Visual Decoding from EEG

A Project Report Submitted by

Atharv Kumar (B21038)

Yashika Gupta (B21174)

in partial fulfilment of the requirements for the award of the degree of

BTech in Data Science and Engineering

Indian Institute of Technology Mandi

School of Computing and Electrical Engineering

December, 2024

Declaration

I hereby declare that the work presented in this Project Report titled Visual Decoding from EEG submitted to the Indian Institute of Technology Mandi in partial fulfilment of the requirements for the award of the degree of BTech in Data Science and Engineering, is a bonafide record of the research work carried out under the supervision of Prof.Arnav Bhaskar and Prof.Padmanabhan. The contents of this Project Report in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

Signature

Atharv Kumar (B21038)

Yashika Gupta (B21174)

Certificate

This is to certify that the Project Report titled Visual Decoding from EEG, submitted by Atharv Kumar(B21038) and Yashika Gupta(B21174) to the Indian Institute of Technology Mandi for the award of the degree of BTech in Data Science and Engineering, is a bonafide record of the research work done by him under my supervision. To the best of my knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Signature

Prof.Arnav Bhaskar

Prof.Padmanabhan

Acknowledgements

We would like to extend our gratitude to all those who contributed to the successful completion of this project "Visual Decoding from EEG."

First and foremost, We would like to thank our supervisor Prof. Arnav Bhaskar and Prof. Padmanabhan for giving us a chance to work on this project. Their constructive feed-back and guidance throughout the project have been invaluable and have played a very important role in the successful completion of the project. Further we would also like to extend our heartfelt thanks to our mentor Dr.Jyoti Nigam for her unwavering support, guidance and feedback throughout the research and completion of the project. Their expertise and mentorship were crucial in shaping the direction of the outcomes of the work.

We are also grateful to IIT Mandi and School of Computing and Electrical Engineering at IIT Mandi for providing us a chance to work on this project and providing access to necessary resources and infrastructure, enabling us to explore more about advanced technologies like stable diffusion and use them in our project.

We would also like to acknowledge the authors of the base papers which we have followed "DreamDiffusion: Generating High-Quality Images from Brain EEG Signals" whose work and methodology inspired us in many aspects of this project. Their contributions laid the base for exploring EEG-to-image reconstruction tasks.

Lastly we are very much thankful to our friends for their support and encouragement throughout the project.

Thank you all for the successful completion of the project !!

Abstract

The project tries to solve the challenging task of visual decoding from EEG signals, aiming to reconstruct images as perceived by the brain using EEG (with high temporal resolution) and deep learning techniques. We have worked on a dataset which comprise of EEG Signals, corresponding images and labels. We have developed a comprehensive pipeline. EEG signals were encoded into embeddings using a Variational Auto-Encoder (VAE). Image features were extracted in two forms: depth maps obtained via MiDaS and textual descriptions generated using the BLIP model. The textual data was further transformed into embeddings using a VAE. To align EEG and text embeddings in a shared latent space, we employed the CLIP model, yielding optimized embeddings. Further we send these embeddings along with a noisy image to the stable diffusion model to reconstruct the original image. We have used metrics like cosine similarity to check the quality of the optimized embeddings and SSIM score to check the quality of the reconstructed image.

The proposed framework not only reconstructs images from EEG signals, but also provide insights into how different modalities like textual prompts, depth maps and EEG embeddings contribute to visual decoding.

We conducted experiments with a pre-trained Stable Diffusion model to test different combinations of inputs, including noisy images, depth maps, and textual prompts. Results revealed that textual descriptions from BLIP served as the most effective prompt for image reconstruction. Finally, optimized EEG embeddings were projected as prompts into the Stable Diffusion pipeline alongside noisy images to achieve reconstruction. Future work aims to integrate depth map embeddings aligned with EEG signals for further enhancement. This methodology demonstrates a novel approach to decoding visual information from neural activity, potentially advancing applications in neuroscience and assistive technologies.

Contents

List of Figures

List of Tables

Visual Decoding from EEG

1 Introduction and background

The human brain encodes a very large amount of information about the world, and decoding this information will give insights into cognitive processes and neural representations. Among various brain-computer interface technologies, electroencephalography (EEG) has emerged as a practical tool due to its non-invasive nature and temporal resolution. EEG signals represent brain activity in real-time, making them a valuable resource for exploring how humans perceive, process, and interpret visual information. However, translating EEG signals into interpretable formats, such as reconstructed images, remains a challenging problem.

Some recent works, such as MinD-Vis attempt to reconstruct visual information based on fMRI (functional Magnetic Resonance Imaging) signals, which is another way to measure brain activities. They have demonstrated the feasibility of reconstructing high-quality results from brain activities. However, they are still far away from our goal of using brain signals to create conveniently and efficiently. 1) Since fMRI equipment is not portable and needs to be operated by professionals, it is difficult to capture fMRI signals. 2) The cost of fMRI acquisition is high. They greatly hinder the widespread use of this method in the practical artistic generation. In contrast, EEG (electroencephalogram) is a non-invasive and low-cost method of recording electrical activity in the brain. Portable commercial products are now available for the convenient acquisition of EEG signals, showing great potential for future art generation.

This project "Visual Decoding from EEG" aims to bridge the gap between neural signals and visual representations. Leveraging the advanced deep learning models like Variational Autoencoders (VAE), BLIP, CLIP and Stable Diffusion, this work explores an innovative approach to reconstructing images solely from EEG data. The methodology combines embeddings extracted from EEG, textual descriptions and depth maps. We have created a comprehensive pipelines for cross-modal alignment and image reconstruction. This project have potential applications in neuroscience and human-computer interaction.

Background

EEG-based image reconstruction has gained significant attention in recent years, driven by advancements in machine learning and multimodal integration.

