BrainVis: Exploring the Bridge between Brain and Visual Signals via Image Reconstruction

Honghao Fu[1,2] Zhiqi Shen[2] Jing Jih Chin[2] Hao Wang[1†]

[1]The Hong Kong University of Science and Technology (Guangzhou)

[2]Nanyang Technological University

https://brainvis-projectpage.github.io

Abstract

Analyzing and reconstructing visual stimuli from brain signals effectively advances understanding of the human visual system. However, the EEG signals are complex and contain a amount of noise. This leads to substantial limitations in existing works of visual stimuli reconstruction from EEG, such as difficulties in aligning EEG embeddings with the fine-grained semantic information and a heavy reliance on additional large self-collected dataset for training. To address these challenges, we propose a novel approach called BrainVis. Firstly, we divide the EEG signals into various units and apply a self-supervised approach on them to obtain EEG time-domain features, in an attempt to ease the training difficulty. Additionally, we also propose to utilize the frequency-domain features to enhance the EEG representations. Then, we simultaneously align EEG time-frequency embeddings with the interpolation of the coarse and fine-grained semantics in the CLIP space, to highlight the primary visual components and reduce the cross-modal alignment difficulty. Finally, we adopt the cascaded diffusion models to reconstruct images. Our proposed BrainVis outperforms state of the arts in both semantic fidelity reconstruction and generation quality. Notably, we reduce the training data scale to 10% of the previous work.

1. Introduction

The brain-computer interface (BCI) establishes a connection between the brain and external devices [13], enabling people to obtain signals of brain activity. Meanwhile, the advancement in deep learning (DL) has significantly improved the feasibility of complex brain signal decoding [1]. These build foundations for research works [3, 21, 23] that study brain activity in response to external stimuli, including the correlation between brain signals and sensory perception, such as hearing [10], smells [38] and vision [32]. Particularly, exploring the brain visual processing is crucial for enhancing our understanding of human brain's visual and related functions. However, the visual processing has extensive involvement of neurons, making the explaining of brain's visual process challenging [15].

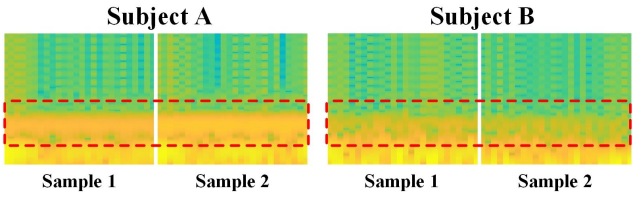

Figure 1. Spectrograms of EEG samples from two different subjects, showing the usefulness of frequency domain information. Samples from one subject, e.g. *Samples 1 and 2 of Subject A*, exhibit distinct local features and similar global features. While different subject samples, e.g. *Samples of Subjects A and B*, present various global features.

To promote the understanding of human brain's visual function, in recent years, some studies have utilized deep neural networks to extract visual information from brain signals and reconstruct images [9, 16, 23, 27, 31, 33, 35]. Electroencephalography (EEG) is one of the brain signals which captures overall brain activity using scalp electrodes, yielding multi-channel 2D temporal sequences [32]. Despite its advantages of low cost and ease of processing [18], EEG faces several challenges, including limited spatial resolution, imprecise brain region localization abd noise from non-brain factors [20]. These issues make image reconstruction from EEG signals highly challenging.

Previous attempts [16, 31, 35] proposed to integrate long short-term memory (LSTM) model and generative adversarial networks (GAN). Because of the scarcity of training samples, they only presented poor semantic accuracy and reconstruction quality. A more recent study [2] pro-

†The corresponding author.

posed an alternative method adopting masked autoencoder (MAE) [14] and latent diffusion model (LDM) [27], demonstrating improved reconstruction performance. However, we observe this method still has the following limitations: (1) Learned EEG feature embedding heavily relying on the additional collected training dataset; (2) Only extracting time-domain features, without considering the significant inter-subject heterogeneity in the frequency domain, as shown in Figure 1; (3) Only conducting coarse-grained category reconstruction, struggling to capture fine-grained semantic details in the reconstructed images.

To overcome these limitations, we propose **BrainVis**, a novel pipeline for image reconstruction from EEG. It comprises four modules: (1) Time encoder, which divide the EEG into various units and apply a self-supervised approach on them to obtain EEG time-domain features, in an attempt to ease the training difficulty; (2) Frequency encoder, responsible for converting EEG signals from time domain to frequency domain and extracting frequency-domain feature, enhancing the EEG representations; (3) Cross-modal EEG alignment, it aligns EEG time-frequency embeddings with the interpolation of the coarse and fine-grained semantics in the CLIP [26] space, to highlight the primary visual components and reduce the cross-modal alignment difficulty; and (4) Cascaded diffusion models, employed for reconstruction. The aligned EEG fine-grained embeddings, along with the coarse-grained embeddings of EEG classification results, serve as conditions of cascaded diffusion models for multi-level semantic visual reconstruction.

