BrainDiffusion: Reconstructing Visual Semantics from Non-Invasive Neural Activity Readings

Rajath Rao*[1], Lei Zhou†[1], and Dimitris Samaras‡[1]

[1]Stony Brook University

October 14, 2024

Abstract

This paper introduces BrainDiffusion, a novel framework for reconstructing visual semantics from non-invasive EEG signals. By leveraging deep learning techniques, particularly masked autoencoders and contrastive learning, we propose a method for aligning neural activity with visual stimuli. The model uses a two-stage training approach: pre-training an EEG encoder to extract semantic representations and fine-tuning it alongside Stable Diffusion for image generation. Despite the inherent noise in EEG signals and their subjective nature, BrainDiffusion demonstrates promising results in producing images that resemble perceived visual content. Through experiments with various loss criteria, including cosine similarity and supervised contrastive learning, we explore embedding-space alignment and achieve improved image generation accuracy. These results mark a significant step toward bridging the gap between abstract neural representations and concrete visual outputs, with potential applications in communication aids and virtual reality. Challenges such as data noise, limited batch sizes, and computational constraints are also discussed.

*rajath.rao@stonybrook.edu

†Advisor, lei.zhou@stonybrook.edu

‡Advisor, samaras@cs.stonybrook.edu

1 Introduction

The mind stands as one of the most intricate biological entities, capable of producing the most extraordinary thoughts, emotions, and images. Throughout history, the idea of directly translating mental imagery and thoughts into tangible visual representations has always felt like science-fiction for most.

The central objective of this endeavor is to contribute knowledge in bridging the gap between the abstract domain of thoughts and the tangible realm of images. By implementing and studying different deep learning techniques for deciphering the neural patterns underlying mental imagery, we strive to develop techniques that can translate these patterns into visual representations. Such technology holds the potential to revolutionize various fields, from aiding individuals with communication disabilities to enhancing virtual reality experiences.

However, despite the promise of this technology, several challenges lie ahead. One of the primary obstacles is the inherent complexity of the brain's neural activity. Although electroencephalogram (EEG) signals are informative and cheap to record, they are often noisy and difficult to interpret with precision [1]. Additionally, the subjective nature of mental imagery poses a significant challenge, as individual experiences and interpretations vary widely. Moreover, ethical considerations surrounding privacy and consent must be carefully addressed, particularly when dealing with sensitive mental processes.

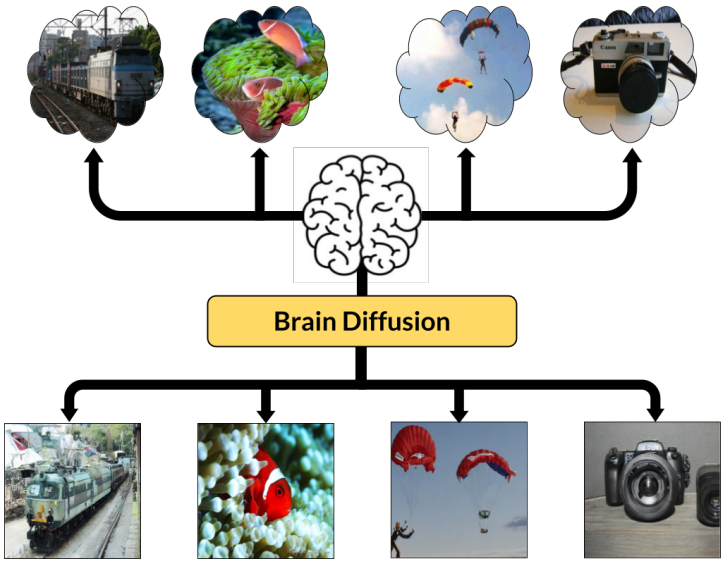

Figure 1: Brain Diffusion is capable of generating photo-realistic images from EEG waves based on perceived visual stimuli from a brain.

2 Background

2.1 Measuring Brain Activity

EEG signals are discrete measurements of superficial neural activities with high temporal resolution (> 30 Hz) but weak spatial resolution [2]. As previously mentioned, EEG signals tend to be very noisy and variant across many samples and datasets alike. This is partly influenced by the lack of ubiquity in measurement devices used in different research studies.

2.2 Previous Works

The work in this report is heavily inspired by the novel breakthroughs made in DreamDiffusion [1]. The authors of [1] deal with the noisy property of EEG waves by leveraging the deep learning paradigms of autoencoders with masked modeling. Rich, contextual representations of EEG waves can be used to also learn transferable modalities through text/images [3].

