ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240732

Hidden States in LLMs Improve EEG Representation Learning and Visual Decoding

Aoyang Liu[a,b,c], **Haodong Jing**[a,b,c], **Yulong Liu**[a,b,c,d], **Yongqiang Ma**[a,b,c,*] **and Nanning Zheng**[a,b,c]

[a]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
[b]National Engineering Research Center of Visual Information and Applications
[c]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[d]The Hong Kong University of Science and Technology
ORCID (Aoyang Liu): https://orcid.org/0009-0009-3930-3221, ORCID (Haodong Jing): https://orcid.org/0000-0001-6643-7588, ORCID (Yulong Liu): https://orcid.org/0000-0001-9840-1512, ORCID (Yongqiang Ma): https://orcid.org/0000-0002-6063-5601, ORCID (Nanning Zheng): https://orcid.org/0000-0003-1608-8257

Abstract. Analyzing brain signals and reconstructing visual stimuli from the brain can facilitate further exploration on cognitive functions of the human brain, which have attracted strong interest in neuroscience and artificial intelligence. However, due to defects such as complex noises and the lack of alignment accuracy, efficient methods for extracting information from Electroencephalogram (EEG) signals are still very limited, making it difficult to perform EEG visual decoding tasks. Our study shows a way to handle the issues by proposing a new method for EEG representation learning and visual decoding, thus completing end-to-end image reconstruction tasks from EEG signals. We utilize the ability of semantic extraction and prediction of large language models (LLMs) to enhance the performance of EEG feature extraction. For semantic representation learning, we align EEG signals with target semantic embeddings, which are obtained from hidden states of Large Language Model Meta AI 2 (LLaMa-2) by inputting descriptions of images into the model. We also extract visual features from EEG signals to improve the quantity of the reconstructed images at low levels. Then we fuse semantic features and visual features by applying a pre-trained diffusion model and finally generate the corresponding images. We are the first to incorporate the LLM into EEG visual decoding tasks. Our method achieves the state-of-the-art result of EEG classification accuracy and the quality of reconstructed images on ImageNet-EEG datasets. In one word, our work is an important step forward in the field of exploiting the relationship between language models and human visual cognition. Our codes are available at https://github.com/lay-atsa/llm4eeg.

1 Introduction

Rapid development of science and technology is making the artificial general intelligence realizable. The intersection of neuroscience and computer science has brought unprecedented opportunities, particularly in understanding and simulating cognitive processes of the human brain. As a non-invasive neurophysiological technique used to record the neural activity of the brain, Electroencephalogram (EEG) has attracted many researchers. By capturing the potential changes in brain electrical signals, EEG can provide high-temporal-resolution information on brain activity, which reveals dynamic properties associated with various neurophysiological processes. These characteristics make EEG an important tool to study the relationship between neural activity and cognitive functions[20][13]. In recent years, numerous researchers have successfully utilized deep neural networks to decode visual information from brain signal representations[37][28]. However, these studies are all facing common challenges and limitations. Firstly, the noise and sparsity collections of EEG signals limit the precision of information extraction. Secondly, the decoding capability of existing methods for semantic information is relatively limited in complex contexts, especially for the representation of abstract concepts. Moreover, a comprehensive understanding of the dynamic changes in semantic information within EEG signals is lacking, restricting adaptability to different tasks and cognitive states.

Difficulties in semantic information extraction from EEG may be alleviated by the current development of large language models (LLMs). Some key efforts focus on using pre-trained LLMs, such as the generative pre-trained transformer(GPT) series[33][1], BERT[10], etc., to extract semantic information from text and transform it to perform other tasks. These models can capture associations among words, phrases, and sentences, encoding rich semantic information at the textual level, which can be also transferred to downstream tasks. Therefore, as an effective encoding result of textual information, it may be further used in EEG semantic decoding tasks, such as brain-to-text generation and brain-to-semantic recognition.

However, beyond semantic information extraction, another challenge in this field is how to reconstruct stimuli from visual information and enhance reconstruction performance. By combining the obtained semantic information with generation models, researchers attempt to generate images that match the textual descriptions, ensuring consistency between the generated images and the provided semantic features[2]. The advantages of these methods lie in their ability to generate images from text in an end-to-end manner without explicitly providing image samples. This provides a new pathway for image generation tasks based on semantic information.