Our contributions can be summarized as:

- We improve the self-supervised embedding method for EEG time-domain features, eliminating the reliance on additional self-collected large-scale dataset.
- We propose the first EEG visual stimuli reconstruction method that integrates time and frequency-domain features, enhancing the EEG representations.
- BrainVis generates images with higher accuracy and better quality, and also overcomes previous limitation that was limited to only coarse-grained reconstruction.

The experimental results demonstrate that our proposed BrainVis outperformed the state-of-the-art in both semantic reconstruction and generation quality. Additionally, to address the lack of quantitative baseline evaluation in this field, we introduced Structural Similarity Index Measure (SSIM) and Cosine Similarity (CS) as evaluation metrics for semantic and generation quality.

2. Related Works

2.1. Reconstructing Visual Stimuli from fMRI

Compared to EEG, functional magnetic resonance imaging (fMRI) has higher spatial resolution, lower noise and greater information capacity. As a result, fMRI-based reconstruction is more popular [8]. Early studies [12, 17, 37] utilized linear regression with traditional or DL features for visual classification or visual reconstruction using deep decoding networks [4]. With the development of GAN [11], there has been a growing trend in using GAN as a core framework for visual reconstruction [9, 30].

Recently, a study [23] used pre-trained CLIP model [26] to obtain embeddings of images and their artificial captions, aligning fMRI with these CLIP embeddings, then used it as the conditions for image generation by StyleGAN2. A study [33] transferred this latent embedding alignment to LDM [27], achieving higher-quality reconstruction. Subsequently, works that centered around CLIP and LDM [5, 24, 25, 34] continued to improve the accuracy and image quality of visual reconstructions, including introducing vector quantized-variational autoencoder (VQ-VAE) [36] and masked brain modeling (MBM) based on MAE [14] for better fMRI feature embedding. However, applying fMRI in practical settings is limited due to high cost, large data scale, and high processing complexity.

2.2. Reconstructing Visual Stimuli from EEG

Compared to fMRI, EEG has the advantages of being easy to obtain and cost-effective. However, its lower spatial resolution and information capacity, along with significant noise interference, make reconstructing visual stimuli from EEG more challenging. In early work [16], LSTM was used for EEG classification and feature embedding, then variational autoencoder [19] and GAN were used to generate images from EEG feature embeddings. Their results showed that EEG contained visual semantic information, and that GAN generated higher-quality images compared to VAE. However, due to the inherent limitations of EEG and the small size of the EEG-Image pairs datasets, the results of studies using LSTM and GAN frameworks [31, 35] were not ideal.

Inspired by MBM [5], the recently proposed DreamDiffusion [2] embeds time-domain features of EEG using MAE and aligns them with CLIP-encoded images, as conditions of pre-trained LDM to achieve coarse-grained category reconstruction. However, it only captures the time-domain features of EEG signals, and the ability of its models' unsupervised feature embedding relies on pre-training on an additional large-scale self-collected EEG dataset (approximately 120,000 samples), indicating that it requires further enhancement in the representation capability of EEG features. Moreover, the condition of well pre-trained stable diffusion is text embedding, but DreamDiffusion directly aligns EEG features with image embeddings in CLIP space, which may introduce additional semantic noise, since text and image embeddings are not entirely equivalent.