Traditional methods rely on fMRI-image paired data to train the model to predict image features from fMRI. These images are fed into GANs for stimulus reconstruction during testing. However recent studies have proposed unsupervised approaches, such as a reconfigurable autoencoder designs. The drawback of using the fMRI signals is that fMRI equipment is not portable and needs to be operated by professionals. Secondly cost of fMRI acquisition is high and also they have high spatial resolution but low temporal resolution.

Generating images from EEG has also been explored using deep learning techniques like Brain2Image, ThoughtViz etc. Overall these approaches demonstrate the potential of using brain signals to generate

images and advance the field of brain-computer interfaces.

State-of-the-art methods, such as Dream Diffusion and BrainVis, have demonstrated that aligning EEG embeddings with visual or textual features can facilitate image generation. For instance, guided diffusion models have shown the ability to map neural signals to both coarse and fine-grained image features, while text-image alignment frameworks like CLIP enable semantic understanding of EEG embeddings. Despite these advances, challenges remain in capturing the rich temporal and frequency-domain information of EEG signals and aligning them effectively with visual features

This project builds on the foundational ideas of EEG-based multimodal learning, integrating insights from recent research. The use of BLIP for generating textual captions from images and MiDaS for extracting depth information represents innovative steps toward improving reconstruction fidelity. Additionally, leveraging the capabilities of Stable Diffusion, a powerful generative model, offers the potential to decode not only semantic but also stylistic aspects of visual stimuli.

We seek to push the boundaries of visual decoding from brain activity by combining EEG embeddings with textual and depth map features. The project tries to offer a robust framework for reconstructing images from neural signals with improved accuracy.

2 Problem definition and Objective

The problem of visualizing the brain activity is a a long standing challenge in neuro-science and artificial intelligence. Despite significant advances in brain-computer interfaces reconstructing images from non-invasive EEG signals remains difficult due to several inherent limitations like Low Spatial Resolution, High dimensionality, Cross Modal Alignment and Generative complexity. Traditional methods use fMRI in-place of EEG signals for brain computer interaction but fMRI faces several issue like 1) fMRI equipment is not portable and needs to be operated by professionals, it is difficult to capture fMRI signals. 2) The cost of fMRI acquisition is high. They greatly hinder the widespread use of this method in the practical artistic generation. In contrast, EEG (electroencephalogram) is a non-invasive and low-cost method of recording electrical activity in the brain. Portable commercial products are now available for the convenient acquisition of EEG signals. Thus using EEG signals for visual decoding gives us a robust and portable solution to brain-computer interfaces.

Objectives

The objectives of our project can be summed down to major three :

Develop a Visual Encoding Pipeline Using EEG Signals: Create a robust framework that can extract meaningful features from EEG signals and use them to represent and understand visual stimuli effectively.

Integrate Multi-Modal Techniques for Image Reconstruction: Leverage image captioning (using BLIP) and image generation (using Stable Diffusion) to reconstruct high-quality images based on brain activity data.

Enable Direct "Thought-to-Image" Conversion: Achieve seamless image generation from EEG signals without relying on textual intermediates, demonstrating the potential of brain activity-driven visual content creation.

3 Methodology

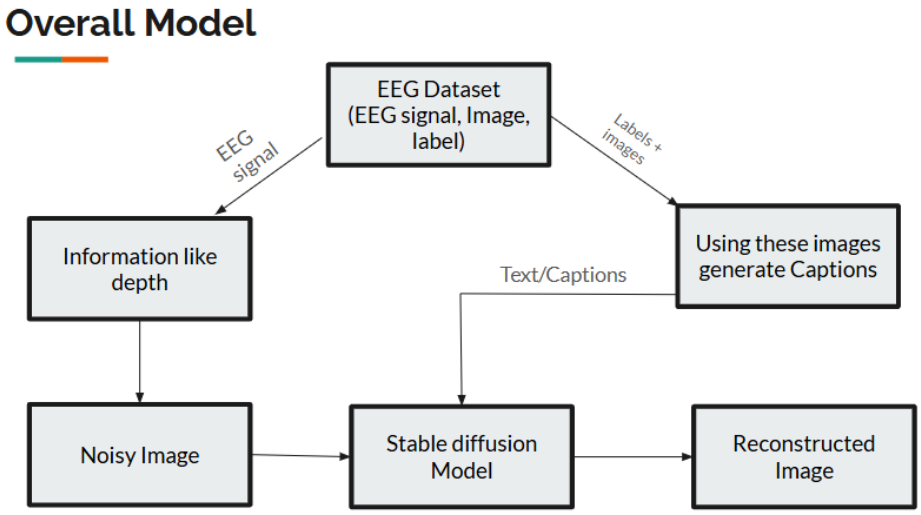

Figure 3.1: Overall Model Followed (note that at testing time we have only EEG signals)

The methodology for the project "Visual Decoding from EEG" involves a multi-modal approach combining EEG signals, depth maps, and textual embeddings to reconstruct images (decode visual stimuli using EEG). The whole process is divided into two organized pipelines one pipeline being extracting depth from the EEG (EEG to Depth) where we bring the depth map embeddings and the EEG embeddings closer. The second pipeline involves caption generation from the images and then get the text embeddings and then bring both the text and the EEG embeddings closer. Later both the pipelines come together by using a projection layer and are fed as a input into a stable diffusion model which produces the reconstructed image and learns to reconstruct image based on modified EEG embeddings. Thus in the testing phase we can take EEG embeddings, modify it and then put forward that modified embeddings in the stable diffusion model to get the reconstructed image, based on the EEG data. (visual decoding using EEG).

The whole process is structured into the following steps:

1. Data Acquisition and Pre-processing

The dataset consists of EEG signals, corresponding visual images, and labels. Each data modality undergoes specific preprocessing:

EEG Signals: Raw EEG data.The structure of the data is as follows : We have 11965 samples, with 128 channels and 491 time-points.

Images: Images corresponding to the EEG signals are resized and normalized to match the input requirements of various models.

Labels: Textual descriptions and categories serve as supplementary data for multi-modal alignment.

