BRAID: Brain Representation to Artificial Image via Diffusion

Teo Imoto-Tar
Department of Mathematics
University of California, San Diego
La Jolla, CA, USA
timototar@ucsd.edu

***Abstract*—Reconstructing visual experience from the brain is a compelling and challenging interdisciplinary task that is at the intersection of neuroscience and machine learning. Here we introduce BRAID: Brain Representation to Artificial Image via Diffusion. By leveraging a student-teacher architecture, we avoid the heavy lifting required to train the network and at the same time greatly simplify the data-pipeline. We propose utilizing a pre-trained Latent Diffusion Model (LDM) to train a student fMRI encoder in place of the encoder used by Segmind Stable Diffusion 1B (SSD-1B) Variational Autoencoder (VAE). The student fMRI encoder will then generate latents similar to the original latent encoder. From there, the reconstruction of the visual stimuli is handled by the LDM.**

I. Introduction

A. Research Problem

Our primary concern is to answer the question, how much information can be extracted from the brain? Is there a way to quantify this? We aim to address these questions by reconstructing visual experiences from brain activity. More specifically, we intend to develop a model whereby taking functional Magnetic Resonance Imaging (fMRI) features we can generate images resembling the original stimuli. In both neuroscience and machine learning, translating brain signals and decoding meaningful signals for visual reconstruction is a complex task.

The goal is to explore how much meaningful information we can extract from brain signals, mainly from the primary visual cortex (V1) and higher cortices. By decoding information directly, we hope to open new methods for efficient cross-model generation and more broadly gain insight into how the brain encodes and preserves perceptual experiences.

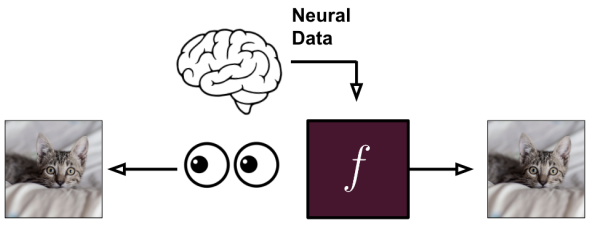

Fig. 1. Intuition for task. We want to decode neural data to reconstruct the original image stimuli seen by the subject. We are tasked with distinguishing an f that allows for this mapping.

B. Importance

The visual system, while studied extensively, still holds many intriguing complexities and remains one of the brain's greatest wonders. Reconstruction of visual stimuli from neuronal activity in an interpretable manner could help open a window into better understanding of cognition and perception. Object decoding is a major leap forward in brain-computer interface (BCI) with limitless potential applications. For individuals with conditions that limit verbal or motor output, such as ALS or severe paralysis, being able to reconstruct what they see or imagine from brain activity could offer a powerful way to communicate.

While the technology is still developing, advances in neural decoding could one day support tools that translate internal experiences into visual or linguistic outputs, providing a more intuitive connection between thought and expression. This project is motivated by recent advances in both state-of-the-art generative models, namely Stable Diffusion, and brain imaging technologies.

C. Related Works

Some notable works that share a similar codebase and approach–utilizing autoencoders and transformers–include MindEye [5], MindEye2 [6], Brain-Streams [4], and NeuroPictor [3], all of which feature fMRI-to-image models designed to reconstruct the original visual stimuli. One of the primary drawbacks in methods expressed in these papers is that they tend to depend on multi-modal inputs, which require more pre-processing and data collection. Although they provide more semantics and surface information for the latent space, this introduces complexities that make it difficult to scale and consequently limits broader applicability. Additionally, many of these models claim to perform well with limited data, but this often requires subject-specific training, which impedes generalization and likewise constrains scalability.

II. Methodology

A. Dataset

We use Natural Scenes Dataset (NSD) [1] , a large-scale 7T (7 Tesla) fMRI dataset containing thousands of natural scene images ($\sim$8.3TB betas and $\sim$37GB stimuli) collected from 8 handpicked subjects. This extensive dataset contains many different modalities of neuro-imaging including anatomical (T1 and T2), Echo-Planar Imaging (EPI), BOLD signals, brain

masks (localized Region of Interest or ROIs), and much more. In particular, we use features extracted from fMRI single trial responses ("betas") using generalized linear model (GLM).