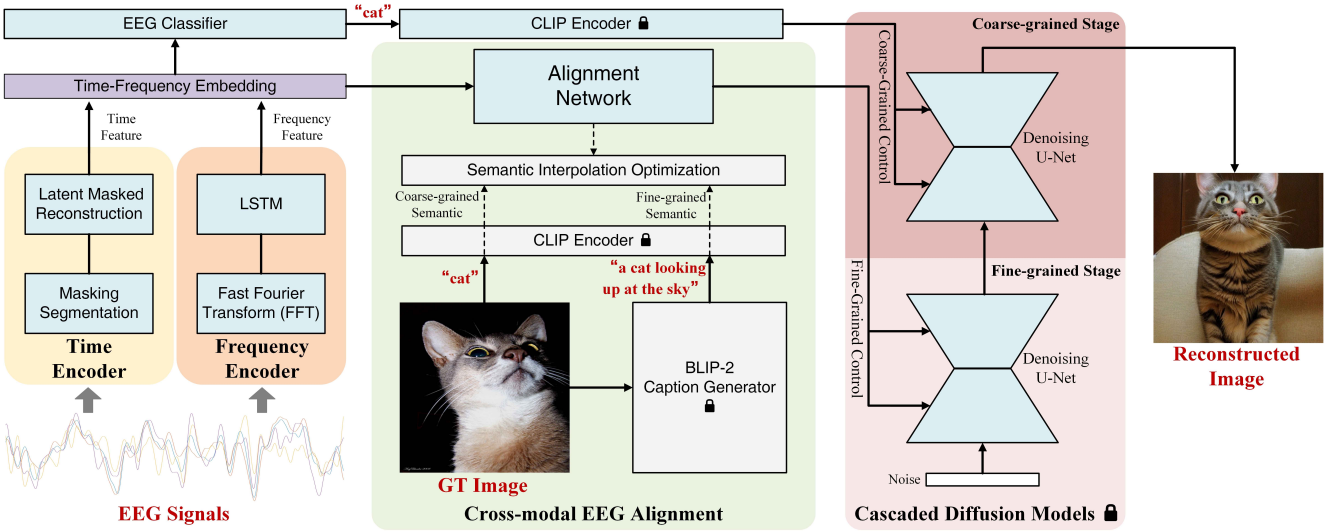

Figure 2. Overview of our proposed **BrainVis**. We first aim to obtain the time and frequency features for the given EEG signals, in which the time encoder leverages self-supervised pre-training approach with masking segmentation and latent reconstruction, while the frequency encoder employs LSTM to extract features with Fast Fourier Transform (FFT). Then the time and frequency encoders are fine-tuned simultaneously with EEG classifier to obtain time-frequency dual embedding. Following this, an alignment network aligns the EEG time-frequency dual embedding with the fine-grained semantic of visual stimuli image in CLIP space, which is semantic interpolation of the label of visual stimuli image and its caption. Finally, the aligned EEG fine-grained embedding, along with the coarse-grained embedding of classification result, serve as conditions of cascaded diffusion models for multi-level semantic visual reconstruction.

3. Method

3.1. Preliminary

CLIP [26] is a pre-trained model for image-text similarity measurement, which embeds them in the same latent space with the contrastive learning. It mainly consists of an image encoder and a text encoder.

Stable diffusion [27] is a text-to-image generation model, which uses CLIP embedding of text as its conditional input. It guides the denoising process from random noise by controlling the conditional input of the U-Net, realizing the generation of images with specified semantics.

BLIP-2 [22] is a visual-language pre-trained model that can generate captions from images. It utilizes pre-trained image and language models with fixed parameters, and uses a query transformer to fill the gap between their modalities.

3.2. Overview

In Figure 2, we show our proposed BrainVis, which comprises four components: time encoder, frequency encoder, cross-modal EEG alignment, and cascaded diffusion models. Time encoder leverages self-supervised pre-training approach with masking segmentation and latent reconstruction, embedding fine-grained time-domain features extracted from EEG. Then frequency encoder uses LSTM to obtain frequency-domain features with Fast Fourier Transform (FFT). Before EEG alignment, the time encoder and frequency encoder are fine-tuned together with EEG classifier under visual semantic supervision to obtain time-frequency dual embedding, which is then fed into an alignment network. The alignment network undergoes overall fine-tuning supervised by the CLIP embeddings, which represent the semantic interpolation of the labels of corresponding visual stimuli and their captions generated by BLIP-2 [22]. This enables the fine-grained semantics alignment of the EEG embeddings with the CLIP space. Finally, the aligned EEG fine-grained embeddings, along with the coarse-grained embeddings of classification results, serve as conditions of cascaded diffusion models for multi-level semantic visual reconstruction.

3.3. Time-Frequency Dual Embedding

To learn EEG time-domain features, prior research [2] utilized a masked autoencoder [14]. However, it demands a substantial volume of training data and struggles to represent complex EEG signals accurately. Drawing inspiration from the prior art [6], our approach leverages window slicing and a random masking strategy to partition the signal into various semantic units. Instead of directly reconstructing the entire signals in the time domain, our approach focuses on reconstructing the embeddings of single semantic units, which eases the reconstruction difficulty. Meanwhile, each semantic unit of the signal has a high-dimensional embedding, increasing the information density.

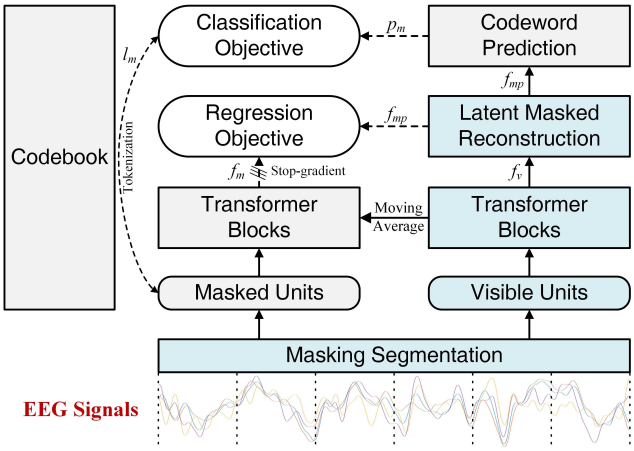

Figure 3. Self-supervised pre-training of time encoder. First, the EEG is segmented into visible and masked slices, then the masked slices are tokenized as l_m. Next, the visible features f_v are extracted by transformer blocks, which are used for predicting the masked features f_{mp} and their codewords. The probability representation of predicted codewords is p_m. Meanwhile, a non-trainable model obtained by moving average from the transformer blocks is used to extract real features of masked slices f_m. Finally, regression and classification objectives are performed.

Technically, as shown in Figure 3, the EEG signal $x \in \mathbb{R}^{c \times l}$, with c channels and l time series length, is uniformly divided into n time slices as semantic units. Each slice is projected into a d-dimension space to extract local features across channels. This results in a representation of $z \in \mathbb{R}^{n \times d}$. A random masking strategy is further applied with a mask radio r_m, partitioning z into the visible and the masked part. It is worth noting that our self-supervised learning objective is two-pronged. One is the regression objective employed for reconstructing signal feature embeddings. Another is the codebook-based semantic unit classification objective, which facilitates learning the semantic attributes and enhances the EEG feature representations.