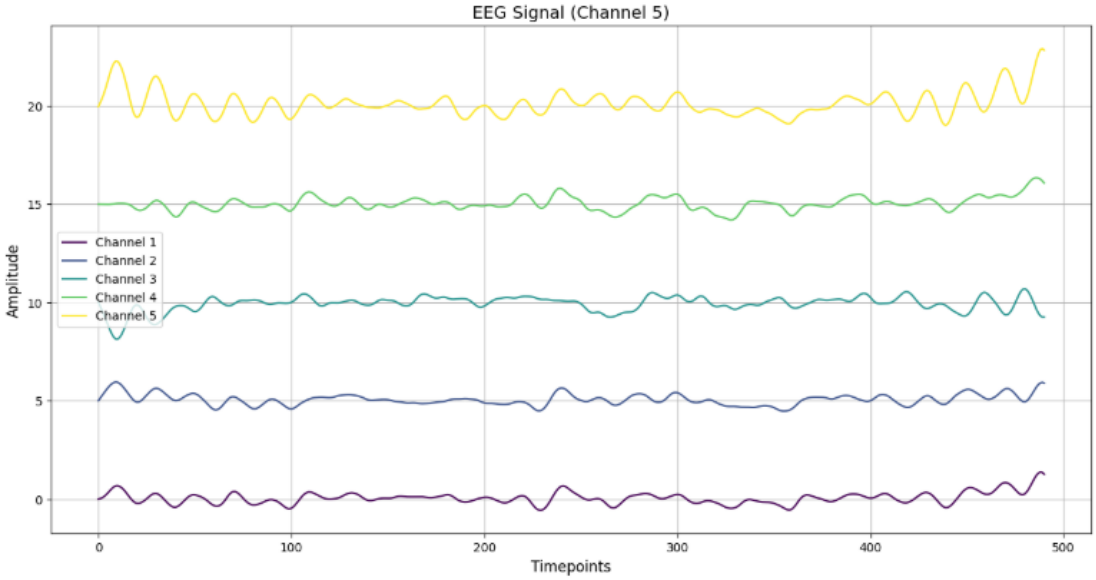

Figure 3.2: EEG data (plotting only 5 out of 128 channels)

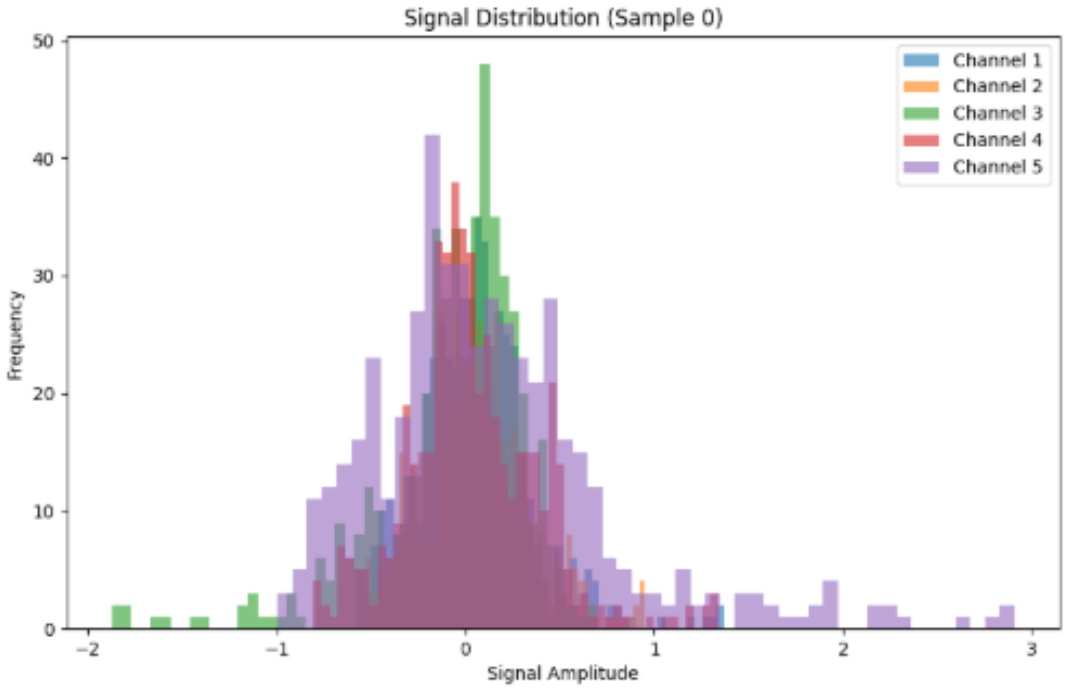

Figure 3.3: Signal Distribution of EEG signal considering 5 out of 128 channels

2. EEG Embedding Extraction

We have used Variational Auto-Encoder to convert the EEG signals into embeddings.A VAE model encodes EEG signals into latent embeddings, capturing essential neural features while reducing dimensionality. This step transforms the temporal and spatial complexity of EEG data into a compact and meaningful representation.

