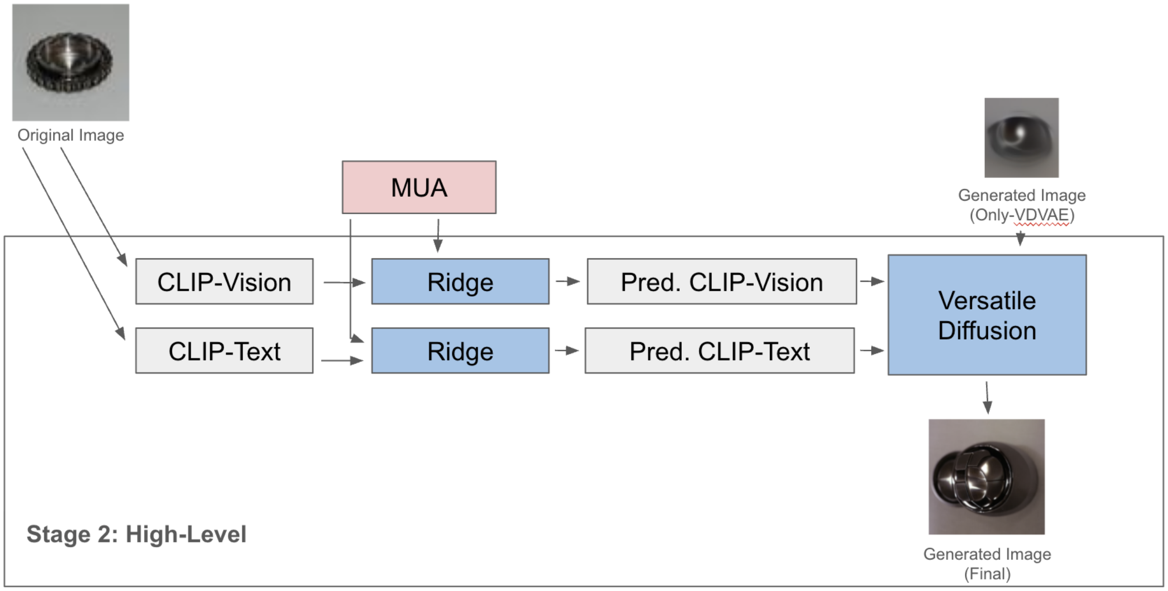

Figure 10. Stage 2 - High-Level Flow

In our current implementation, we completed the reconstruction pipeline using predicted CLIP-Vision embeddings along with VDVAE images as input to Versatile Diffusion. This combination enabled the model to enhance structural reconstructions with semantic detail.

3.6 Evaluation Metrics

To assess the quality of the reconstructed images generated by the decoding models, we employed a comprehensive set of evaluation metrics. These metrics quantify the similarity between the original ground truth images and the generated images at both low and high levels of visual representation.

Low-Level Similarity Measures:

- PixCorr (Pixel-wise Correlation): This metric calculates the pixel-wise correlation between the generated and ground truth images. A higher value indicates greater similarity. For this measure, generated images were downsampled from 512x512 to 425x425 pixels.
- SSIM (Structural Similarity Index): SSIM measures the perceptual closeness between two images, considering luminance, contrast, and structure. A higher SSIM score suggests better structural similarity. For this measure, generated images were downsampled from 512x512 to 425x425 pixels.
- AlexNet(2) & AlexNet(5): These metrics assess feature-wise similarity using the activations from the second (conv2) and fifth (conv5) convolutional layers of a pre-trained AlexNet model. Higher percentages indicate that the generated images share more similar low-to-mid level features with the ground truth images.

High-Level Similarity Measures:

- Inception: This metric compares images using features extracted from a pre-trained InceptionV3 model, focusing on higher-level object features. A higher score indicates better semantic similarity.
- CLIP (Contrastive Language-Image Pre-training): This involves a 2-way cosine similarity of image embeddings generated by the CLIP-Vision model. It measures how well the generated image aligns semantically with the ground truth image from the perspective of a multimodal model. A higher value is better.
- EffNet-B (EfficientNet-B1) & SwAV (Swapping Assignments between multiple Views): These metrics calculate the feature space distance using EfficientNet-B1 and SwAV-ResNet50 models, respectively. Lower distances indicate greater similarity in the high-level feature representations learned by these deep neural networks.