To achieve the regression objective, we first use transformer blocks to learn the visible features f_v. Then the masked features f_m are acquired by another non-trainable model obtained by moving average from the f_v transformer blocks. Further, new transformer blocks are introduced for latent masked reconstruction, which take f_v as input and generate the predicted value f_{mp} for f_m. The mean squared error (MSE) between f_m and f_{mp} is computed to derive regression objective, and the loss L_{reg} is define as:

$$L_{reg} = \frac{1}{d}||f_m - f_{mp}||_2^2. \tag{1}$$

In term of the classification objective, a tokenizer assigns the one-hot code l_m to each masked slice as self-supervised labels through the codebook, where l_m is also defined as the codeword. The total number of codewords is denoted as n_t. We employ a linear layer to convert each masked slice into a probability representation consisting of n_t codewords. By classifying f_{mp}, we obtain the probability representation p_m, and the cross-entropy error between p_m and l_m is computed to derive classification objective. The classification loss L_{cls} is define as:

$$L_{cls} = -l_m \cdot \log(p_m) \tag{2}$$

The overall learning objectives of the masked autoencoder can be defined as:

$$L_{te} = L_{reg} + L_{cls}. \tag{3}$$

The learned transformer blocks are the EEG time encoder.

After the self-supervised learning process, we also incorporate the EEG paired image labels into the time encoder training, in which we adopt the cross-entropy loss.

Frequency embedding learning. In terms of the frequency-domain features, we first convert EEG from the time to the frequency domain using Fast Fourier Transform (FFT). However, due to the limited training data scale, learning EEG frequency-domain features through complex networks may lead to challenging overfitting issues. Therefore, we use a LSTM model to extract the visual semantic frequency-domain feature embedding. To this end, we also use EEG paired image labels as the supervision to train the LSTM model, which is the frequency encoder. We use the cross-entropy loss for training.

Time-frequency embedding fusion. To build the unified time-frequency features, we further fine-tune the time and frequency encoders jointly. Specifically, we concatenate the output of time and frequency encoders, and adopt the cross-entropy loss, with the image labels as supervision. The predicted label of the EEG signals will be further used in the image reconstruction phase.

3.4. Cross-modal EEG Alignment

To achieve high-quality image reconstruction, we use a pre-trained stable diffusion [27] model as the backbone generator. Previous work [2] proposed to align the EEG embeddings with the image CLIP embeddings, and then input the EEG embeddings to the stable diffusion for image generation. It is notable that the original stable diffusion model requires textual descriptions as the conditional input. Due to the gap between the image and text CLIP embeddings [39], they [2] failed to reconstruct images containing fine-grained information. To allow the stable diffusion model adaptable in our task, we propose to align our EEG signal embeddings with the corresponding text semantic features.

To be specific, for each instance, we only have coarse category annotation, e.g. "*cat*", which is not enough to represent the complex semantics in the given image. Therefore, we propose to generate pseudo captions to incorporate

more fine-grained semantics. We generate captions through BLIP-2 [22]. Then, we obtain the CLIP encoding of the captions as c_{cap} through CLIP text encoder.

However, we empirically observe the accurate alignment between the EEG features and c_{cap} is hard to achieve. To overcome this issue, we propose to align the EEG features with the interpolation of the fine and coarse-grained information, which represent the generated captions and category annotation respectively.

To this end, we first obtain the CLIP encoding c_{label} of EEG category annotations. We then add fully connected layers with residual connections as alignment network to align the EEG time-frequency embeddings, obtained from the last hidden layer of the EEG classifier, with c_{cap} and c_{label}, and proceed to fine-tune the entire model. Technically, we maximize the cosine similarity between the output of the alignment network c_{EEG} and c_{label} and c_{cap}. This aims to identify the directions of category semantics, highlight the primary components of the captions, and ease the alignment difficulty. We define the loss function L_{si} for the semantic interpolation alignment as follows:

$$L(a, b) = 1 - \frac{a \cdot b}{|a||b|}, \tag{4}$$

$$L_{si} = L(c_{cap}, c_{EEG}) + L(c_{label}, c_{EEG}). \tag{5}$$

3.5. Cascaded Diffusion Models

Due to the inevitable imperfect alignment, there exist information loss and semantic noise in our reconstruction process. To mitigate the negative impact, we propose cascaded diffusion models. We adopt different levels of semantics as conditional inputs for the pre-trained stable diffusion [27] model to guide the image reconstruction process.

First, we use the aligned c_{EEG} as condition, performing reverse diffusion to reconstruct images. We observe the quality of the reconstructed images is low, due to the information loss and noise in c_{EEG}. Therefore, we propose to further refine the reconstructed images, which serve as the new input for the diffusion model. During this refinement process, we propose to use the EEG classification labels obtained from Section 3.3 as the new condition. Since the annotations of these predicted labels are unambiguous, hence the reconstructed results from the refinement stage are more precise and clear. By doing so, the reconstruction accuracy is mainly affected by EEG classification results, which is more correct than c_{EEG}.