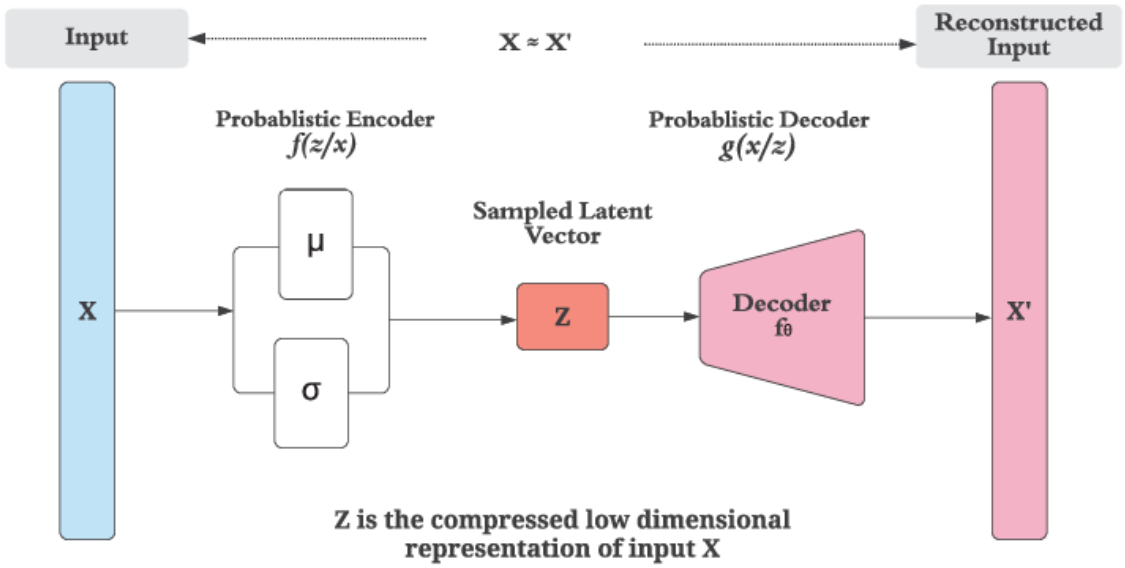

Figure 3.4: Architecture of Variational Autoencoder (VAE) to extract embeddings

3. Multi-Modal Feature Extraction

Depth Map Generation: We have used MiDaS, a depth estimation model to extract depth maps from images. These depth maps provide spatial information about the visual stimuli and are visualized alongside the original images and validate correspondence.

Next we extract the embeddings from the depth map using Variational Autoencoder (VAE) and then tried to bring the EEG embeddings and Depth Map embeddings closer and get the modified EEG embeddings.

Caption Generation: The BLIP model generated textual descriptions of the images. These captions add semantic context, enriching the representation of visual information. BLIP: Bootstrapping Language Image Pre-training is used for unified vision-language understanding and generation. BLIP is a new VLP framework which enables a wider range of downstream tasks than existing methods.

Data perspective: most state-of-the-art methods (e.g., CLIP (Radford et al., 2021), ALBEF (Li et al., 2021a), SimVLM (Wang et al., 2021)) pre-train on image-text pairs collected from the web. Despite the performance gain obtained by scaling up the dataset, BLIP paper shows that the noisy web text is

suboptimal for vision-language learning.

It introduces two contributions from the model and data perspective :

(a) **Multimodal mixture of Encoder-Decoder (MED):** a new model architecture for effective multi-task pre-training and flexible transfer learning. An MED can operate either as a unimodal encoder, or an image-grounded text encoder, or an image-grounded text decoder. The model is jointly pre-trained with three vision-language objectives: image text contrastive learning, image-text matching, and image conditioned language modeling.

(b) **Captioning and Filtering (CapFilt):** a new dataset boostrapping method for learning from noisy image-text pairs. We finetune a pre-trained MED into two modules: a captioner to produce synthetic captions given web images, and a filter to remove noisy captions from both the original web texts and the synthetic texts.

BLIP achieves state-of-the-art performance on a wide range of vision-language tasks, including image-text retrieval, image captioning, visual question answering, visual reasoning, and visual dialog. We also achieve state-of the-art zero-shot performance when directly transferring our models to two video-language tasks: text-to-video retrieval and videoQA.

Therefore we are using BLIP model, because of its improved accuracies and state of the art results to generate captions from the Corresponding Images of the EEG.

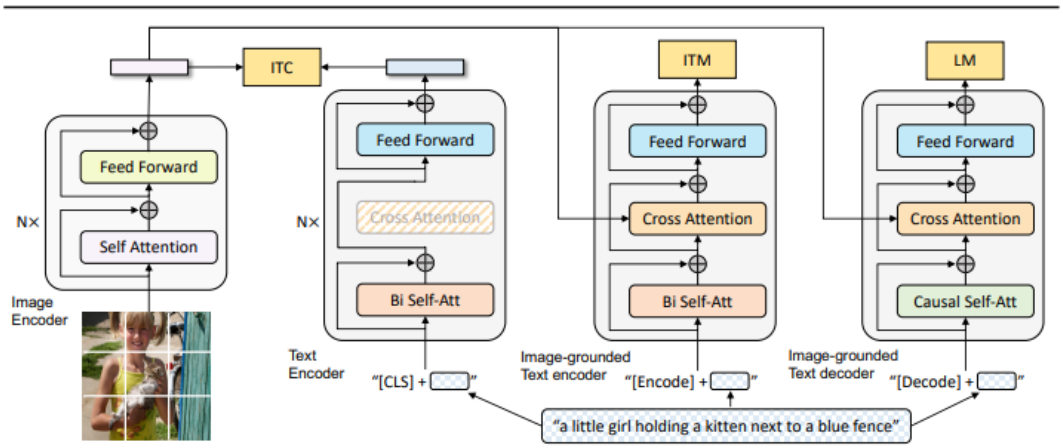

Figure 3.5: Pre-training model architecture and objectives of BLIP (same parameters have the same color). We propose multimodal mixture of encoder-decoder, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations. (2) Image-grounded text encoder uses additional cross-attention layers to model vision-language interactions, and is trained with a image-text matching (ITM) loss to distinguish between positive and negative image-text pairs. (3) Image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers, and shares the same cross-attention layers and feed forward networks as the encoder. The decoder is trained with a language modeling (LM) loss to generate captions given images.

Text Embedding Conversion Captions are converted into embeddings using another VAE, aligning the textual data in a latent space comparable to EEG embeddings.

4. Cross Modal Alignment

The EEG embeddings and the text embeddings are bought into a common latent space and both the embeddings are bought closer using the CLIP model. This alignment ensures that embeddings derived from EEG Signals correspond to meaningful textual features, facilitating effective multimodal integration.

Optimized embeddings are derived from the alignment process.

