Image Reconstruction from Human Brain Activity Using Diffusion Models

Sarabesh Neelamegham Ravindranath
Ismail Ben Seddik
Babasanmi Adeyemi
Tauqeer Saleem

December 2024

Introduction

Reconstructing visual experiences from human brain activity is a compelling challenge at the intersection of computational neuroscience and computer vision. Our project addresses this problem by leveraging latent diffusion models (LDMs), specifically Versatile Diffusion, to reconstruct high-resolution images from functional magnetic resonance imaging (fMRI) data. The core of our framework involves deriving latent representations directly from fMRI signals using VDVAE, which are then integrated into the diffusion process. Additionally, CLIP's ability to align vision and text embeddings enhances the reconstruction. By combining these advanced techniques, our methodology addresses the challenges of transforming neural representations into high-quality images, bridging a crucial gap in the field.

Problem and Motivation

The primary problem addressed in this project is reconstructing images from human brain activity using diffusion models. This is a challenging task in computational neuroscience and machine learning, aiming to bridge the gap between neural signals and visual perception [1, 6]. Understanding and reconstructing the visual experiences encoded in brain activity is crucial for advancing brain-computer interfaces, diagnosing neurological disorders, and enhancing our understanding of human cognition. This research aligns with the growing interest in utilizing machine learning to interpret complex biological signals, particularly in neuroimaging and fMRI studies [1].

Recent advancements in generative AI, particularly in diffusion-based models, have made this endeavor more feasible and effective. Diffusion models, such as Stable Diffusion [2] and Versatile Diffusion [3], have shown impressive capabilities in generating high-resolution images, videos, and multimodal content by progressively refining random noise into coherent visual or textual representations. The ability of these generative models to work across different modalities—such as images, text, and video—means that they can effectively bridge the gap between neural signals and visual content [4, 5].

Dataset

The **Natural Scenes Dataset (NSD)**, as described by Allen et al. [1], is a groundbreaking ultra-high-field (7T) fMRI dataset designed to study brain activations based on perceived images. The dataset contains functional imaging data from **8 healthy adult participants**, collected across multiple sessions while the subjects viewed thousands of natural scenes.

- **Number of Sessions and Trials:** Data was collected over **30-40 scanning sessions** for each participant, ensuring robust coverage of visual experiences across multiple sessions.
- **Number of Images:** Participants viewed over **10,000 unique natural images**, providing extensive visual stimuli to capture a wide range of visual experiences.
- **High Temporal and Spatial Resolution:** The dataset offers **1.8-mm isotropic resolution** across the entire brain, with a **1.6-second sampling rate**, enabling the tracking of dynamic neural responses with high temporal fidelity.
- **Scale and Diversity:** Designed to capture a diverse set of visual experiences across subjects, sessions, and trials, making it an ideal resource for reconstructing complex visual representations from brain activity. It uses COCO image dataset across all the subjects
- **Availability:** The dataset is publicly accessible through the **Center for Magnetic Resonance Research (CMRR)** at the **University of Minnesota**.

Due to its comprehensive scale and high resolution, the NSD dataset serves as a crucial foundation for projects aimed at reconstructing high-fidelity visual experiences from neural signals.

Approach and Novelty

Our approach leverages latent diffusion models to reconstruct images from brain activity using the Natural Scenes Dataset (NSD). We draw inspiration from recent works, such as MindEye2 [4], Ozcelik et al. [5], and Takagi et al. [6]. In these studies, the process of reconstructing visual experiences from fMRI data has highlighted the potential of combining generative models with cross-modal representations.

Following the ideas from MindEye2, we aim to map fMRI signals to visual representations by first creating a robust base image. Unlike direct methods, where fMRI signals are fed straight into the diffusion model, we generate this base image using an autoencoder trained to capture essential visual features. We also draw on techniques from Ozcelik et al. [5], who employ generative latent diffusion models to reconstruct natural scenes from fMRI signals. Additionally, we build upon Takagi et al. [6], which demonstrates high-resolution image reconstruction from brain activity using latent diffusion models.