To summarize, introducing the cascaded diffusion models for refinement effectively remedies the noised image reconstruction, yielding higher-quality reconstruction results.

4. Experiments

4.1. Dataset and implementation

Dataset. Our experimental setup employed an EEG-image pairs dataset [32], encompassing 11,466 EEG sequences recorded from 128-channel electrodes. These sequences were associated with 40 distinct image classes. To elicit neural responses, six subjects were presented with a total of 2000 object images. These visual stimuli were selected from a subset of the widely used ImageNet [28] dataset and comprised 40 classes, each containing 50 easily recognizable images. To ensure the exclusion of potential interference from previously shown images, the first 40 ms of each recorded EEG sequence were discarded, and the following 440 ms of EEG data were utilized for the experiments. The dataset was split into training, validation, and testing sets using an 8:1:1 ratio. For comprehensive dataset characteristics, please refer to [32]. Besides, our primary objective focused on achieving accurate image semantic reconstruction. Therefore, we conducted evaluations at the semantic level rather than measuring pixel-level accuracy.

Implementation details. We follow the optimal parameter settings of prior related works [2, 5] and empirically set our parameters in Section 3: $c = 128$, $l = 440$, $n = 110$, $d = 1024$, $r_M = 0.75$. We show the ablative analysis on the total number of codewords n_t in Table 1. Based on the classification accuracy is highest, we set $n_t = 660$. Moreover, the pre-trained stable diffusion model is v1-5-pruned-emaonly, while the pre-trained CLIP encoder is ViT-L/14. Transformer for visible feature extraction consists of 8 self-attention blocks, and for latent masked reconstruction comprises 4 cross-attention blocks. In each block, the attention heads are set to 16, and feed-forward network dimension is fixed at 4096. Additionally, the LSTM size is specified as 128. We employ a learning rate of 0.001 with a batch size of 128. Pre-training epochs are 300 for time encoder, while training epochs are 900 for frequency encoder. Fine-tuning for time encoder lasts 80 epochs, and joint fine-tuning for time and frequency encoder lasts 30 epochs. Cross-modal EEG alignment fine-tuning is performed for 200 epochs.

4.2. Evaluation metrics

EEG Classification Accuracy (CA). CA measures the accuracy of visual classification performed on EEG signals. It reflects the extent to which EEG's latent features are correlated with coarse-grained visual semantics.

N-way Top-K Classification Accuracy of Generation (GA). Following the methodology of [2, 5], we utilized N-way Top-K accuracy to evaluate the semantic correctness of the reconstruction results. For multiple trials, the top-K and classification accuracies were calculated based on $N - 1$ randomly selected classes, along with the correct class. Both the ground-truth and generated images were ini-