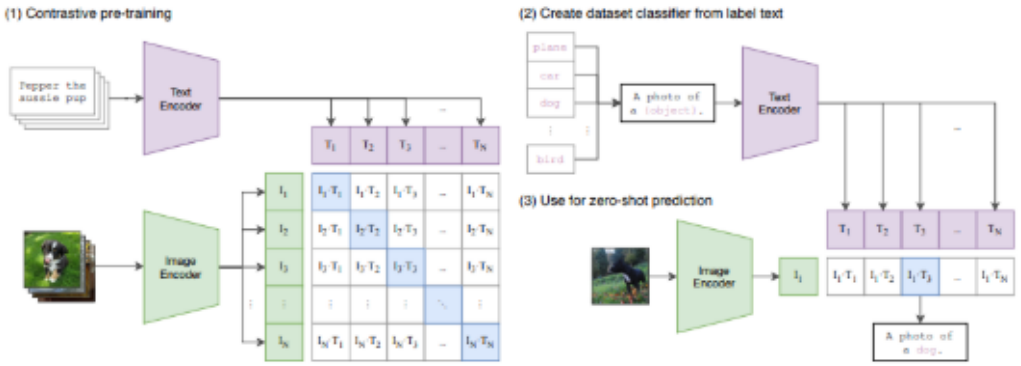

Figure 3.6: CLIP architecture to bring both EEG and Text Embeddings closer

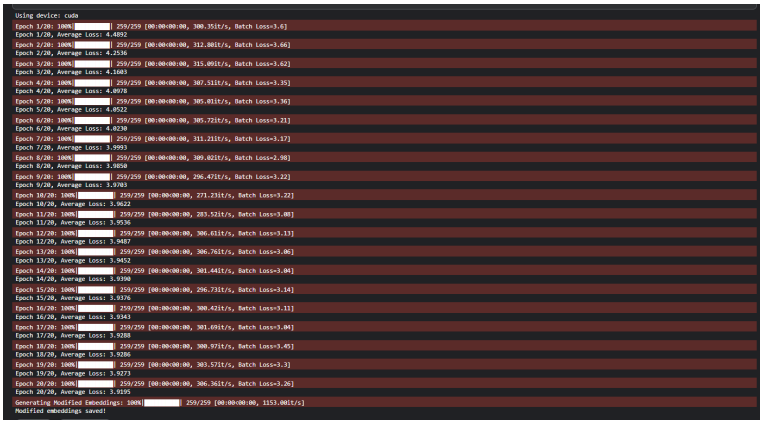

Figure 3.7: Figure shows the average loss of the CLIP decreasing with every epochs indicating the EEG embeddings and the Text embeddings are bought closer

5. Image Reconstruction with Stable Diffusion

We divided the step of Stable Diffusion into two phases:

1. Experimental Phase: Multiple experiments are conducted to evaluate the reconstruction capabilities of the stable diffusion model and the combination of inputs which generate the best result from the stable diffusion model.

Noisy Image Input: Inputs include noisy images paired with text, labels, or empty prompts.
Depth Map Input: Depth maps are paired with text, labels, or empty prompts. These experiments reveal that BLIP-generated captions, when used as prompts, produce the most accurate reconstructions.

2. Final Reconstruction Pipeline: Optimized EEG embeddings are passed through a projection layer to serve as prompts. These prompts , combined with noisy images are fed into the stable diffusion model to reconstruct the target images. (Visual decoding using EEG)

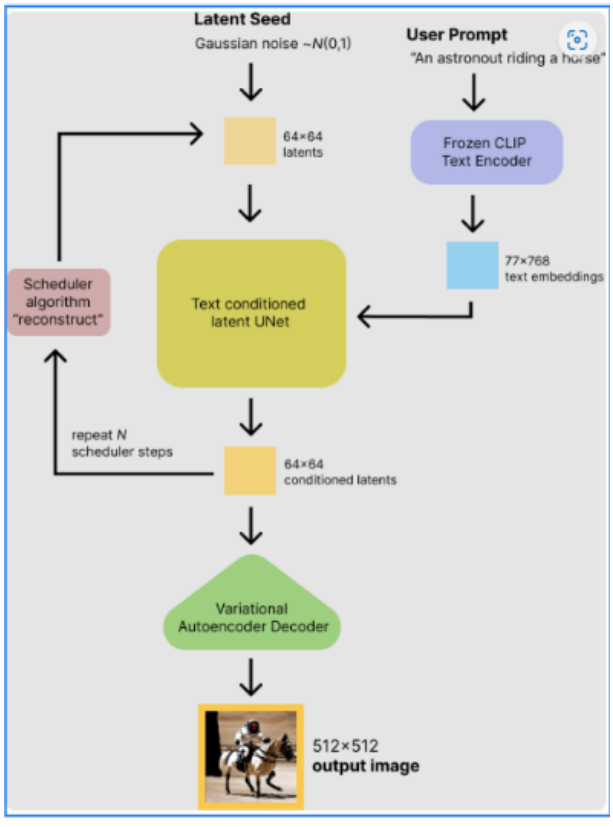

Figure 3.8: CompVis/val-2 stable diffusion model architecture to reconstruct the image

6. Future Methodology Enhancements

Depth Map and EEG Alignment: We align depth map embeddings with EEG embeddings in the shared latent space to explore their joint contribution to image reconstruction.

Testing Phase: During testing, the model is expected to reconstruct images solely from EEG embeddings, leveraging their alignment with textual and spatial information.

7. Evaluation metrics

Structural Similarity Index(SSIM): We use SSIM evaluation metric as a measure of the visual similarity between reconstructed and original images.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

with:

- μ_x the average of x;
- μ_y the average of y;
- σ_x^2 the variance of x;
- σ_y^2 the variance of y;
- σ_{xy} the covariance of x and y;
- $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ two variables to stabilize the division with weak denominator;
- L the dynamic range of the pixel-values (typically this is $2^{\#bits\ per\ pixel} - 1$);
- $k_1 = 0.01$ and $k_2 = 0.03$ by default.