Our method incorporates CLIP embeddings that capture visual and textual representations of the input images. Visual embeddings are extracted directly from the images, while textual embeddings are obtained from image captions. These embeddings are then integrated into the diffusion process to conditionally refine the base image, ensuring a more accurate mapping between brain activity and visual content.

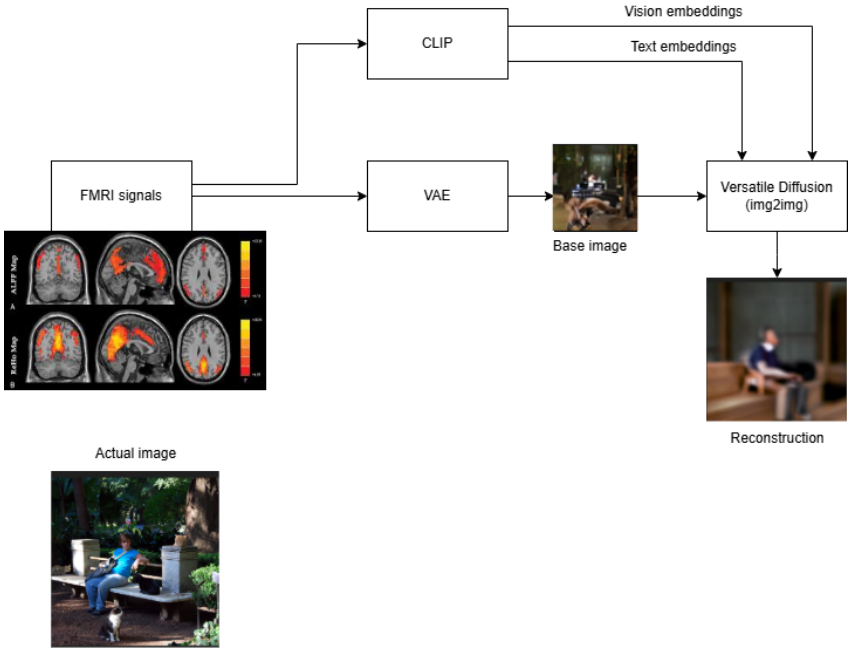

Figure 1: Image Creation Pipeline. The process includes fMRI input, base image generation using an autoencoder, CLIP embedding extraction (visual and textual), and conditional refinement using the diffusion model.