Recently, researchers at Meta AI [4] present a framework using contrastive learning to decode magnetoencephalogram (MEG) waves. MEG is a non-invasive medical imaging technique which offers much higher spatial and temporal resolution ($\approx 5,000$ Hz) [4]. However, collecting MEGs incur significant costs as it requires skilled operators within medical facilities. Their lack of portability further compounds these challenges. Rather, the incentive to generate images from EEG waves presents a much more appealing and potentially revolutionary alternative.

2.3 Masked Signal Modeling (MSM)

With the advent of transformers [5, 6] in this era of deep learning, the task of learning contextual representations of input modalities has become quintessential. Similar to how autoencoder transformers, like BERT [7], are pre-trained on the task of masked language modeling (MLM) in natural language processing (NLP), previous works have shown that masked autoencoders also have the capability to reconstruct visual semantics [8, 9], known as masked image modeling (MIM) [10, 11]. Transitively, the same concept can be applied to EEG signals, namely masked signal modeling (MSM).

2.4 Latent Diffusion Models

Diffusion models, popular for their efficacy in generating high-quality content [12], utilize bi-directional Markov Chains to align with image-like data biases, resulting in strong generative capabilities [12, 13]. Optimal synthesis quality often involves reweighted objectives during training to balance image fidelity and compression efficiency. However, evaluating and refining these models in pixel space proves resource-intensive and time-consuming [14, 15]. Latent diffusion models (LDMs) [16] alleviate these challenges by operating within a compressed

latent space, enhancing efficiency and synthesis quality through the integration of UNet-based denoising models with attention mechanisms. This approach facilitates enhanced conditioning of image generation [16].

3 Data

As this framework requires a two-stage training process of pre-training and fine-tuning, there are two primary dataset categories for this task.

3.1 Eclectic EEG Signals

The authors of [1] amass approximately 120,000 EEG data samples ranging from a variety of cognitive tasks pertaining to motor imagery, positive deflection, visual evoked potentials, and event-related potentials [17]. Eclectic EEG signals are chosen from over 400 subjects so that robust representations of most EEG waves can be learned accordingly. This also potentially enables an EEG encoder to generalize across multiple subjects, indicating its proficiency in both *interspecific* and *intraspecific* learning.

3.2 EEG-Image Pairs

The dataset in [18, 19] comprises of EEG signals from six subjects, following a recording protocol that involved presenting 40 object classes, each containing 50 images sourced from ImageNet [20]. Thus, a total of 2,000 images were presented making a grand-total of 12,000 EEG-Image pairs. Each image was presented for 0.5 seconds, with a 10-second interval of a black screen maintained between class blocks while EEG data continued to be recorded.

Each EEG segment is characterized by 128 channels, recorded over a duration of 0.5 seconds at a sampling rate of 1 kHz, and is represented as a matrix with dimensions of $128 \times L$, where $L = 512$ denotes the interpolated number of samples in each segment on each channel [18, 19].

4 Method

The methodology proposed in this paper follow three primary steps (two of which happen concurrently):

1. Pre-Training EEG Encoder (Section 4.1)

2. Fine-Tuning Phase
   (a) EEG-Image Embedding Alignment (Section 4.2)
   (b) Fine-Tuning Stable Diffusion (Section 4.3)

Initially, it is imperative to devise a methodology for extracting semantic representations from EEG waves. This preparatory step becomes essential as it lays the foundation for subsequent procedures necessitating EEG embeddings for alignment with CLIP image embeddings. To facilitate this process, random portions of the raw EEG wave are masked (like masked tokens). An autoencoder [7, 8] is pre-trained for the task of reconstructing masked EEG signals, thereby ensuring the fidelity and integrity of the derived semantic representations [1].