For PixCorr, SSIM, AlexNet features, Inception, and CLIP, higher scores denote better performance, while for EffNet-B and SwAV, lower scores are preferable

3.7 Results

The decoding pipeline was evaluated in stages to understand the contribution of different components and feature types. The performance was benchmarked against a reference reconstruction that utilized only CLIP-Vision features extracted directly from the ground-truth images, representing an upper bound for the generative capability of the Versatile Diffusion model without any brain signal input.

Stage 1: VDVAE-Only Decoding

- Performance: In this stage, only the VDVAE latent variables predicted from Multi-Unit Activity (MUA) were used to reconstruct images. This method achieved the highest scores on low-level metrics such as PixCorr (0.4435), SSIM (0.5915), and AlexNet(2) (71%). Conversely, it performed the poorest on high-level semantic metrics like Inception (15.5%), CLIP (10%), EffNet-B (0.73785), and SwAV (0.54495).
- Interpretation: These results indicate that decoding based solely on VDVAE latents successfully captures structural details of the visual stimuli, such as layout, shape, and texture, but largely misses high-level semantic content like object identity or category meaning. This stage serves as a baseline for decoding performance focusing on structural fidelity.

Stage 2: VDVAE + CLIP-Vision Features Decoding

This stage enhanced the reconstruction by incorporating predicted CLIP-Vision features along with the predicted VDVAE latents from MUA, feeding both into the Versatile Diffusion model.

- Performance: This combined approach yielded intermediate scores across both low-level and high-level metrics. For instance, PixCorr was 0.3185, SSIM was 0.3324, AlexNet(5) was 56.5%, Inception was 17.5%, and EffNet-B was 0.8568.
- Interpretation: The inclusion of CLIP-Vision features improved object detail and semantic content compared to the VDVAE-only stage, demonstrating that combining low-level structural information (from VDVAE) with high-level semantic content (from CLIP-Vision) leads to a more balanced image reconstruction

Reference: CLIP-Vision Features Only (No Brain Data)

This method used CLIP-Vision features extracted directly from the ground truth images (not from MUA) as input to Versatile Diffusion to generate images.

- Performance: It scored highest on high-level semantic metrics: Inception (90.0%), CLIP (47.0%), and achieved the best (lowest) scores for EffNet-B (0.5247) and SwAV (0.3205). However, it performed relatively poorly on low-level metrics like PixCorr (0.1313) and SSIM (0.2583).
- Interpretation: This confirms the Versatile Diffusion model's capability to generate semantically accurate images when provided with clear high-level visual features and serves as an upper bound for the diffusion model's performance in this context.

Method	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Only CLIP-Vision (Reference)	0.1313	0.2583	48%	85.00%	90.00%	47.00 %	0.5247	0.3205
Only VDVAE	0.4435	0.5915	71%	27.50%	15.50%	10%	0.73785	0.54495
VDVAE + CLIP-Vision	0.3185	0.3324	49%	56.50%	17.50%	8.50%	0.8568	0.58325

Table 1: Quantitative Comparison of Decoding Methods by Evaluation Metrics

Figure 11 provides a qualitative comparison of these image reconstruction approaches. The top row displays the original ground truth images. The second row shows reconstructions from Stage 1 (VDVAE latent features predicted from MUA), illustrating the capture of low-level structural information but a general lack of semantic detail. The third row presents reconstructions from Stage 2 (combining predicted VDVAE latent features and CLIP-Vision features from MUA via Versatile Diffusion), demonstrating improved object detail and semantic content compared to Stage 1. The bottom row shows the reference reconstruction using only CLIP-Vision features from the ground truth images (no brain data), highlighting the generative capability of the diffusion model with ideal input. These visual results align with the quantitative metrics, where Stage 1 reconstructions tend to preserve basic structure and the reference reconstructions achieve high semantic fidelity, while Stage 2 offers a promising balance when decoding from brain-derived features.