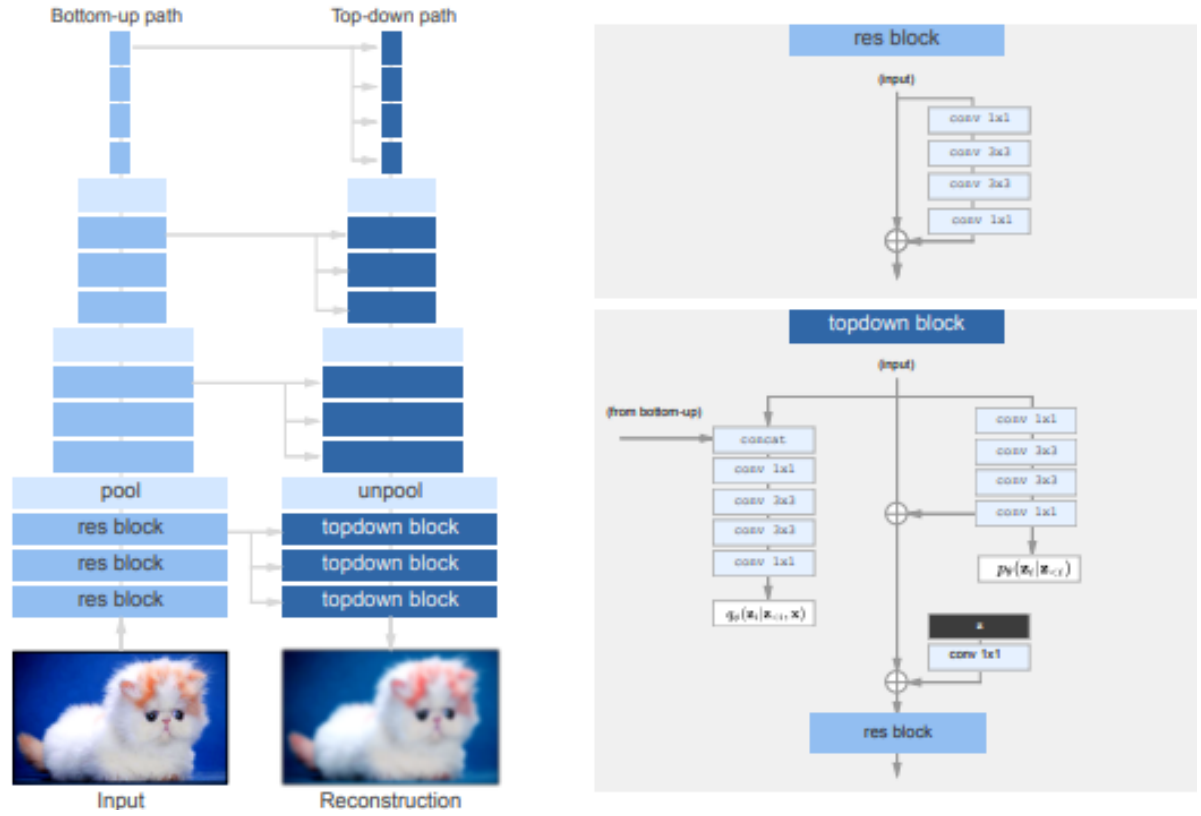

Figure 2: A diagram of the top-down VDVAE architecture for learning complex distributions.

Methodology

VAE for Base Image Generation

We utilize a Variational Autoencoder (VAE) to generate an initial base image (reconstruction) from fMRI input. The VAE serves as a crucial step in our pipeline, capturing essential visual features that form the foundation for our diffusion-based image reconstruction. The use of a very deep VAE is particularly important for handling natural images, as these images contain intricate structures, textures, and patterns. A deeper VAE, as introduced and demonstrated in the research by Child et al. [**?**], has the ability to capture complex visual features more effectively than shallow models. Their findings show that very deep VAEs can generalize autoregressive models and perform well for natural image reconstruction, ensuring more coherent and high-quality visual representations.

In our approach, the VAE maps the input fMRI data to a latent space that retains the significant visual patterns present in the original scenes. This latent representation is then used to generate a base image that accurately reflects the visual content encoded in brain activity. By doing so, we provide a meaningful starting point for our diffusion

model, which subsequently refines this base image by incorporating additional visual and textual information through CLIP embeddings.

The generated base image retains crucial details and structures, serving as an optimal input to our diffusion process. This enables the reconstruction of high-fidelity images from neural signals, leveraging the strengths of deep generative models to bridge the gap between brain activity and visual perception.