4.1 Pre-Training EEG Encoder

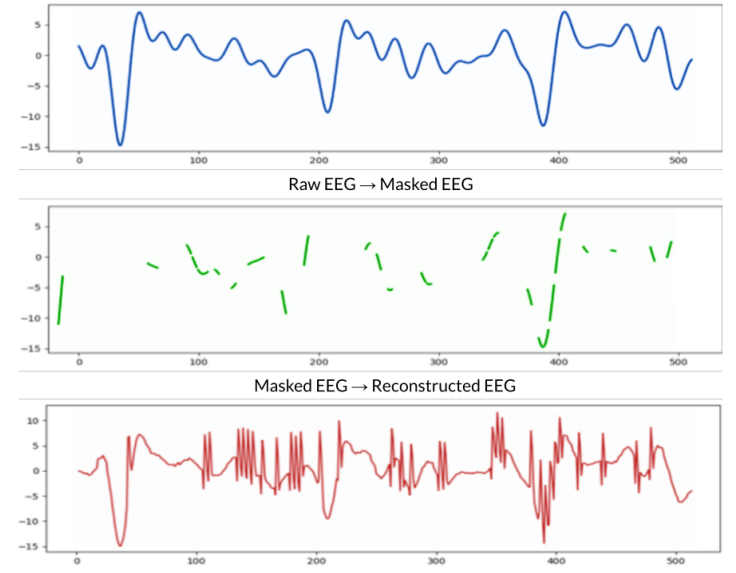

Figure 2: The noisy, eclectic EEG signals are randomly masked (75%) and reconstructed during the pre-training of EEG Encoder.

A Masked Auto-Encoder (MAE) [8], using a vision transformer (ViT) [6] as its backbone, is pre-trained to reconstruct eclectic EEG signals with high accuracy. During learning, the attention mechanism layers of the transformer learn deep, semantic representations of EEG signals in a latent space which serve an embedding for any given EEG signal. Referring to Fig. 2, a masking ratio of 75% has been shown to yield the highest overall accuracy for this task [1] and for reconstructing visual semantics in general [8].

4.2 EEG-Image Embedding Alignment

During fine-tuning, we experiment with different loss criteria to better align embeddings outputted from the EEG Encoder with embeddings outputted from

the pre-trained CLIP Encoder. We also study the effects of using a contrastive learning objective for the task of embedding-space alignment.

The authors of [1] define and apply only one loss criteria—cosine similarity.

4.2.1 Cosine Similarity

(***CosSim***)

$$\mathcal{L}^{sim} = \sum_{i \in I} \mathcal{L}_i^{sim}$$

$$\mathcal{L}_i^{sim} = 1 - \text{sim}(z_i, z_{j(i)}) \tag{1}$$

In (1), let $i \in I \equiv \{1...N\}$ be the index of all EEG embeddings and their corresponding image embeddings. z_i refers to the source EEG embedding, $z_{j(i)}$ refers to its associated CLIP image embedding, and sim() essentially denotes the cosine similarity. To align the embedding spaces, we aim to maximize the cosine similarity between corresponding embedding pairs [1]. To minimize loss function during optimization, the loss criteria for cosine similarity is defined as 1 minus the cosine similarity.

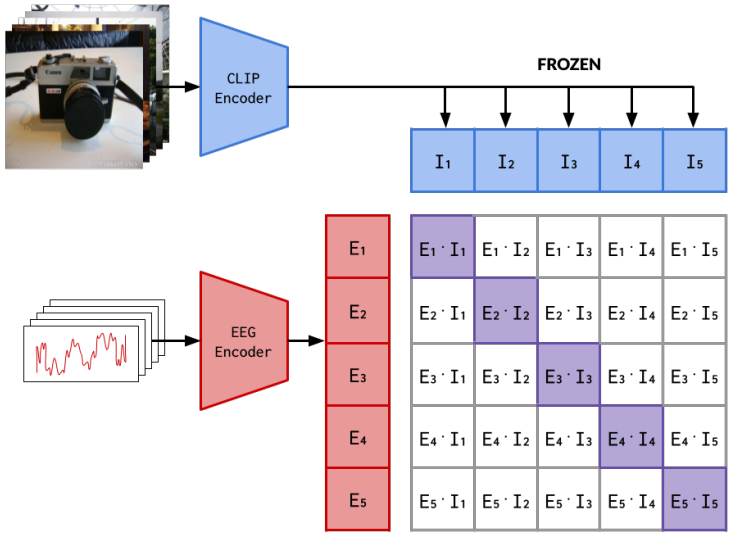

Figure 3: During fine-tuning, the parameters of the EEG Encoder are learned through a contrastive objective to reduce the cosine similarity between pairwise embeddings. The diagonal of the similarity matrix (highlighted in purple) distinguishes the pairwise similarities. More detail is covered in Section 4.2.2.