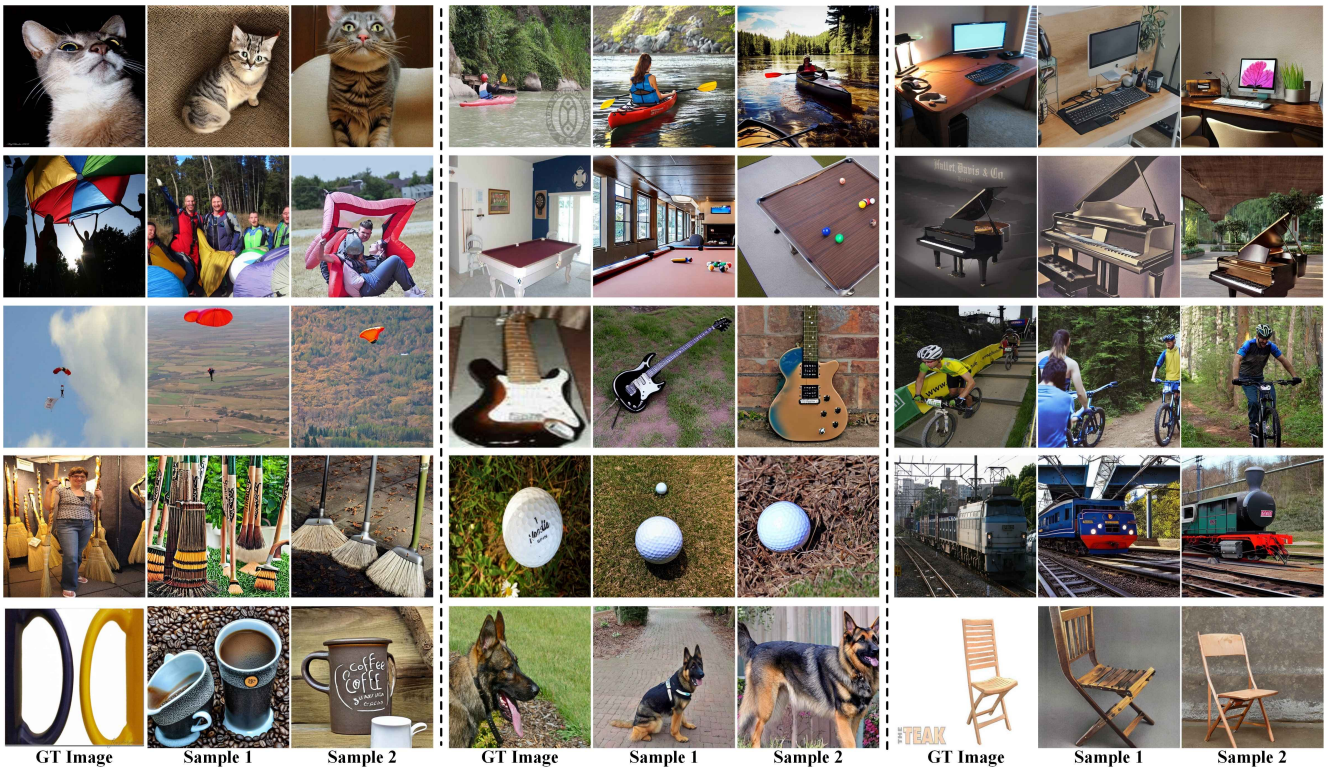

Figure 4. Visual stimuli reconstruction results (partial) and the ground truth (GT) images. We are able to not only reconstruct the correct types but also capture their fine-grained semantics to some extent, like prominent numerical, color, environmental, or behavioral features.

tially inputted into a pre-trained ImageNet1K classifier [7]. We then examined whether the top-K classification among the N selected classes matched the ground-truth classification. In our case, we set $N = 50$ and $K = 1$.

Inception Score (IS). IS is employed to evaluate the diversity of the reconstructed images, providing insights into their quality. For a detailed definition, please refer to [29].

Structural Similarity Index Measure (SSIM). SSIM quantifies the similarity between images based on criteria such as brightness, contrast, and structure.

CLIP Similarity (CS). Following the methodology of [24], we utilized a pre-trained CLIP ViT-B/32 Encoder to obtain deep feature vectors of both the generated and ground-truth images. The cosine similarity between these vectors was then calculated. This indicator intuitively reflects the semantic similarity between images.

4.3. Image Reconstruction Visualization

Figure 4 presents part of the reconstructed results. We not only reconstruct the coarse-grained categories, but also to a certain extent, preserve their fine-grained information. For instance, considering the five sets of images presented in the first column from top to bottom, we successfully reconstruct the behavior of a cat looking up, distinguish between the states of parachute on the ground and in the air while preserving color similarity, and approximately establish the quantitative relationship between brooms and cups and their respective ground truth (GT) images.

n_t	-	220	330	**660**	990
CA	0.3804	0.4097	0.4056	**0.4130**	0.3286

Table 1. The impact of codewords number n_t on EEG classification accuracy (CA). '-' indicates that no semantic unit classification objective has been performed during the self-supervised pre-training of time encoder. The experiment is conducted on all subjects, taking the average value of CA. In addition, the model used in the experiment is consistent with 'model 3' in Table 2.

4.4. Comparison against other methods

Baselines. Our baselines for comparison include Brain2Image [16] and DreamDiffusion [2]. Brain2Image leverages conventional GAN [11] for image reconstruction, while DreamDiffusion utilizes latent diffusion model [27] and relies on an additional dataset. However, it is worth noting that these baseline methods exhibit limitations in terms of evaluation metrics. For instance, Brain2Image can only be assessed with IS since their definition of accuracy is unclear and no code is provided. DreamDiffusion only re-

Figure 5. Comparison of reconstructed images' quality between our method and prior works.

Figure 6. Comparison of reconstructed images with DreamDiffusion on the same ground truth (GT) image.

Methods	Subj	GA	IS	SSIM	CS
Ours	1	0.7099	7.0325	0.7552	0.6617
	2	0.3905	7.3429	0.7531	0.5854
	3	0.4720	7.9862	0.7522	0.5999
	4	**0.4927**	8.6232	0.7517	0.6002
	5	0.3588	7.5774	0.7518	0.5939
	6	0.3035	7.2144	0.7506	0.5739
	AVG	0.4546	**7.6294**	0.7524	0.6025
B2I [16]	**AVG**	-	**5.0700**	-	-
DD [2]	**4**	**0.4580**	-	-	-

Table 2. Quantitative comparative experiment of reconstruction results. To simplify the description, B2I means Brain2Image, while DD means DreamDiffusion. Besides, GA is N-way Top-K accuracy, IS is inception score, SSIM is structural similarity index measure, CS is CLIP similarity, and AVG is average.

ports N-way Top-K accuracy. In order to provide a more comprehensive analysis, we incorporate additional metrics including SSIM and CS to expands the evaluation framework. Furthermore, DreamDiffusion was solely evaluated on the fourth subject, whereas our experimentation encompasses all six subjects, offering a more comprehensive basis for future studies and facilitating wider applicability of our findings. Notably, results of prior methods used for comparison are directly taken from the original papers.