* Corresponding Author. Email: musayq@xjtu.edu.cn

Aiming to enhance the decoding performance of EEG signals, in this study, we propose a method, introducing the LLM into EEG visual decoding and stimuli image reconstruction tasks. Based on the similarity between the LLM and human cognitive representations, we utilize the contextual understanding ability of LLMs to generate the feature representations as the target semantic features of EEG to extract semantic features from EEG signals. Specifically, we obtain the representations from the hidden states of the pre-trained LLaMa-2 by inputting text descriptions of images, and then perform contrastive learning to align the EEG representation and the LLaMa-2 embedding space. We also extract visual features from EEG signals and subsequently apply a pre-trained diffusion model to generate images. By selecting images from the training dataset and mapping the EEG semantic features to the textual embedding space, We effectively fuse features from two modalities and thus reconstruct images which are similar as stimuli images in both semantic and visual features. The experimental results indicate that our method has achieved optimal levels in both quantitative and qualitative evaluations of semantic classification accuracy and image reconstruction metrics. The overview of our framework is shown in Figure 1.

Our novel approach is designed to combine language models and brain activity data, which overcomes some of the constraints observed in previous studies, offering a new perspective for a more accurate and comprehensive interpretation of information within EEG signals. Overall, the integration of a pre-trained LLM effectively enhances the analysis and fitting of cognitive information, contributing to the further development of general artificial intelligence.

In this study, our main contributions are as follows:

- We are the first to introduce the LLM into EEG visual decoding. By incorporating the LLM, accurate semantic features can be extracted from EEG signals, demonstrating the possibility of applying language models in visual decoding tasks.
- Beyond semantic decoding and classification, we implement visual feature extraction and fuse the two features efficiently by applying the diffusion model, completing end-to-end image reconstruction tasks from EEG signals.
- We validate the effectiveness of method on ImageNet-EEG dataset. Our method achieves the state-of-the-art result of EEG classification accuracy and the quality of reconstructed images, which motivates further exploration on the relationship between language models and human visual processing.

2 Related Works

2.1 Decoding Visual Stimuli from EEG

Decoding visual information from functional magnetic resonance imaging (fMRI) and EEG involves extracting patterns of brain activity associated with visual stimuli or other corresponding information and using machine learning techniques to predict or classify specific visual properties or categories. fMRI measures changes in blood oxygen level-dependent (BOLD) signals, which provides high spatial resolution, allowing researchers to localize brain activity with precision. Reconstructing images from fMRI data has made significant progress[46][16][31][8]. Compared with fMRI, EEG provides excellent temporal resolution, capturing rapid changes in brain activity with millisecond precision. It measures electrical potentials generated by neural activity in the brain by recording the electrical activity of the brain using electrodes placed on the scalp. However, due to the limited spatial resolution, high noises, rapid changes and the lack of data, there are few effective methods for visual decoding, especially reconstructing images from EEG signals.

To tackle these challenges, one approach involves using Long Short-Term Memory (LSTM)[39] networks for EEG information classification and feature embedding. Besides, variational autoencoders (VAEs) and generative adversarial networks (GANs) have been applied to generate images from EEG features[17], but the achieved results have not been sufficiently significant. Some researchers have attempted to bind GANs with conditional progressive growing for perceptual image generation[18]. Some researchers have focused on cross-modal multi-mode processing, achieving coarse-grained classification[4]. With the development of diffusion models, aligning spatiotemporal information from EEG with image features enables high-resolution image reconstruction.