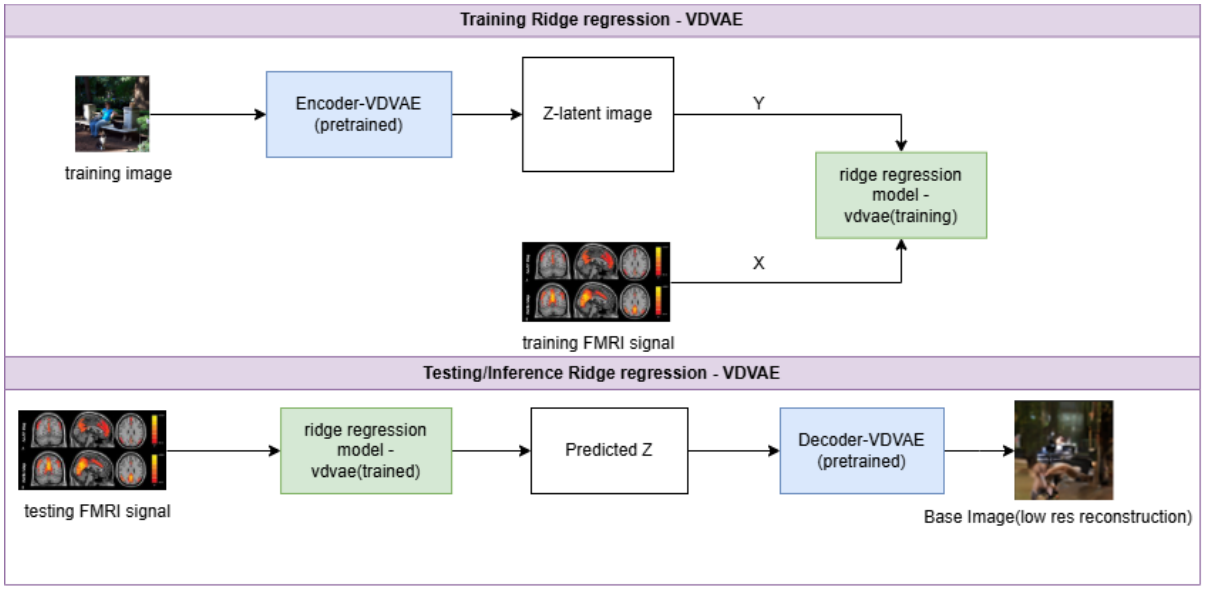

Figure 3: Training and Testing pipeline for VAE-based image reconstruction. During training, fMRI data is paired with latent representations extracted by the VAE encoder, and a regression model is trained to predict latent z. During testing, fMRI data is passed through the regression model, and the predicted latent vectors are decoded back into reconstructed images.

Training FMRI-Z regressor

During the training phase, we follow a structured process to map fMRI signals to the latent representations of the VAE. For each training sample, we first use the encoder component of the VDVAE to process a set of training images. This encoder extracts latent representations z from these images. We then pair these latent representations with the corresponding fMRI signals from the training dataset. Using these pairs, we train a ridge regression model. The objective of this regression model is to predict the latent z for a given input fMRI signal.

In essence, the regression model learns a mapping from the fMRI signal space to the latent representation space of the VAE. This ensures that during inference, we can accurately reconstruct images by mapping test fMRI data back into the latent representations through the trained regression model.

Testing / Inference FMRI-Z regressor

During the testing phase, we take the test fMRI signals and pass them through the trained regression model to predict the corresponding latent representations z. These predicted latent vectors are then sent to the decoder of the VDVAE, which reconstructs the images from the latent space.

The process ensures that the generated images capture the visual content encoded in the test fMRI signals with high fidelity. This approach leverages the relationship between brain activity and visual representations while maintaining the coherence and details captured by the deep generative model.

CLIP Embedding Integration

CLIP (Contrastive Language-Image Pre-Training) embeddings capture both visual and textual information from images. In our approach, we use CLIP to condition the diffusion model by leveraging meaningful representations extracted from both visual and textual sources. We split the CLIP integration into two distinct components: CLIP Vision and CLIP Text.

CLIP Vision

For CLIP Vision, we focus on extracting visual embeddings from input images. During training, we send training images to the CLIP image encoder to obtain visual embeddings. We then create a regression model to map corresponding fMRI signals to these visual embeddings. The objective is to learn a direct mapping between the brain activity data and the visual representations captured by CLIP.

During testing, we send test fMRI signals to the trained regressor, which outputs the corresponding visual embeddings. These embeddings guide the diffusion model to generate images that accurately reflect the visual content encoded in brain activity.

CLIP Text

For CLIP Text, we capture textual information from image captions sourced from the COCO dataset. During training, we first pass image captions through the CLIP text encoder to get textual embeddings. Then, we train a separate regression model to map fMRI signals to these textual embeddings, ensuring that the relationship between brain activity and textual constraints is learned effectively.