Quantitative results. We conducted image generation experiments involving all six subjects in our study, generating and evaluating four images for each EEG sample. The quantitative results are presented in Table 2. Our method achieved a GA of 0.4927 on subject 4, surpassing the performance of DreamDiffusion, which attained a GA of 0.4580. As DreamDiffusion was solely evaluated on subject 4, a direct comparison regarding its performance on other subjects cannot be made. Additionally, our method achieved higher IS compared to the Brain2Image, with an average of 7.6294 across all subjects. These results signify that our method outperforms the current state-of-the-art in terms of both semantic fidelity reconstruction and generation quality. Moreover, our method achieved average SSIM and CS values of 0.7524 and 0.6025, respectively, across all subjects, providing potential new baselines for future research endeavors.

Qualitative results. In Figure 5, we adopt a qualitative comparison approach employed by previous studies to evaluate the quality of reconstructed images. Specifically, we compare our reconstructed results for three distinct image categories (Airliner, Jack-o'-Lantern, and Panda) with the results of Brain2Image and DreamDiffusion. The comparative analysis highlights the higher quality of our generated images, underscoring the effectiveness of our proposed method. In Figure 6, we compare our reconstructed images with DreamDiffusion at the same GT images. Since Brain2Image did not provide corresponding GT images for

Model	TiE	FrE	PT	FA	FS	subj1	subj2	subj3	subj4	subj5	subj6	Average
1	T	F	-	1	-	0.3742	0.2515	0.2927	0.2805	0.2160	0.2547	0.2783
2	T	F	1	-	-	0.0982	0.3313	0.2744	0.3597	0.2596	0.1740	0.2495
3	T	F	1	2	-	0.7362	0.3926	0.3654	0.4146	0.3210	0.2484	0.4130
4	T	F	2	1	-	0.2086	0.3497	0.3720	0.3293	0.3333	0.2919	0.3141
5	T	F	1	-	2	0.1227	0.3313	0.3902	0.3659	**0.3704**	0.2484	0.3048
6	T	F	1	2	3	0.7485	0.3988	0.4329	0.4329	0.3456	0.2795	0.4397
7	F	T	1	-	-	0.5547	0.1484	0.2344	0.2188	0.2031	0.1875	0.2578
8	F	T	1	-	2	0.5938	0.1875	0.2656	0.2578	0.2578	0.2031	0.2943
9	T	T	1	2	-	0.7178	0.4049	0.4024	0.4085	0.2778	0.2919	0.4172
10	T	T	1	-	2	**0.7607**	**0.4172**	**0.4695**	**0.4878**	0.3518	**0.3726**	**0.4766**
11	T	T	1	2	3	0.7546	0.4110	0.4451	0.4695	0.3333	0.3168	0.4551

Table 3. The ablation analysis of EEG classification accuracy (CA). For simplicity, TiE means Time Encoder, FrE means Frequency Encoder, PT means pre-training TiE, FA means TiE and FrE jointly fine-tuning on all subjects, while FS means fine-tuning on single. The number under PT, FA and FS means their order of training. If FS is not '-', each subject corresponds to a separate fine-tuned model.

reconstructed results, we can only compare with DreamDiffusion on limited GT images. Subjectively, our reconstructed results exhibit greater likeness to GT images.

4.5. Ablation Study

In the EEG visual classification experiments, we evaluate the model's representation of visual semantics by considering the time domain, frequency domain, and their fusion. We perform ablations on frequency encoder and time encoder to assess their contributions. We also analyze the impact of time encoder pre-training and the order of pre-training and fine-tuning. Additionally, we optimize the training strategy by controlling data conditions during fine-tuning. Classification experiments provide clear assessments with less computational resources compared to generating images. Table 3 illustrates the highest average classification accuracies: 0.4397 for the time domain, 0.2943 for the frequency domain, and 0.4766 for their fusion. This indicates that combining the time and frequency domains enhances the model's ability to embed visual semantic features of EEG. Furthermore, both the combined temporal-frequency model and the frequency domain model are more suitable for fine-tuning on individual subjects, unlike the model that only considers the temporal domain. This can be attributed to the frequency domain features, which introduce more shared and personalized features. Additionally, pre-training of time encoder is necessary, and pre-training followed by fine-tuning proves to be more effective. Considering various options, we select the optimal approach (model 10) for CLIP alignment and image generation.

4.6. Discussion

Figure 7 shows part of the incorrect reconstructed results. There are noticeable resemblances between these generated

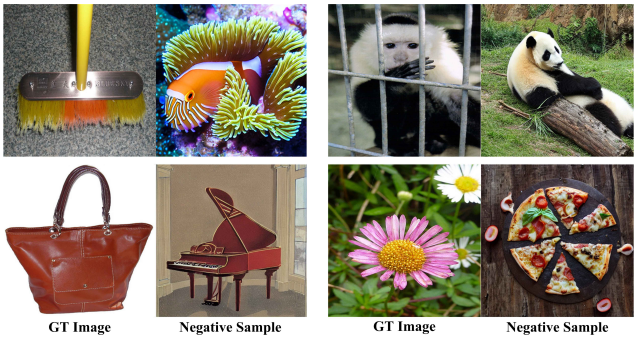

Figure 7. Examples of incorrect reconstruction results.

samples and real images, particularly in terms of color or shape. This implies that the visual information encoded in human EEG may prioritize fundamental properties of objects rather than specific object categories. The classification of visual stimuli from EEG involves decoding and categorizing information pertaining to color, shape, environment and various other features. Building upon this hypothesis, decoding individual properties of visual information, such as color and shape, in isolation may offer an effective avenue for EEG analysis and visual reconstruction.