2.2 Diffusion Models

For EEG visual decoding tasks, in addition to just obtaining the features of the images, we would like to further reconstruct the images. To solve this problem, generative models is an effective tool for image reconstruction. As a class of generative models, diffusion models have gained attention for their ability to model complex data distributions, particularly in the realm of natural images and text. These models operate by iteratively applying a series of transformations to a random noise input, gradually generating samples that approximate the true data distribution[15]. Diffusion models have emerged as a groundbreaking approach in multimodal content generation[43], natural language processing[44], and decision-making[3]. Implicit Diffusion Models (IDM)[12] were introduced in image generation, where pre-trained autoencoders effectively addressed the pixel-wise estimation shortcomings in the latent space, leading to a noticeable improvement in the quality of generated images, which has been employed in visual decoding. MinD-Vis[8], has achieved high-quality and credible results in fMRI image reconstruction by using IDM. BrainCLIP[25] aligned fMRI data with visual and textual data to implement the fMRI-to-image generation task by the fmri-guided diffusion model. However, similar approaches have not been effectively applied in the reconstruction of visual images from EEG signals.

Diffusion models can be used for image-to-image generation by adding noises and removing noises.[35] Given an input sample image as the initial image x_0, noises ϵ are added to it step by step, and the noised image can be estimated as follows:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where x_0 refers to the initial image, $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$, β_t refers to the variance schedule.

During the reverse process, the model predicts the added noise ϵ_θ and the denoised image can be obtained by the following equation:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta). \tag{2}$$

2.3 Large Language Models

Large language models (LLMs), which are trained on a large scale of text datasets to predict the probability distribution of the next word in a sequence, can perform various language-related tasks, such as translation, text summarization, question answering, and more[45]. LLMs can capture rich semantic and syntactic information from the text data and effectively learn a powerful language representation. Generative pre-trained transformer(GPT)[33] is one of the