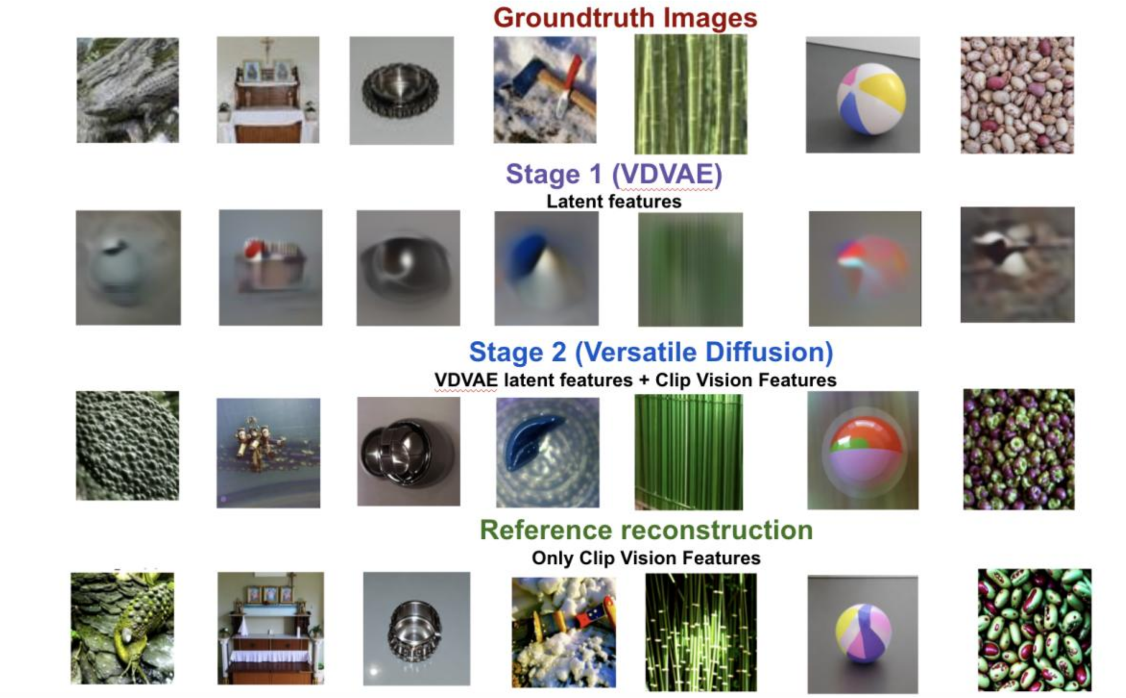
Groundtruth Images
Stage 1 (VDVAE)
Latent features
Stage 2 (Versatile Diffusion)
VDVAE latent features + Clip Vision Features
Reference reconstruction
Only Clip Vision Features

Figure 11: Qualitative comparison of image reconstruction results

Final conclusions regarding the full Brain-Diffuser decoding pipeline are pending the incorporation of CLIP-Text features predicted from MUA. Ongoing experiments are exploring this, as well as a CLIP-only pipeline (CLIP-Vision + CLIP-Text without VDVAE) and a text-guided latents approach (VDVAE + CLIP-Text without CLIP-Vision) to fully assess the integrated contribution of structural, visual, and textual embeddings.

4 Discussion

4.1 Summary and Conclusion

Our study demonstrates the effectiveness of using deep convolutional neural network features from AlexNet to predict neural response in the primate visual cortex. Unlike the prior studies relying on fMRI, which offers limited and temporary resolution, we leveraged normalized multi-unit activity (MUA) – a high-resolution, direct measure of population-level spiking activity – to probe the neural code. Our findings reveal that the mid-to-deep layers of AlexNet mostly strongly align with neural responses in higher-order cortical areas such as IT, supporting for object recognition recapitulate the hierarchical structure of visual processing observed in the primate brain.

In our decoding analysis, we adapted the Brain-Diffuser framework to reconstruct visual stimuli from MUA signals. Our two-stage decoding model first recovered coarse structural information using latents variables from pre-trained VDVAE, and then refined the reconstructions using Versatile Diffusion conditioned on predicted CLIP-Vision and CLIP-Text features extracted from original images. The approach enabled high-fidelity reconstruction that preserved both spatial layout and semantic content, demonstrating that MUA encodes rich information about both low-level visual structure and high-level semantics. Our results highlight the representational similarity between deep neural networks and primate brain, and demonstrate the potential of high-resolution neural recordings to reconstruct perceptual experiences.