4.2.2 Self-Supervised Contrastive Learning

(***SimCLR***)

$$\mathcal{L}^{self} = \sum_{i \in I} \mathcal{L}_i^{self}$$

$$\mathcal{L}_i^{self} = -\log \frac{\exp(\text{sim}(z_i, z_{j(i)})/\tau)}{\sum_{b \in B(i)} \exp(\text{sim}(z_i, z_b)/\tau)} \tag{2}$$

In (2), $\mathcal{L}_i^{self}$ refers to a self-supervised contrastive loss criteria in which pairwise cosine similarities are calculated. The similarity of two embedding pairs can be represented as the diagonal of the similarity matrix, as shown in Fig. 3. Let $B(i) \in I$ where $B(i)$ represents all other CLIP image embeddings in the multiviewed batch such that $z_b \neq z_{j(i)}$ [21, 22]. The variable z_b denotes the index of an arbitrary, negative CLIP image embedding sample and τ represents a temperature, scalar parameter to be tuned during training.

The self-supervised contrastive loss criteria (2) serves to increase the similarity between a positive pairwise example while simultaneously penalizing the negative combinations of pairs [21]. However for this specific task, a major drawback of (2) is the assumption that there cannot be more than 1 positive pair in the batch [22]. The same EEG embedding can represent similar visual semantics indicating that there can be multiple positive pairs in a batch during fine-tuning — which is where a supervised loss criteria can help.

4.2.3 Supervised Contrastive Learning

(***SupCLR***)

$$\mathcal{L}^{sup} = \sum_{i \in I} \mathcal{L}_i^{sup}$$

$$\mathcal{L}_i^{sup} = \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(z_i, z_{j(i)})/\tau)}{\sum_{b \in B(i)} \exp(\text{sim}(z_i, z_b)/\tau)} \tag{3}$$

In (3), $P(i) \equiv \{p \in A(i) : \widetilde{y_p} = \widetilde{y_i}\}$ defines the indeces of all positive samples within the multiviewed batch, where $|P(i)|$ is its cardinality—the total number of positive samples [22]. The outer summation allows us to aggregate losses for multiple positives in the same batch instead of calculating a singular pairwise losses, as in (2).

Note that if $|P(i)| = 0$, or equivalently, if there are no other positive samples in the batch aside from the associated CLIP image embedding, (3) is no different than (2).

The loss objective of (3) enables the model to learn patterns between similar samples pertaining to the same ground truth class. Say, we have two independent EEG embeddings for a subject envisioning a "dog" in their mind. During fine-tuning, the model should learn to increase pairwise similarities between two

EEG embeddings of a "dog" since they both are semantically the same. By aggregating losses from the same positive samples, the model can learn similarities (for positives) and contrasts (for negatives) without penalizing EEG-Image pairs of the same class [22].

4.3 Fine-Tuning Stable Diffusion

After pre-training the EEG Encoder to extract semantic representations of EEG waves described in 4.1, we fine-tune a pre-trained checkpoint of Stable Diffusion while simultaneously updating the parameters of the EEG Encoder to align the embedding-space of EEG waves with that of images, as mentioned in 4.2.

Stable Diffusion (DF) operates within the latent space, where it encodes an image x using a Vector Quantization (VQ) encoder [1, 16], $\mathcal{E}(\cdot)$, to yield its latent representation $z = \mathcal{E}(x)$. The UNet architecture employs a cross-attention mechanism to introduce conditional signals, allowing integration of additional information, such as EEG data. Specifically, the output of the EEG encoder y is projected using a projector τ_θ to yield an embedding $\tau_\theta(y)$ in $M \times d_\tau$ dimensions [1], which is then incorporated into the UNet via a cross-attention layer implementing $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V$. While fine-tuning, we simultaneously optimize the EEG Encoder and cross-attention heads of the de-noising UNet while freezing the remaining parameters of the Stable Diffusion framework. The following loss function is employed during the fine-tuning process for the pre-trained SD checkpoint.

$$\mathcal{L}_{\mathcal{SD}} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(x_t, t, \tau_\theta(y))||_2^2\right] \tag{4}$$

where ϵ_θ represents the de-noising UNet [1].

5 Results

There is not much evaluation to be done on the pre-trained EEG Encoder due to insufficient time. Instead, the results of this paper will focus on presenting findings and analysis on the experiments conducted during the fine-tuning phase.