Figure 3.9: SSIM (Structural Similarity) Formula and Significance

4 Theoretical/Numerical/Experimental findings

Experimental phase:

Output 1: Given the generated caption and the Image with added noise to the stable diffusion model

SSIM : 0.4418

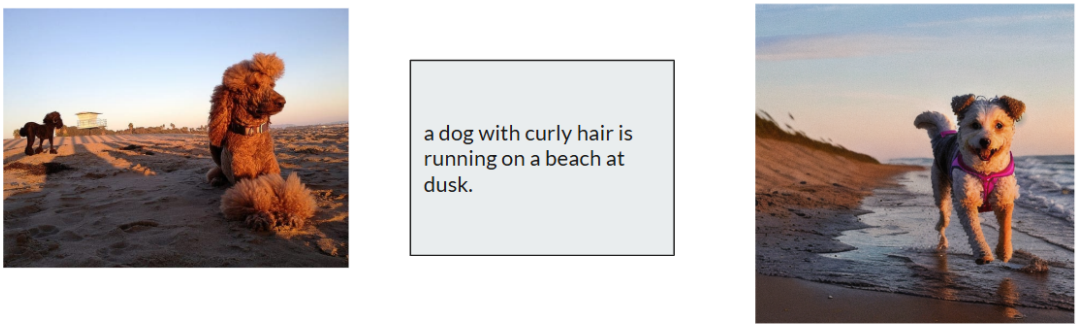

Figure 4.1: Reconstructed image when generated caption and Image with added noise is given

Experimentation : Table shows the SSIM values corresponding to the various types of prompts given with the noisy image to the diffusion model

Input - 1	**Input - 2**	**SSIM value (in percentage)**
Noisy Image	Prompt = ""	10.312
Noisy Image	Prompt = Class Label	15.55
Noisy Image	Prompt = BLIP Caption	40.32

Table 4.1: Experimentation Table for testing different combinations of prompts and noisy image

Results of the above Experimentation

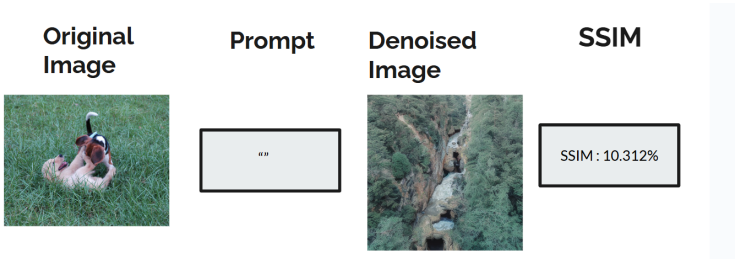

Figure 4.2: Reconstructed image when noisy image and prompt = ””

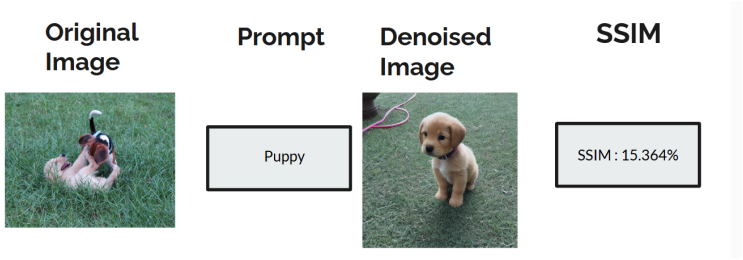

Figure 4.3: Reconstructed image when noisy image and prompt = Class Label

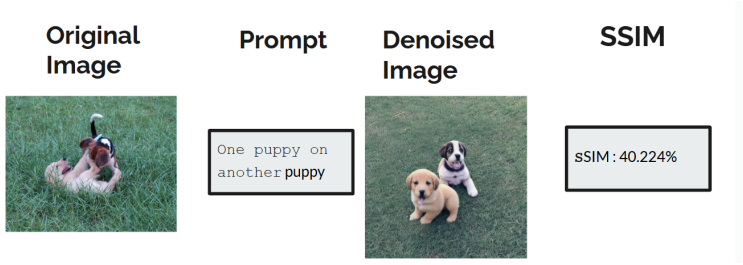

Figure 4.4: Reconstructed image when noisy image and prompt = BLIP Caption

Next We have used MiDAS model to convert the corresponding images of the EEG into depth map. Here are the findings of the same with the image and its corresponding depth maps.