For more stable signals, we use ROI masks to extract specific betas. This way, features extracted consider the Hemodynamic Response as BOLD signals lag behind the active neural signal. Since the stimuli are shown every 4 seconds with only 3 seconds of rest, this leads to overlapping BOLD signals which make decoding difficult. However, by modeling the Hemodynamic Response Function (HRF), the GLM can compensate for these temporal dilations.

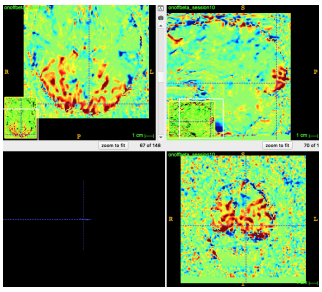

Fig. 2. Example ON–OFF beta map from NSD. This visualization reflects voxelwise activation in response to image presentation across a full session, estimated using a canonical HRF. Bright areas indicate stronger stimulus-driven responses.

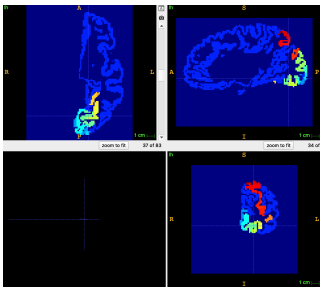

Fig. 3. Example ROI visualization from the NSD dataset viewer. Each hemisphere is independently labeled with integer-coded ROI masks (e.g., V1, V2, V3), provided in multiple functional and anatomical spaces. Shown here is an axial and sagittal view of a left hemisphere ROI volume in func1pt8mm space.

Courtesy of the NSD dataset, `nsd_stimuli.hdf5`, the compressed dataset formatted in Hierarchical Data Format version 5, contains all the image stimuli across all subjects, sessions, and trials. There are 73,000 images in this datafile. Each subject had a total of 30,000 total trials with 40 sessions. Each session consists of 750 trials or image stimuli with ∼10,000 randomly sampled unique images. This means that an image per subject is shown a maximum of three times but not every subject saw every image. The 30,000 total trials were repeated for 8 subjects for a grand total of 240,000 total images.

B. Data Preparation: fMRI preprocessing

The NSD is located on AWS which makes interaction and downloading images from it efficient and scalable. To prepare the dataset for training, all images were extracted from the ∼37GB COCO image dataset stored in Hierarchical Data Format version 5 (HDF5) for organization and allowing for partial I/O. In addition, we also have the .mat which contains the entire experiment design. This aligns the subject, session, and fMRI trials to the actual image stimuli. This file maps the fMRI data to the stimuli images or trials. It is also important in aligning the fMRI data to the latent outputs from the SSD-1B encoder which will later be used as the ground truth to train the fMRI encoder. Using the experimental desin, we obtain the stimulus images based off of the beta single-trial GLM sessions.

Next, we prepare the fMRI feature vectors for the student encoder model. Recall, a single session consists of 750 trials or image stimuli seen by the subject. While 1mm recordings have higher spatial resolution, 1pt8mm GLM recordings are better suited for our task and have broader support as this is the default resolution.

fMRI is a time series of brain volumes or 3D slices over time but in the case of single-trial GLM, it is the averaged recording for that specific stimuli. The fMRI time series is averaged across the exposure of the stimuli becoming single brain volume. The shape of the fMRI data for one session is `(1, 83, 103, 81)` where the 1 represents the single trial with respective X, Y, Z coordinates for the slices. When flattened, the brain volume becomes a 692,469 dimensional vector which as an input for most models is problematic due to its excessively high dimension.

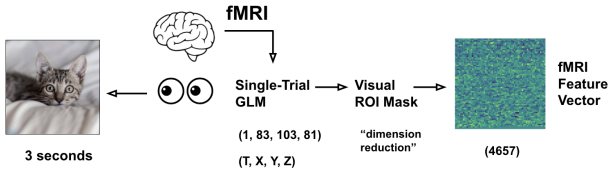

Fig. 4. Intuition for extracting fMRI feature vector. Starting from the left-hand side, the subjects views the image stimulus for 3 seconds. The fMRI is recoded and processing using fMRIPrep to extract the single-trial GLM. We then apply the visual ROI mask which reduces dimension from 692,469 to a concise fMRI feature vector with a shape of 4657. The beta fMRI feature vector is mapped onto a fMRI activation map for the sake of visuals.