4.2 Future Work

We plan to expand our encoding analysis by accessing and analyzing the full raw NUA dataset, rather than relying on the normalized version. The raw dataset contains richer spatiotemporal information, including detailed electrode and Utah array configurations, which may improve both decoding accuracy and resolution. Additionally, we plan to evaluate oracle correlation using repeated trial data (30 trials per image) to more precisely quantify the upper bound of predictability for each electrode. However, managing and preprocessing the raw MUA dataset presents computational challenges due to its large size (54.6GB per monkey, compared to 203.1 MB for normalized MUA).

In the encoding stage, after discussions with the professor, we eventually decided to discard quality-related metrics such as oracle correlation. However, since such metrics must have some intrinsic value, and given that we did not apply any processing to the MUA data during encoding, we believe that in the future, it may be worthwhile to revisit oracle correlation as a way to adjust and correct the MUA signals we obtained.

In parallel, we are conducting a set of targeted experiments to better understand how different representational modalities contribute to brain-based image reconstruction. In the **CLIP-only pipeline**, we aim to isolate the effect of high-level semantic embeddings by combining predicted CLIP-Vision and CLIP-Text features - without structural input from VDVAE - in the Versatile Diffusion framework. In the **Text-guided Latents** setup, we explore whether CLIP-Text embeddings can enhance low-level structure by pairing them with VDVAE generated reconstructions, excluding CLIP-Vision guidance. Lastly, in the **Full Brain-DIffuser model**, we will integrate all three components - VDVAE, CLIP-Vision, and CLIP-Text - to evaluate how structural. Visual, and textual signals jointly influence reconstruction quality. These experiments are expected to clarify the unique and combined contributions of each modality and help us build more interpretable and accurate neural decoding models.

Last but not least, during the midterm, the professor gave us feedback that the connection between the encoding and decoding components of our research needed to be strengthened. In the future, we could explore how to better integrate these two modules and apply them under real-world conditions. For example, we could simulate a monkey's brain entirely using models — generating simulated neural responses from image features and then reconstructing images from this synthetic data. These results could then be used to investigate brain responses in other animals, potentially eliminating the need to implant electrodes into their brains each time, and thus, to some extent, avoiding certain ethical issues.

4.3 Ethical Considerations

This study uses publicly available neural data collected under strictly regulated conditions. While brain implant surgeries raise ethical concerns due to their invasive nature and impact on animal welfare, our work focuses solely on data analysis and does not involve direct experimentation, such as performing surgeries or modifying neural activity.

5 Contribution

- **Michael Zhou (mgz2112)**: Led task delegation and project coordination. Implemented the full decoding pipeline using CLIP features and contributed to the encoding pipeline with AlexNet.

- **Qingwen Liu (ql2549)**: Coordinated internal team meetings, set project milestones, and tracked progress. Implemented heatmaps for the encoding model and developed the decoding model, integrating CLIP-Text and VDVAE. The primary contributor to both presentation slides and the final project report.

- **Yueran Wang (yw4441)**: Main contributor to dataset exploration and literature review. Analyzed the encoding results and developed the evaluation pipeline for the decoding model. Analyzed decoding results using eight low-level and high-level measures and synthesized key insights to assess decoding performance. The primary contributor to the report and final presentation.

- **Chengyan Li (cl4604)**: Implemented visualizations for the encoding model and analyzed the resulting data to extract key findings.The primary contributor to the literature review and the presentation.

References

[1] Amirhossein Farzmahdi, Wilbert Zarco, Winrich A Freiwald, Nikolaus Kriegeskorte, Tal Golan (2024) **Emergence of brain-like mirror-symmetric viewpoint tuning in convolutional neural networks** *eLife* 13:e90256 https://doi.org/10.7554/eLife.90256

[2] Adeli, H., Minni, S., & Kriegeskorte, N. (2023). **Predicting brain activity using Transformers.** bioRxiv, 2023-08. [bioRxiv]

[3] Papale, Paolo & Wang, Feng & Self, Matthew & Roelfsema, Pieter. (2025). **An extensive dataset of spiking activity to reveal the syntax of the ventral stream**. *Neuron*. 10.1016/j.neuron.2024.12.003.

[4] Caron, M., et al, "**Emerging Properties in Self-Supervised Vision Transformers**," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[5] Ozcelik, F., VanRullen, R. **Natural scene reconstruction from fMRI signals using generative latent diffusion**. *Sci Rep* 13, 15666 (2023). https://doi.org/10.1038/s41598-023-42891-8

[6] Child, Rewon. (2020). **Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images**. 10.48550/arXiv.2011.10650.