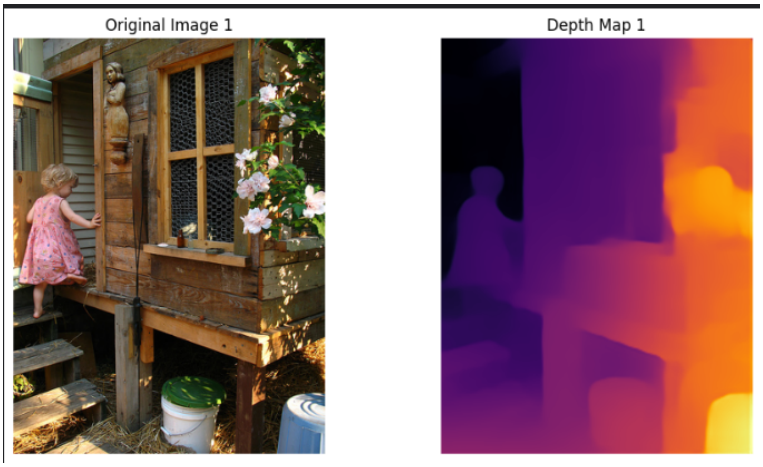

Figure 4.5: Original Image with its corresponding depth map-1

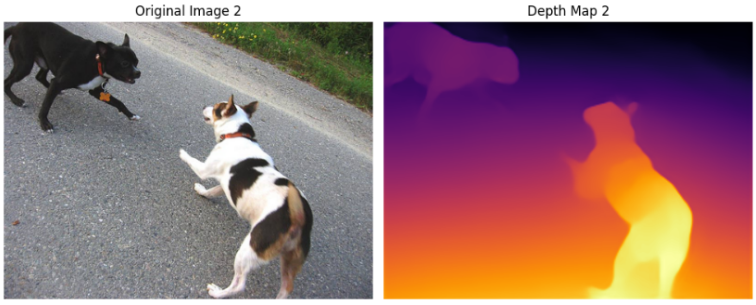

Figure 4.6: Original Image with its corresponding depth map-2

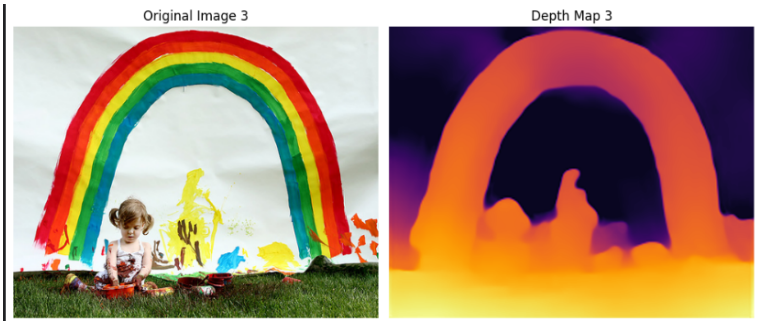

Figure 4.7: Original Image with its corresponding depth map-3

Experimentation 2: Table shows the SSIM values corresponding to the various types of prompts given with the depth map to the diffusion model

Input - 1	Input - 2	SSIM value (in percentage)
Depth-Map	Prompt = ""	15.56
Depth-Map	Prompt = Class Label	23.04
Depth-Map	Prompt = BLIP Caption	41.71

Table 4.2: Experimentation Table for testing different combinations of prompts and Depth Map

Results of the above Experimentation

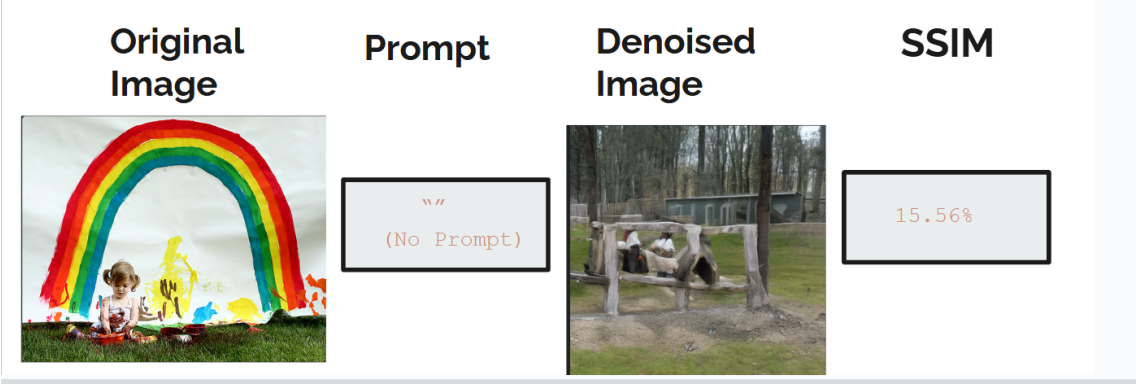

Figure 4.8: Reconstructed image when Depth Map and prompt = ""

Figure 4.9: Reconstructed image when Depth Map and prompt = Class Label

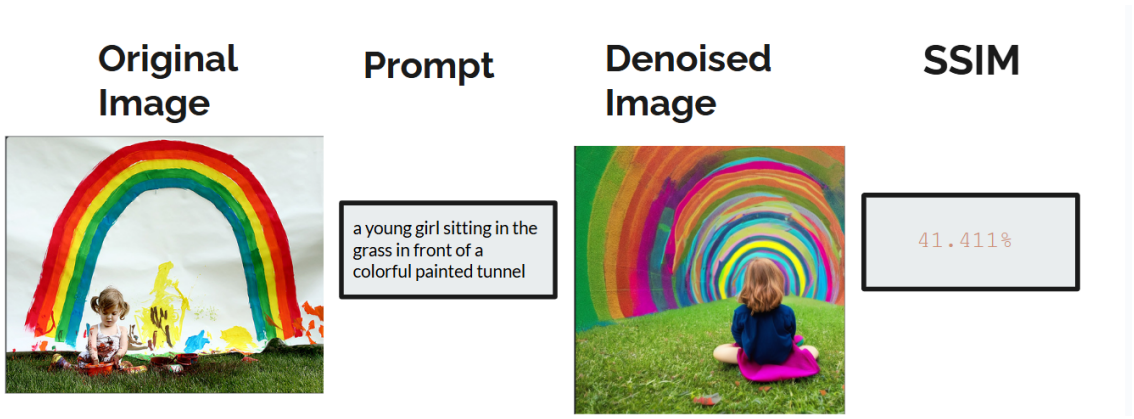

Figure 4.10: Reconstructed image when Depth Map and prompt = BLIP Caption

Final Reconstruction Pipeline

Optimized EEG embeddings are passed through a projection layer to serve as prompts. These prompts, combined with noisy images, are fed into the Stable Diffusion model to reconstruct the target images.