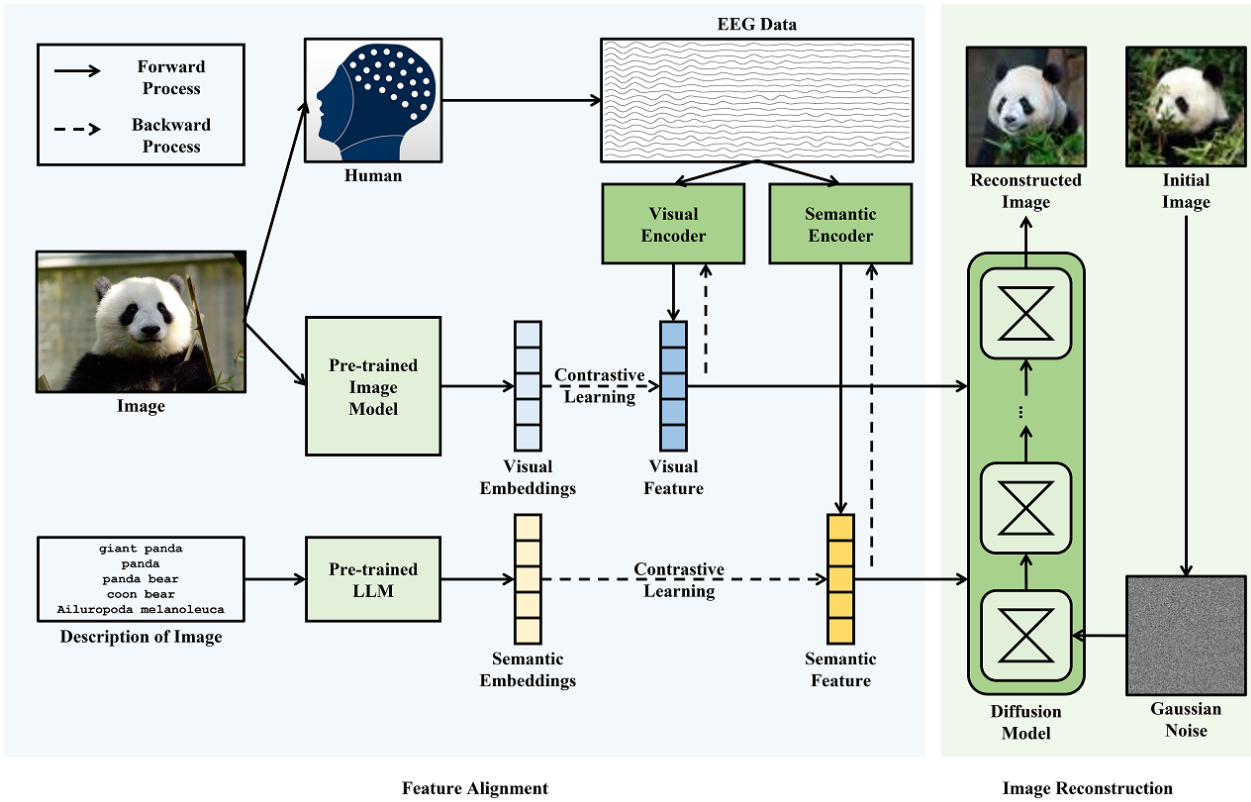

Figure 1. Overview of our architecture. First, we train two EEGNet models through contrastive learning to extract the semantic features and visual features from EEG signals, respectively. Then, we apply the pre-trained diffusion model to generate images that match both semantic features and visual features.

most powerful LLM architectures, which employs multi-head self-attention[42] to combine the representations of each word in a sequence with those of preceding words. Building upon the powerful GPTs, researchers have proposed models such as LLaMa[41], and Vicuna[9]. Multimodal LLMs are designed to handle diverse types of inputs beyond text. Blip[24] utilizes Q-formers to project multimodal inputs into the text space, while others simply train a fully connected layer as a projector.

The LLM can be used for extracting semantic features from language stimuli. To successfully perform the next-word prediction task, it learns to extract features that capture the meaning of the input sequence. Large Language Model Meta AI 2 (LLaMa-2) is one of the generative pre-trained models. It employs a neural network with billions of variables, using the same transformer architecture as the GPTs. However, unlike GPTs, it is an open-source LLM and the pre-trained model is accessible online. LLaMa-2 outperforms open-sourse models and is comparable to GPT-3.5 on several benchmarks.

It is observed that representations from language models share similarities with cognitive representations. For example, semantic knowledge can predict patterns of brain dynamics and thus discriminate stimuli categories[27][29]. For LLMs, pre-trained on larger datasets with a larger scale of parameters, there are more similarities. Semantic representations can predict how well human subjects understand the text[7]. Attention weights can predict brain activity accurately[22]. Word embeddings help the annotation task of semantic cognition[22]. Therefore, we hypothesize that incorporating LLMs can improve the performance of EEG decoding tasks. Researchers have applied LLMs to implement semantic decoding for text comprehension tasks, generating corresponding text from brain activity without markers[11][40]. However, the relationship between LLMs and visual conceptual cognition has not been studied.

3 Method

3.1 Preliminary

Contrastive Language-Image Pre-training (CLIP). CLIP[34], firstly proposed in the image-text pairing task, aims to learn to associate images with their textual descriptions by constrastive learning. For multimodal feature alignment, it is an efficient method that has been used to align brain activities and the feature of stimuli in brain decoding tasks. In our method, we utilize the loss function proposed in CLIP, denoted as CLIP loss.

Given a batch of embeddings of data $A = \{a_i | i = 0, ..., M\}$ and $B = \{b_i | i = 0, ..., M\}$ from two modalities, the contrastive loss is given by

$$Con(A, B; \tau) = -\frac{1}{M} \sum_{i=0}^{M} \log[\frac{\exp(\cos(a_i, b_i)/\tau)}{\sum_{j=0}^{M} \exp(\cos a_i, b_j/\tau)}], \quad (3)$$

where M is the batch size, τ is a temperature hyper-parameter, and the (A_i, B_i) denotes the matched embedding pair of data from two modalities while other $(A_i, B_j)_{j \neq i}$ denote unmatched pairs.

To perform bi-directional contrastive learning between two modalities, the CLIP loss is as follows:

$$L_C(A, B) = \frac{1}{2}(Con(A, B; \tau) + Con(B, A; \tau)). \quad (4)$$

By minimizing the CLIP loss, the model will learn similar representations on cosine measure between two samples from matched meaning but different modalities in the common embedding space, thus we could get an efficient representation from raw data.

Pre-trained models (LLaMa-2 & VGG-19) In our task, we aim to use the large language model to process the semantic information

of the description of the images. There are three different sizes of LLaMa-2 model: 7b, 13b and 70b. Among these sizes, we chose the LLaMa-2 7b, which is swift and suitable for basic tasks like summaries or categorization.