During testing, we input test fMRI signals to the corresponding regressor, which predicts the textual embeddings. These embeddings condition the diffusion model, ensuring that the reconstructed images align with the textual constraints provided by CLIP.

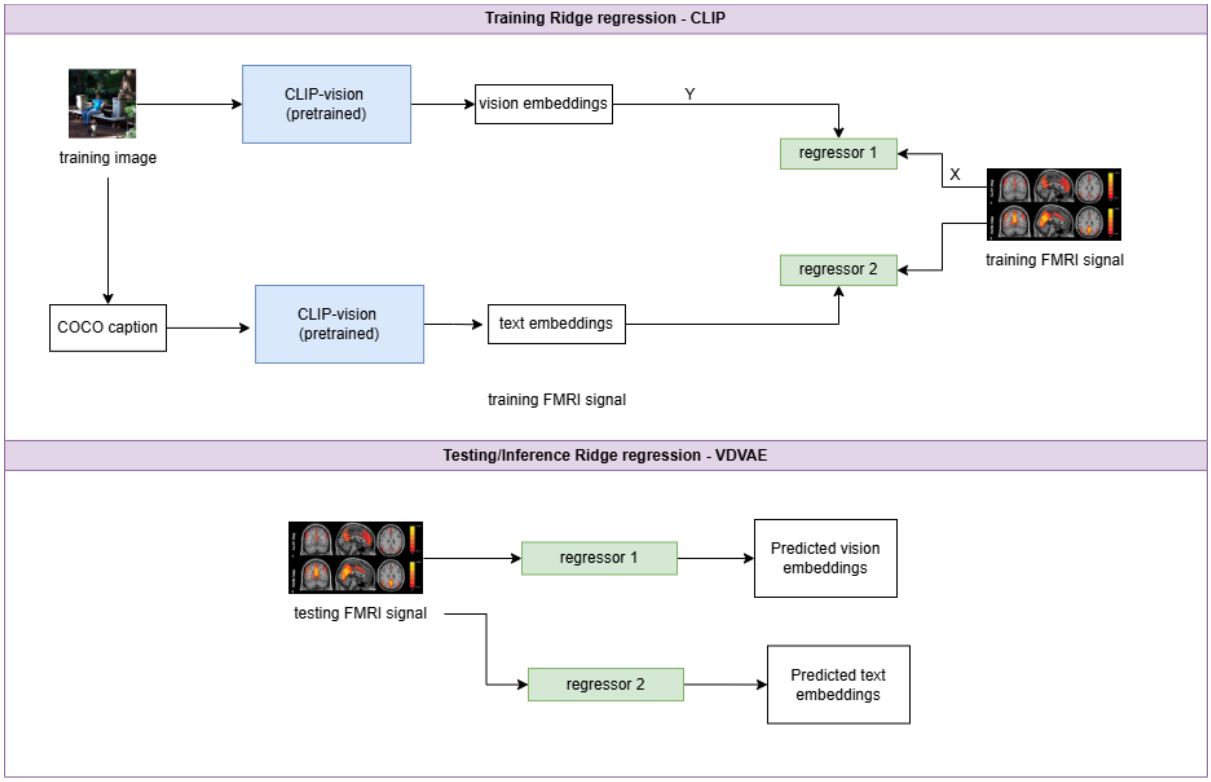

Figure 4: CLIP Embedding Integration pipeline for Vision and Text. For CLIP Vision, the training phase maps fMRI signals to visual embeddings using the CLIP image encoder. For CLIP Text, the training phase maps fMRI signals to textual embeddings obtained from image captions. During testing, test fMRI signals are processed through the regressors to obtain both visual and textual embeddings, which guide the diffusion model to reconstruct images.

Versatile Diffusion Model

To reconstruct high-fidelity images that align with neural activity, we employed the **Versatile Diffusion** model [3], which is designed for generating multimodal outputs conditioned on textual, visual, or a combination of embeddings. The model leverages a shared latent space and diffusion-based refinement to enable flexibility across tasks like text-to-image generation, image-to-image translation, and image variations.