5.1 Embedding Alignment Evaluation

t-SNE (t-distributed Stochastic Neighbor Embedding) is a widely used technique for nonlinear dimensionality reduction that preserves local and global structure in high-dimensional data. By transforming EEG embeddings into a lower-dimensional space using t-SNE, we aim to uncover meaningful semantic representations and cluster patterns for analysis.

Figure 4: Scatterplot of dimensionally reduced EEG embeddings plotted on 2D t-SNE feature space. The points represent EEG embeddings reduced into 2D space and are color-coded by their class label. *(Top)* EEG embedding-space **before** fine-tuning. *(Bottom)* EEG embedding-space **after** fine-tuning.

Since the EEG Encoder was pre-trained for the task of reconstructing the original EEG wave, it does not pick up on cortical semantics or latent representations of one's thoughts. Rather, this happens during fine-tuning when the EEG Encoder recognizes semantic similarities between the two embedding-

spaces which resemble cortical semantics. As seen in Fig. 4, positive EEG embeddings cluster together and separate from their respective negatives.

In order to quantitatively evaluate which loss criteria, ie. (1), (2), (3), best aligns EEG embeddings with CLIP image embeddings, the Dunn Index (DI) [23] can serve as a metric of evaluation for clustering algorithms and dimensionally reduced data. A higher DI indicates superior clustering performance. It operates under the assumption that effective clustering entails compact clusters that are well-distinguished from one another.

$$DI = \frac{\min_{1 \leq i \leq j \leq m} \delta(C_i, C_j)}{\max_{1 \leq k \leq m} \Delta_k} \tag{5}$$

Here, $\min_{1 \leq i \leq j \leq m} \delta(C_i, C_j)$ entails determining the minimum inter-cluster distance amongst all possible combination of cluster pairs (C_i and C_j) where $1 \leq i \leq j \leq m$. The denominator of (5), $\max_{1 \leq k \leq m} \Delta_k$, defines the maximum intra-cluster distance. In simple terms, this is referring to the spread of the largest cluster C_k where $1 \leq k \leq m$.

The ratio of the smallest inter-cluster distance (separation) and the largest intra-cluster distance (spread) yields higher values when clusters are separated with larger gaps and compact with smaller cluster sizes. Using this metric, we analyze the EEG Encoder's outputted embeddings before and after fine-tuning with different training objectives, ie. ***CosSim***(4.2.1), ***SimCLR***(4.2.2), and ***SupCLR***(4.2.3).

Table 1: Cluster Analysis - EEG Embedding Space

	Training Loss Criteria			
	Pretrained	***CosSim***	***SimCLR***	***SupCLR***
Dunn Index	0.04018	0.07640	0.06528	0.06037

Looking at Table 1, it is noticable that ***SupCLR*** did not have superior performance in EEG-image embedding alignment. In fact, the baseline (***CosSim***) outperformed all other potentially improved training objectives. We believe this likely due to the fact that both ***SimCLR*** and ***SupCLR*** are largely reliant on large batch sizes for variant datasets. Due to a lack of computation power, we trained with a maximum, affordable batch size of 8 when there are a total of 40 possible labels when aligning EEG-image pairs.

As mentioned in 4.2.3, (3) is no different than (2) if there are no other positive samples present in the multiview batch. In this case, since the batch size is so small in comparison to the number of classes we have, the probability of having atleast one positive sample is $1 - \frac{39}{40}^7$ which is $\approx 16\%$. Eq. (3) suffers a lack of positive samples to apply a supervised training objective due to a small batch size. To confirm this hypothesis, we subset the entire dataset to only 12 unique classes and keep the batch size as 8 which yields higher probabilities of

positive samples present in a multiview batch. Specifically the probability of having atleast one positive sample is $1 - \frac{11}{12}^7$ which is $\approx 46\%$.