We employ visual ROI (region of interest) mask for dimensionality reduction. Through this method, we are able to reduce the ∼700k dimensional vector to a concise and information rich 4647 dimensional fMRI feature vector. The ROI contains the Primary Visual Cortex, Secondary Visual Cortex, Tertiary Visual Cortex, V4, Ventral Occipital Areas, Lateral Occipital Areas, Temporal Occipital Areas, as well as the Intraparietal Sulcus Areas–a combination of lower and higher cortices from the visual system. We organized these feature vectors by their session and subject. Lastly, we converted these feature vectors into pytorch tensors for accelerated training.

With this pre-processed fMRI feature vectors, the input for our encoder, we can now prepare the initial latent used as the ground truth for our encoder model. We achieve this by passing the image stimulus into the original SSD-1B VAE encoder and retrieve the latents.

Finally, all data is prepared for training. The fMRI feature vectors (single-trial GLM betas with ROI mask) as the input to the model and the initial latents from the original SSD-1B VAE encoder as the target values. We will use this data to train the encoder. Due to limited storage, only the first 10 sessions from the first subject were usable as training data for the fMRI encoder.

C. Baseline Pre-Trained Model

A good starting point was to find an efficient and novel image-to-image latent diffusion model. More specifically, recreating

images from initial latents z_0. Initial experiments utilized Stable Diffusion XL (SDXL) which is primarily known for its novel text to image capabilities. For our use case, we want to find a unimodal model that excels specifically at image to image tasks as a baseline. Stable Diffusion models at their core are latent diffusion models that are conditioned primarily on text. While this deviates from their primary application, the Stable Diffusion model is also capable of image reconstruction from image inputs without text conditioning. For speed and optimization, we opted for the Segmind Stable Diffusion with 1 billion parameters (SSD-1B) which is 50% smaller than the original SDXL but offers a 60% speed up in inference which is perfect for our use case [2]. Initial experiments tested image reconstruction results as a sanity check showed that this model was able to capture core features of the original image. From the results of the SSD-1B, we can see reconstructing the original image from the pure latents is viable.

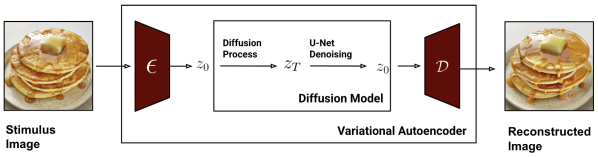

Fig. 5. Overview of the SSD-1B model. From the left-side, we input an image into the encoder which outputs the initial latent. We take this initial latent and pass it through the diffusion model adding noise for T steps and denoising for an additional T steps with U-Net to approximate the initial latent representation. Finally, we pass the latent through the surrounding Variational Autoencoders decoder to reconstruct the original image. In this example, a pancake image stimuli passed into the left-side of the model and is successfully reconstructed on the right-hand side.

Fig. 6. Example images reconstructed using SSD-1B from image stimuli images.

D. Model Overview

In regards to the two main issues raised–1) the use of multi-modal data which is more involved 2) subject-specific training–we aim to address these problems through a cross-modal approach which greatly simplifies the architecture while simultaneously confronting these issues. Furthermore, we propose to leverage knowledge distillation with contrastive learning to help the student encoder to learn and produce meaningful latents, similar to that of the teacher. SSD-1B uses the SDXL VAE which is a ResNet inspired deep Convolutional Neural Network (CNN) with attention and downsampling as the parent encoder. This approach will help encourage the fMRI child encoder to generate meaningful latents from the fMRI feature vector. For our child encoder, we start with a simple baseline MLP model. The purpose being to demonstrate how even relatively simple models paired with complex ones can still produce meaningful output and contribute to a multimodal setup. This method will eliminate the need for auxiliary data modalities which will reduce pipeline complexity while also promoting generalization across subjects through a shared latent space.

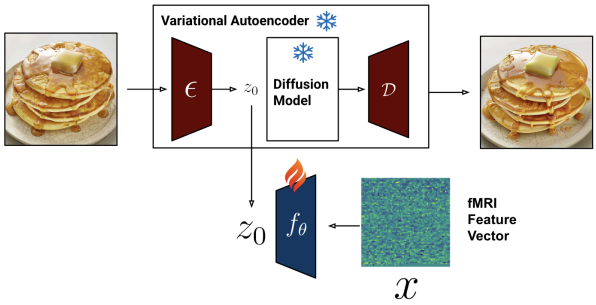

Fig. 7. Training of child fMRI encoder f_θ utilizing knowledge distillation. Using initial latent z_0 from the SSD-1B VAE encoder as the ground truth, we can feed corresponding fMRI feature vectors as input to f_θ for training. As depicted, SSD-1B VAE and diffusion model are frozen.