1. **Base Image Creation**:

    - The fMRI signal is first sent to the trained fMRI-Z regressor, which outputs the predicted latent variable Z.
    - The predicted Z is then sent to the decoder of the pretrained VDVAE, which generates the base image. This image acts as the initial starting point for the diffusion process capturing low-level details of the image.

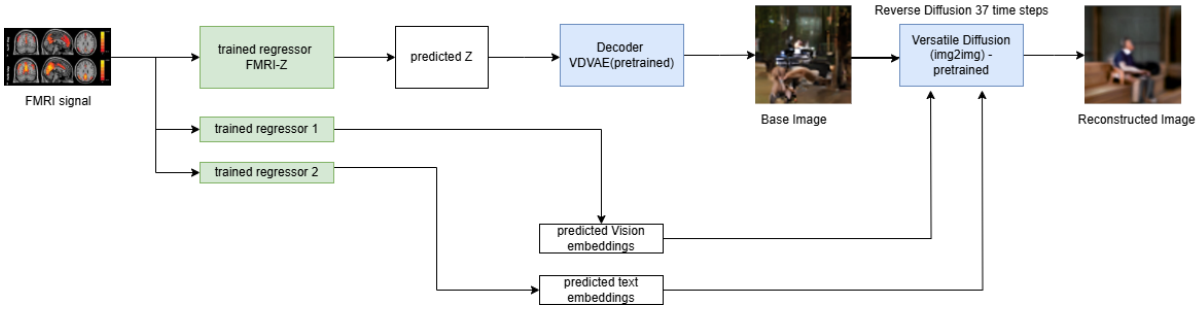

Figure 5: Pipeline for Versatile Diffusion with Autoencoder. fMRI signals are processed to predict latent Z (VDVAE-based), which generates a base image. The base image is encoded into latent space, refined using reverse diffusion conditioned on CLIP embeddings, and finally decoded to produce the reconstructed image.

2. **CLIP Embedding Predictions**:

   - The same fMRI signal is sent to:
     (a) **Regressor 1**: Predicts the corresponding **CLIP vision embedding**, which captures visual features of the target image.
     (b) **Regressor 2**: Predicts the corresponding **CLIP text embedding**, which represents textual attributes derived from the image captions.

The process continues with:

- **Latent Extraction and Reverse Diffusion**:
  - The base image is sent to the **Versatile Diffusion Autoencoder**, which encodes the image into a latent representation Z_{latent}.
  - A reverse diffusion process is initiated, starting with the Z_{latent} and conditioned on:
    * Predicted CLIP vision embeddings (obtained from Regressor 1 trained on fMRI signals and CLIP vision features).
    * Predicted CLIP text embeddings (obtained from Regressor 2 trained on fMRI signals and CLIP text features).
  - The reverse diffusion process iteratively updates Z_{latent}, ensuring it incorporates multimodal information from the predicted CLIP embeddings. The updated latent representation aligns the image with the constraints derived from fMRI signals.

- **Reconstruction Using the Autoencoder Decoder**:
  - The refined latent Z_{latent} is passed through the AutoKL decoder of the Versatile Diffusion model.

– The AutoKL decoder generates the final reconstructed image, which integrates the visual, textual, and neural constraints while maintaining high fidelity and coherence.

Versatile Diffusion Conditioning with Cross-Attention

To refine the base image into a high-fidelity reconstruction, we utilize the **Versatile Diffusion** model [3]. This model conditions the image generation process on both **CLIP text and vision embeddings** through a cross-attention mechanism.

- **Cross-Attention Mechanism:** - At each reverse diffusion step, the latent representation Z_{latent} acts as the *query*, while the CLIP embeddings serve as *keys* and *values*. - Cross-attention computes a weighted sum of the embeddings:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

  - This updates Z_{latent}, injecting text and vision constraints into the generation process.

- **Reverse Diffusion and Decoding:** - Over multiple timesteps, the conditioned Z_{latent} is iteratively refined. - The final latent representation is decoded using the AutoKL decoder to produce the reconstructed image.