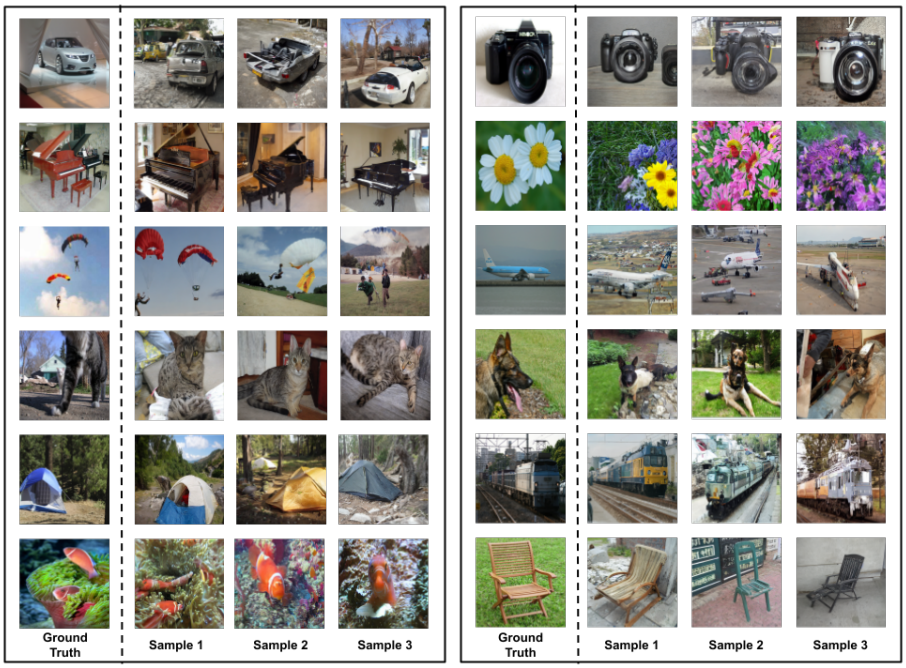

Figure 5: Sampled results for BrainDiffusion using self-supervised contrastive learning. The columns labeled "Ground Truth" portray the original image that a subject visually perceived. The columns labeled "Sample [1,2,3]" depict selected samples from the generated images.

Table 2: Cluster Analysis - EEG Embedding Space

	Training Loss Criteria			
	Pretrained	***CosSim***	***SimCLR***	***SupCLR***
Dunn Index	0.04018	0.07527	0.07905	0.08089

The above table, (Table 2) confirms our hypothesis that reducing the number of unique classes enables ***SimCLR*** and ***SupCLR*** to improve the alignment of EEG embeddings with CLIP image embeddings, provided a small batch size.

5.2 Image Generation Evaluation

We evaluate the fidelity and accuracy of the generated images by using a pre-trained image classification model. Specifically, we use "*google/vit-base-patch16-224*" [6] which has a very high accuracy on the ImageNet-1K classification task [20].

Table 3: Generated Image - Accuracy Table

	Accuracy Metric	
	Top-1 Acc	***Top-3 Acc***
CosSim	0.0500	0.1806
SimCLR	0.0863	0.1417
SupCLR	0.0627	0.2214

Table 3 validates this because the highest accuracy obtainable by Dream Diffusion was 22.14% using ***SupCLR***. About 1 in 5 generated images are visually the same as the ground truth when comparing, although this is highly dependent on the class of the ground truth.

As expected, the *Top-1 Acc* scores for all three training objectives are much lower than that of the *Top-3 Acc*. This likely due to the fact that EEG waves are very noisy and it is difficult to obtain discrete, semantic representations of different visual image classes.

CosSim consistently performs worse than ***SimCLR*** and ***SupCLR*** on the subset of data reduced to 12 unique classes.

Table 3 does indeed corroborate better accuracy results for ***SimCLR*** and ***SupCLR*** in *Top-3 Acc* and *Top-1 Acc*, respectively.

5.3 Experiments on Measures of Similarity

The authors of [1] solely use cosine similarity to align the embeddings of EEG-Image pairs. We explore the possibility of using a normalized, scalar dot product defined as follows.

$$\mathcal{L}^{sim} = \sum_{i \in I} \mathcal{L}_i^{sim}$$

$$\mathcal{L}_i^{sim} = \frac{z_i \cdot z_{j(i)}}{||z_i|| \cdot ||z_{j(i)}||} \tag{6}$$

where $||z_i||$ and $||z_{j(i)}||$ are the norms of the two embedding pairs.

Figure 6: Scatterplot of dimensionally reduced EEG embeddings plotted on 2D t-SNE feature space. The points represent EEG embeddings reduced into 2D space and are color-coded by their class label. *(Top)* EEG embedding-space with **cosine similarity**. *(Bottom)* EEG embedding-space with **normalized dot product similarity**.

Qualitatively, one can notice similar cluster patterns recurrent in both EEG embedding spaces using either cosine similarity or a normalized dot-product similarity. As shown in Fig. 6, there is a set of points alone at the top of the

graph near t-SNE Feature 1 ≈ 0 and t-SNE Feature 2 ≈ 35 that are resembled in both embedding spaces. This leads us to infer that they are both valid similarity metrics and can be used for this task.