5. Conclusion

The paper proposes BrainVis, a novel pipeline for image reconstruction from EEG signals. BrainVis leverages self-supervised learning to extract time-domain features from EEG and employs LSTM to capture frequency-domain features from the FFT results. It aligns the EEG time-frequency features with semantically interpolated CLIP features. The aligned results and the predicted category embedding are served as the conditions of cascaded diffusion

models for image reconstruction. BrainVis surpasses the state-of-the-art technique in terms of semantic reconstruction and generation quality. Furthermore, it eliminates the need for additional data and overcomes previous limitation that was limited to only coarse-grained reconstruction. We also uncover possibly visual components in EEG from negative samples, contributing to a deeper understanding of how the brain processes visual stimuli.

References

[1] Nibras Abo Alzahab, Luca Apollonio, Angelo Di Iorio, Muaaz Alshalak, Sabrina Iarlori, Francesco Ferracuti, Andrea Monteriù, and Camillo Porcaro. Hybrid deep learning (hdl)-based brain-computer interface (bci) systems: a systematic review. *Brain sciences*, 11(1):75, 2021. 1

[2] Yunpeng Bai, Xintao Wang, Yanpei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023. 1, 2, 3, 4, 5, 6, 7

[3] A Banaszkiewicz, J Matuszewski, M Szczepanik, B Kossowski, P Mostowski, P Rutkowski, M Śliwińska, K Jednoróg, K Emmorey, A Marchewka, et al. The role of the superior parietal lobule in lexical processing of sign language: Insights from fmri and tms. *Cortex*, 135:240–254, 2021. 1

[4] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[5] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023. 2, 5

[6] Mingyue Cheng, Qi Liu, Zhiding Liu, Hao Zhang, Rujiao Zhang, and Enhong Chen. Timemae: Self-supervised representations of time series with decoupled masked autoencoders. *arXiv preprint arXiv:2303.00320*, 2023. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. 6

[8] Bing Du, Xiaomu Cheng, Yiping Duan, and Huansheng Ning. fmri brain decoding and its applications in brain–computer interface: A survey. *Brain Sciences*, 12(2):228, 2022. 2

[9] Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, 33:13038–13048, 2020. 1, 2

[10] Simon Geirnaert, Servaas Vandecappelle, Emina Alickovic, Alain de Cheveigne, Edmund Lalor, Bernd T Meyer, Sina Miran, Tom Francart, and Alexander Bertrand. Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. *IEEE Signal Processing Magazine*, 38(4):89–102, 2021. 1

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2, 6

[12] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. 2

[13] Bin He, Han Yuan, Jianjun Meng, and Shangkai Gao. Brain–computer interfaces. *Neural engineering*, pages 131–183, 2020. 1

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3

[15] Fabian Hutmacher. Why is there so much more research on vision than on any other sensory modality? *Frontiers in psychology*, 10:2246, 2019. 1

[16] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817, 2017. 1, 2, 6, 7

[17] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. 2

[18] Ashima Khosla, Padmavati Khandnor, and Trilok Chand. A comparative analysis of signal processing and classification methods for different applications based on eeg signals. *Biocybernetics and Biomedical Engineering*, 40(2): 649–690, 2020. 1

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[20] Chi Qin Lai, Haidi Ibrahim, Mohd Zaid Abdullah, Jafri Malin Abdullah, Shahrel Azmin Suandi, and Azlinda Azman. Artifacts and noise removal for electroencephalogram (eeg): A literature review. In *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pages 326–332. IEEE, 2018. 1

[21] Young-Eun Lee, Seo-Hyun Lee, Sang-Ho Kim, and Seong-Whan Lee. Towards voice reconstruction from eeg during imagined speech. *arXiv preprint arXiv:2301.07173*, 2023. 1

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3, 5

[23] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636, 2022. 1, 2

[24] Yizhuo Lu, Changde Du, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. *arXiv preprint arXiv:2303.14139*, 2023. 2, 6

[25] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 4, 5, 6

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. pages 211–252. Springer, 2015. 5

[29] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. 2016. 6

[30] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019. 2

[31] Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. Eeg2image: Image reconstruction from eeg brain signals. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1, 2

[32] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6809–6817, 2017. 1, 5

[33] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023. 1, 2

[34] Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. *arXiv preprint arXiv:2306.11536*, 2023. 2

[35] Praveen Tirupattur, Yogesh Singh Rawat, Concetto Spampinato, and Mubarak Shah. Thoughtviz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 950–958, 2018. 1, 2

[36] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[37] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12):4136–4160, 2018. 2

[38] Minchao Wu, Wei Teng, Cunhang Fan, Shengbing Pei, Ping Li, and Zhao Lv. An investigation of olfactory-enhanced video on eeg-based emotion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 1602–1613, 2023. 1

[39] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 4