This cross-attention-based conditioning allows the model to dynamically incorporate semantic and visual information, ensuring that the reconstructed image aligns with both modalities and accurately reflects the input fMRI signal.

Challenges and Solutions

Challenges

One of the main challenges was finding the correct diffusion models to run. Initially, we tried using Stable Diffusion and Stable Diffusion XL following the MindEye2 framework. However, we encountered significant issues with out-of-memory (OOM) errors even during inference. These issues made it difficult to successfully execute diffusion models, as our local hardware setup with a 3050 GPU (4 GB VRAM) and 32 GB RAM was unable to handle the memory requirements. Additionally, we faced difficulties in accessing and properly utilizing data from the NSD dataset. The process of reading, preprocessing, and integrating this data into our models was cumbersome and time-consuming. This further complicated our attempts to train or run inference on diffusion models, as our local environment lacked sufficient resources to support these operations.

Solutions

To address these issues, we ultimately chose to settle on VDVAE and Versatile Diffusion models, as they are comparatively smaller and more manageable in terms of size and memory requirements. For working with the NSD dataset, we employed Nibabel, a library that allowed us to efficiently read and process neuroimaging data. This approach made it possible to seamlessly handle the data integration and preprocessing steps. To overcome the computational limitations of our local setup, we utilized a Jetstream instance with impressive specifications. The setup included 32 CPU cores, 117 GB of RAM, a 990 GB root disk, and an NVIDIA A100-SXM4-40GB GPU with 40 GB VRAM. These resources provided the necessary computational power and memory capacity to train and run the diffusion models efficiently, allowing us to successfully implement the experiments without encountering the memory and performance issues that were present in our local environment.