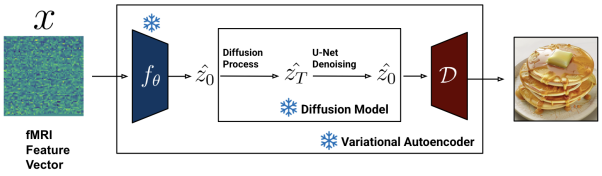

Fig. 8. SSD-1B with fMRI encoder f_θ during inference. We freeze f_θ along with SSD-1B VAE and diffusion model. From the left-side, the fMRI feature fector is fed into f_θ which gives a predicted initial latent z_0. This z_0 is passed through the diffusion model, decoded using the SSD-1B VAE decoder to reconstruct the original image from the fMRI feature vector.

E. Mathematical Formulation: Latent Diffusion

We inject an fMRI-predicted latent into a pre-trained LDM. Let $f_\theta : \mathbb{R}^{D_{\text{fMRI}}} \to \mathbb{R}^k$ be the learned encoder that maps the fMRI feature vector x to a latent code $\hat{z}_0 = f_\theta(x)$. Unlike the original LDM we do not begin the denoising process from pure Gaussian noise. We instead treat $\hat{z}_0$ as a reference latent and run our image-to-image variation (SDXL img2img pipeline) of the sampler.

$$z_{t_0} = \sqrt{\bar{\alpha}_{t_0}}\, \hat{z}_0 + \sqrt{1 - \bar{\alpha}_{t_0}}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

Here $t_0 \in \{0, \ldots, T\}$ is strength, a user-defined parameter $\lambda \in [0, 1]$ (with $\lambda = 0$ meaning no extra noise): $t_0 = \lfloor \lambda T \rfloor$. From z_{t_0} we apply the pre-trained noise predictor $\epsilon_\phi(z_t, t)$ for all remaining timesteps $t = t_0, \ldots, 1$:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\phi(z_t, t)\Big) + \sigma_t\,\eta, \qquad \eta \sim \mathcal{N}(0, \mathbf{I}).$$

The final latent z_0^{out} is then decoded with the frozen VAE decoder:

$$\hat{I} = \mathcal{D}_{\text{img}}(z_0^{\text{out}}).$$

During training only f_θ is updated, using an L2 loss in latent space:

$$\mathcal{L}_{\text{enc}} = \|f_\theta(x) - z_0\|_2^2,$$

while all LDM and VAE parameters ϕ remain fixed. This will allow us to leverage the rich image prior of a large-scale generative model without the need for additional diffusion training.

III. Training Objectives

A. Novelty and Significance

The novelty of this work lies in the proposed solution to the problem, particularly the use of cross-modal architecture as opposed to the conventional multi-modal inputs. In multi-modal approaches, multiple modalities (text, audio, and images) are processed by the model during training and inference. The goal of this method is for each modality to contribute distinct, complementary information, thereby enhancing model accuracy. However, it also introduces added complexity in data handling and model design.

In contrast, cross-modal architectures involve learning representations across different modalities - such as text to generate images - effectively transferring knowledge from one modality to another. This powerful framework functions without the need for auxiliary data modalities making the model more direct. While the use of knowledge distillation for this type of task is not revolutionary, the effect–the elimination of multi-modal data and generalization as a result-is where the significance of this solution lies.

B. Training and Setup

We employ mean squared error (MSE) as the loss function to train the fMRI encoder f_θ. We only had the capacity for 10 sessions for one subject; this amounted to 7,500 samples for training. Due to the limited sample size the training and testing we decided to split out dataset 90/5/5 for training, testing, and validating. We utilized a 2 layer MLP that maps the 4,657 dimensional fMRI feature vectors to the 16,384 dimensional SSD-1B latent (using the shape of the SDXL VAE encoder output). We used the Adam optimizer with a learning rate of $1e-4$ and trained the MLP for 200 epochs. The training loss is depicted below and results are described in the next section.