Table 4: Cluster Analysis - Embedding Alignment

	Training Loss Criteria	
	CosSim	***DotSim***
Dunn Index	0.06671	0.06528

Quantitatively, Table 4 confirms that they both are equivalent for aligning EEG embeddings to CLIP image embeddings due to approximately the same Dunn Index values.

5.4 Limitations

There are many failure cases and limitations with the generated images of Brain Diffusion due to the simple fact that EEG waves are extremely noisy.

Figure 7: **Failure Results** for BrainDiffusion using contrastive learning. The columns labeled "Ground Truth" portray the original image that a subject visually perceived. The columns labeled "Sample [1,2]" depict selected samples from the generated images.

As seen in Fig. 7, there are many such examples of failure during generation of images using EEG embeddings. The belief is that core, low-level features such as color and shape are almost instantaneously noticed by the brain which may be encoded in the EEG signal. This can result in similar shapes or colors in the generated image, but not necessarily the same visual semantics. Additionally, the latent representations of EEG embeddings may not entirely be aligned accurately with that of CLIP image embeddings. As seen in Fig. 6, there can be clusters of positive samples that are not separated and tightly clustered together even after reaching a local optima after fine-tuning.

6 Conclusion

In conclusion, BrainDiffusion reconstructs visual semantics from non-invasive EEG waves using a supervised contrastive loss criteria during the alignment of EEG embeddings with CLIP image embeddings. Our findings show that with an ample batch size, ***SupCLR*** does indeed outperform using solely ***CosSim*** for similarity measures. This not only leads to better embedding alignment but also higher accuracy and fidelity in image generation itself with our fine-tuned Stable Diffusion.

Implications This project aims to bridge the gap between abstract thoughts and tangible images by leveraging deep learning techniques, particularly contrastive learning, to decode neural patterns underlying mental imagery. While promising, challenges such as the complexity of EEG signals, subjective nature of mental imagery, and ethical considerations surrounding privacy and consent remain significant hurdles. The appeal of generating images directly from cheap and easily accessible EEG waves potential revolutionary implications for various applications, from aiding communication disabilities to enhancing virtual reality experiences.

References

[1] Y. Bai, X. Wang, Y. Cao, Y. Ge, C. Yuan, and Y. Shan, "Dreamdiffusion: Generating high-quality images from brain eeg signals," *arXiv preprint arXiv:2306.16934*, 2023.

[2] C. Amo, L. De Santiago, R. Barea, A. López-Dorado, and L. Boquete, "Analysis of gamma-band activity from human EEG using empirical mode decomposition," *Sensors (Basel)*, vol. 17, no. 5, p. 989, Apr. 2017.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[4] Y. Benchetrit, H. Banville, and J.-R. King, "Brain decoding: toward real-time reconstruction of visual perception," 2024.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021.

[9] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *ArXiv*, vol. abs/2106.08254, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235436185

[10] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "A unified view of masked image modeling," 2022.

[11] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," 2022.

[12] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," 2015.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[14] Z. Kong and W. Ping, "On fast sampling of diffusion probabilistic models," 2021.

[15] R. San-Roman, E. Nachmani, and L. Wolf, "Noise estimation for generative diffusion models," 2021.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022.

[17] B. Aristimunha, I. Carrara, P. Guetschel, S. Sedlar, P. Rodrigues, J. Sosulski, D. Narayanan, E. Bjareholt, B. Quentin, R. T. Schirrmeister, E. Kalunga, L. Darmet, C. Gregoire, A. Abdul Hussain, R. Gatti, V. Goncharenko, J. Thielen, T. Moreau, Y. Roy, V. Jayaram, A. Barachant, and S. Chevallier, "Mother of all BCI Benchmarks," 2023. [Online]. Available: https://github.com/NeuroTechX/moabb

[18] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, J. Schmidt, and M. Shah, "Decoding brain representations by multimodal learning of neural activity and visual features," 2020.

[19] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, M. Shah, and N. Souly, "Deep learning human mind for automated visual classification," 2019.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[21] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020.

[22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," 2021.

[23] M. Gagolewski, M. Bartoszuk, and A. Cena, "Are cluster validity measures (in) valid?" *Information Sciences*, vol. 581, p. 620–636, Dec. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2021.10.004