Results and Analysis

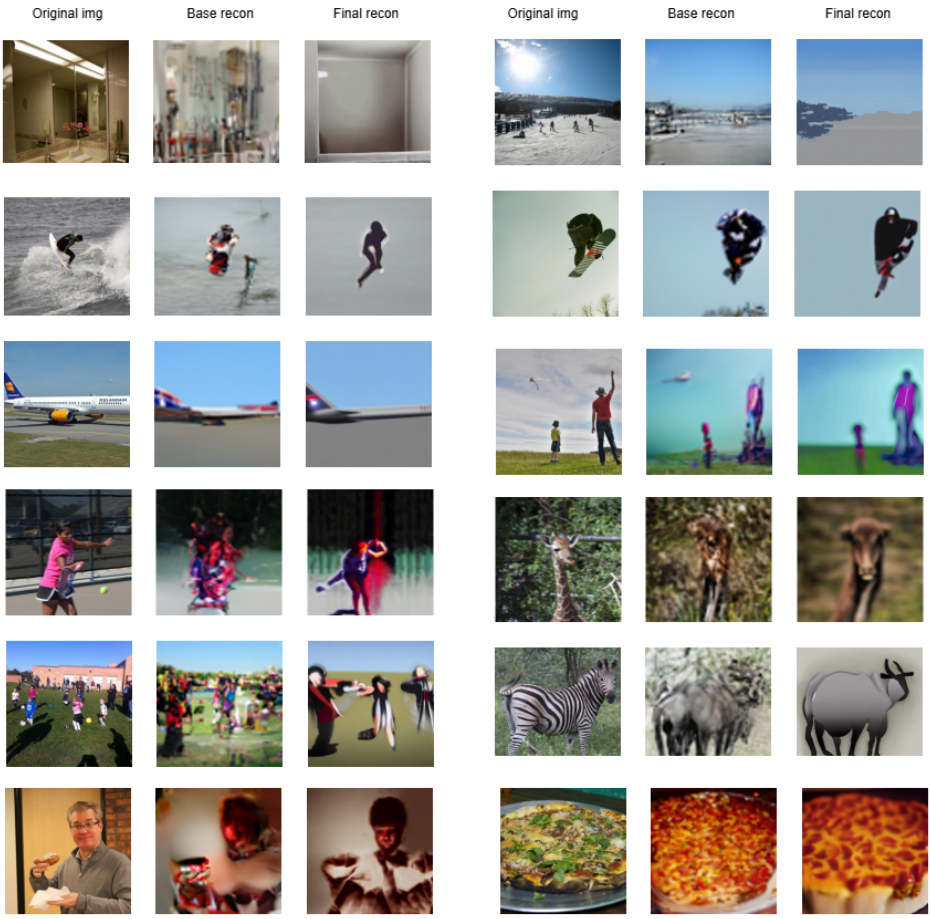

Figure 6: Example output: Reconstructed images alongside ground truth.

The results demonstrate that our approach effectively reconstructs images with high fidelity and detail. We were able to achieve decent reconstructions for some samples. The method performed well for images containing single targets, but the reconstruction quality degraded for more complex images filled with multiple items. We hypothesize that performing a region-of-interest (ROI) extraction before reconstruction could potentially improve the outcomes in such cases.

Additionally, the diffusion model showed better performance for common items such as flights, trains, and snowboarding people. This is likely due to the training of the diffusion model on a dataset that contained many such examples, enabling it to generate more accurate reconstructions for these specific categories. However, our method did not work well for a significant number of images. We believe that experimenting with different diffusion models, such as Stable Diffusion, Stable Diffusion XL, DALL-E 2, or DIT transformer-based diffusion models, might improve the reconstruction quality. Nevertheless, trying these models would require substantial computational resources and more training time, which could be a limiting factor.

Code

Below is the link to the GitHub repository containing the code for neural reconstruction:

`https://github.com/sarabesh/Neural-Recon`

Conclusion

In conclusion, we successfully achieved reconstructions that closely follow the methods and results presented in the referenced papers. Our approach highlights the potential of leveraging diffusion models to decode and reconstruct brain activity into meaningful visual representations. We demonstrated that by swapping out or experimenting with different diffusion models, we could further enhance the reconstruction quality. This flexibility opens up opportunities to utilize large pre-trained models, which are increasingly available and well-trained on massive datasets.

Our findings also contribute to a better understanding of how the brain processes and represents images and objects. Decoding complex visual information from brain activity provides insights into cognitive processes, visual perception, and object recognition. Such advancements have the potential to drive research in neuroscience, improve brain-computer interface technology, and aid in the diagnosis and treatment of vision-related disorders.

Possible applications of this work include developing tools for neuroimaging analysis, enhancing virtual reality interfaces driven by brain activity, and contributing to assis-

tive technologies for individuals with visual impairments. By refining these models and approaches, we can explore new avenues to bridge the gap between neural representations and cognitive understanding of visual stimuli, offering deeper insights into human perception and brain functionality.

References

[1] Allen, E.J., St-Yves, G., Wu, Y. et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25, 116–126 (2022). 10.1038/s41593-021-00962-x.

[2] Ho, Jonathan, Jain, Ajay. Denoising Diffusion Probabilistic Models. arXiv preprint, arXiv:2204.06125 (2020).

[3] Xu, Xingqian, Wang, Zhangyang, Zhang, Gong, Wang, Kai, Shi, Humphrey. Versatile Diffusion: Text, Images, and Variations All in One Diffusion Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7754–7765, 2023.

[4] Scotti, Paul S., Tripathy, Mihir, Torrico Villanueva, Cesar K., Kneeland, Reese, Chen, Tong, Narang, Ashutosh, Santhirasegaran, Charan. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. arXiv preprint, arXiv:2403.11207 (2024).

[5] Ozcelik, Furkan, VanRullen, Rufin. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

[6] Takagi, Yu, Nishimoto, Shinji. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, doi: 10.1101/2022.11.18.517004 (2022).

[7] Child, Rewon. Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images. arXiv preprint, arXiv:2011.10650 (2020).

[8] OpenAI. ChatGPT: Language Model for Conversational AI. *OpenAI*, 2023. https://www.openai.com/chatgpt.