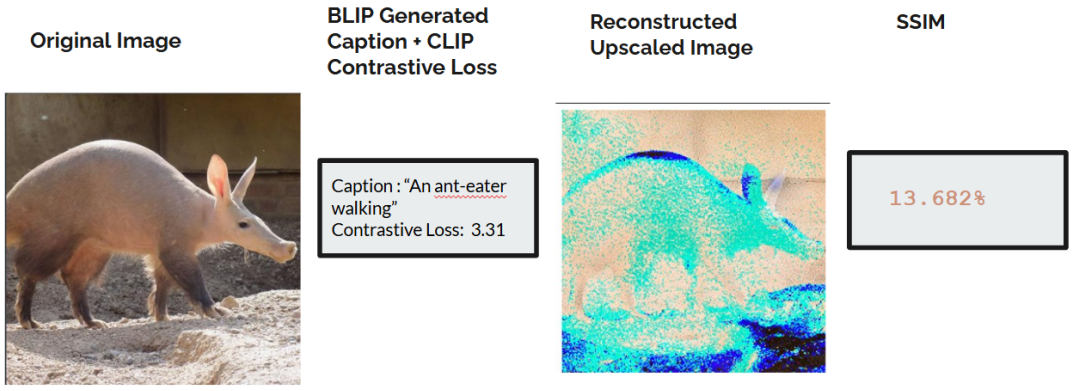

Figure 4.11: Reconstructed image when EEG Embeddings and Noisy Image is given - 1

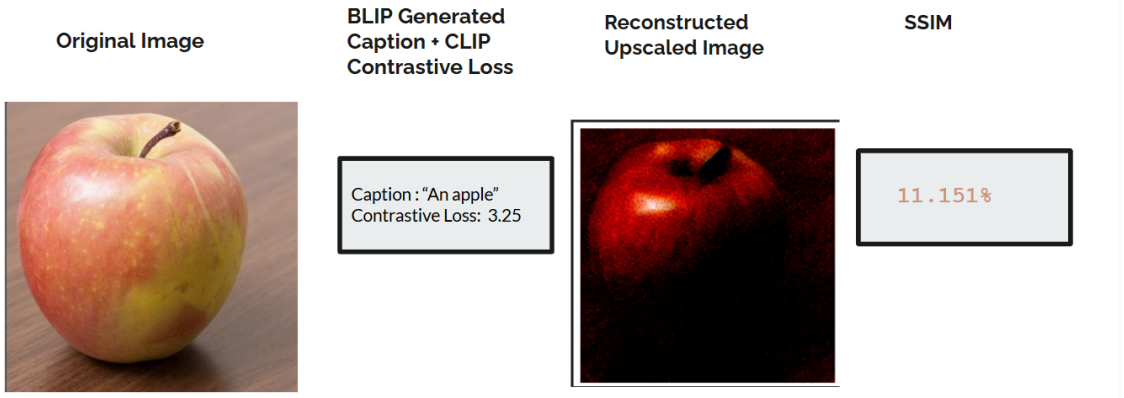

Figure 4.12: Reconstructed image when EEG Embeddings and Noisy Image is given - 2

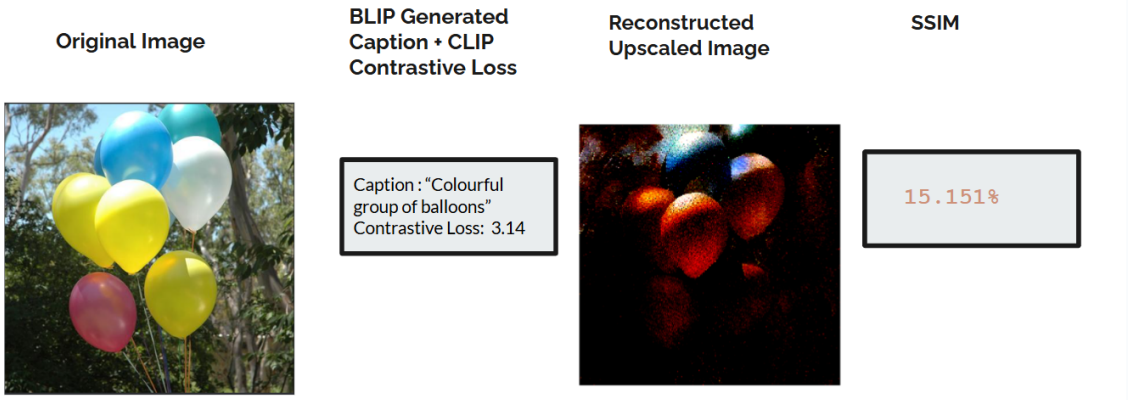

Figure 4.13: Reconstructed image when EEG Embeddings and Noisy Image is given - 3

5 Summary and Future plan of work

This project, "Visual Decoding from EEG," explored a novel pipeline for reconstructing visual stimuli from EEG signals by leveraging deep learning and generative models. The methodology incorporated multiple stages:

1. Extraction of EEG embeddings using Variational Autoencoders (VAEs).

2. Integration of multimodal data, including depth maps and textual captions, to enhance visual representation.

3. Cross-modal alignment of EEG and textual embeddings using the CLIP model.

4. Image reconstruction using the Stable Diffusion model, with experiments revealing that BLIP-generated textual captions provided the most accurate results when used as prompts.

The final step involved passing optimized EEG embeddings through a projection layer to act as prompts for the Stable Diffusion model, reconstructing images using only EEG signals. This work demonstrated the feasibility of aligning neural signals with visual and semantic modalities, paving the way for advancements in Brain-Computer Interface systems.

Future Plan of Work

1. Depth Map Integration : Align depth map embeddings with EEG signals in the shared latent space and then evaluate the combined contribution of depth and textual features to reconstruction of image.

2. Imporove Results: Try to improve the accuracy by using more advanced stable diffusion models and experimenting out other methods to improve upon the SSIM value and getting better results and better reconstructed image.

3. Real-World Applications: Develop practical applications, such as brain-computer interfaces for visually impaired users, and integrate the model into neuro-feedback systems.

References

[1] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. Proceedings of the 39th International Conference on Machine Learning, PMLR 162, 2022. Available: `https://github.com/salesforce/BLIP`.

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. Proceedings of the 40th International Conference on Machine Learning, PMLR 202, 2023. Available: `https://github.com/salesforce/LAVIS/tree/main/projects/blip2`.

[3] Yunpeng Bai, Xintao Wang, Yan-Pei Cao, Yixiao Ge, Chun Yuan, Ying Shan *Dream Diffusion: Leveraging Guided Diffusion for Visual Decoding from EEG Signals*.

[4] Simone Plazzo, Concetto Spampinnato, Issac Kavasidis, Daniela Gaierdao *Decoding Brain Representations by Multimodal Alignment of Neural Activity and Visual Features*.