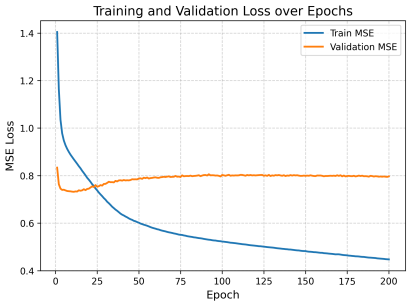

Fig. 9. MLP training over 200 epochs. The training loss steadily decreases while the validation loss initially dips, increases, and then gradually levels out.

IV. Results

Using the simple fMRI MLP encoder, we were able to get promising results. We evaluated the model on samples from both the training and testing datasets for inference.

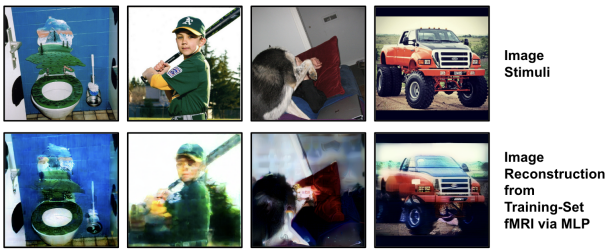

Fig. 10. Image reconstruction from training-set fMRI feature vectors. Here we used the frozen MLP fMRI encoder along with the the rest of the SSD-1B frozen diffusion model and decoder for inference. This demonstrates that the MLP is able to successful encode fMRI features into a meaningful latent space allowing the key features from the original image stimuli to be reconstructed.

Although, the BRAID model was able to accurately re-create the key features of the original stimuli, sampling from the testing dataset was an entirely different story and unfortunately will require further investigation in the future.

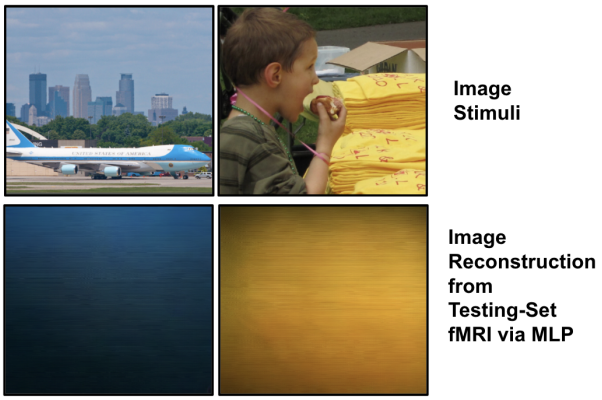

Fig. 11. Image reconstruction from testing-set fMRI feature vectors. Here we see that the reconstructed image is noise. However, the model is able to capture some semantics such as color.

V. Future Work

This approach of knowledge distillation utilizing powerful pre-trained models and training simple lightweight, adapting them effectively for specific tasks is promising. Although we were only able to experiment with the MLP fMRI encoder, we intend to test out different encoders in the future. We have briefly experimented with simple 1D Convolution networks and Transformer models but do not have yet results to share. A major bottleneck in this project was the limited amount of data samples for training which most likely led to the model overfitting to a degree. This is an issue that can be easily overcome with more storage and efficient IO of data. Overall, we intend to 1) gather more training data 2) try different encoder models and compare their performance.

Code Availability

All code and implementation details for this project is available at: https://github.com/teooi/BRAID

References

[1] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, J. B. Hutchinson, T. Naselaris, and K. N. Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25:116–126, 2022.

[2] Y. Gupta, V. V. Jaddipal, H. Prabhala, S. Paul, and P. V. Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss, 2024.

[3] J. Huo, Y. Wang, X. Qian, Y. Wang, C. Li, J. Feng, and Y. Fu. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. *arXiv preprint arXiv:2403.18211*, 2024.

[4] J. Joo, T. Jeong, and S. Hwang. Brain-streams: fmri-to-image reconstruction with multi-modal guidance. *arXiv preprint arXiv:2409.12099*, 2024.

[5] P. S. Scotti, A. Banerjee, J. Goode, S. Shabalin, A. Nguyen, E. Cohen, A. J. Dempster, N. Verlinde, E. Yundler, D. Weisberg, K. A. Norman, and T. M. Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36, 2023.

[6] P. S. Scotti, M. Tripathy, C. K. Torrico Villanueva, R. Kneeland, T. Chen, A. Narang, C. Santhirasegaran, J. Xu, T. Naselaris, K. A. Norman, and T. M. Abraham. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.