Different latent image representations can affect the efficiency of EEG decoding on visual stimuli. Those image models pre-trained on supervised tasks may lead to good performance. For example, the Visual Geometry Group (VGG) pre-trained model[38] refers to a family of convolutional neural network (CNN) architectures. These models are widely used in computer vision tasks, particularly in image classification. Their simplicity and uniform architecture, which consists of stacking multiple convolutional layers followed by max-pooling layers allows the models to capture increasingly complex features from the input image as information flows through the network. Here, in order to extract the visual features of the images more efficiently, we choose the pre-trained VGG-19 model, which is proved to achieve high performance on image retrieval tasks on brain visual decoding[6].

3.2 Task Definition

Our task is to decode visual stimuli from the brain activity recorded when subjects are presented a set of images. Let the dataset of (EEG, image) pairs be $\Omega = (X_i, Y_i)$, $X_i \in S^{C\times T}$ be the recorded piece of EEG signals and $Y_i \in I^{H\times W\times 3}$ be the presented images simultaneously, where C is the channel number of EEG signals, and T is number of time points in a time window. EEG data are recorded when subjects are watching the stimuli images. The target of our research is to generate an image from the piece of EEG signals, aiming to generate images that closely resemble the real images at both high and low levels of features. The schematic of the EEG visual decoding tasks is shown in Figure 2.

3.3 Multimodal Feature Extraction

Both low-level visual features (colors, shapes, structures) and high-level semantic features (categories and content) can elicit corresponding patterns in brain activities. Therefore, it is feasible to extract features at both levels from raw EEG signals to perform image reconstruction. Here we use deep neural networks to accomplish automatic feature learning and extraction by training them.

The EEG encoder networks need to process the temporal and spatial information on multi scales. EEGNet[23], a compact convolutional neural network, which contains a convolutional block, a depthwise convolutional block and a separable convolutional block, shows high accuracy in many EEG classification tasks without specifying the paradigm of the tasks. There are fewer parameters in the depthwise block and the separable block compared with common convolutional layers, which helps the model to learn a more efficient representation and perform higher generalization. We modified the network structure of EEGNet by adding a convolutional block between the temporal convolutional block and the depthwise block, in order to integrate information from different channels. We adopt this architecture on both the semantic encoder and the visual encoder, and denote as E_S and E_V, respectively.

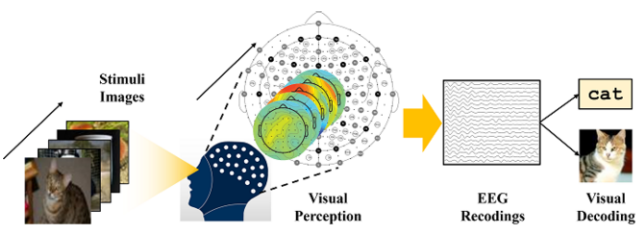

Figure 2. The schematic of the EEG visual decoding tasks. EEG data are recorded when subjects are watching the stimuli images. Then, merely by analyzing the EEG recordings, the category labels of the images are decoded and the images are reconstructed.

The output feature dimension is set to the same as the target representations, which are obtained from the pre-trained models. After training, the encoder networks project the EEG signals to the target embedding space and predict the corresponding features from EEG signals. The setting details of convolutional layers in the networks are shown in Table 1.

Table 1. Setting details of four convolutional layers in the two encoder networks. Top: semantic encoder network. Bottom: visual encoder network.

Location		Channels	Kernel Size	Padding	Groups
Semantic Encoder Network	1	$1 \rightarrow 64$	(1,41)	(0,20)	1
	2	$64 \rightarrow 2048$	(128,1)	0	64
	3	$2048 \rightarrow 2048$	(1,15)	(0,7)	2048
	4	$2048 \rightarrow 64$	(1,1)	0	1
Visual Encoder Network	1	$1 \rightarrow 16$	(1,41)	(0,20)	1
	2	$16 \rightarrow 64$	(14,1)	0	1
	3	$64 \rightarrow 64$	(1,15)	(0,7)	64
	4	$64 \rightarrow 16$	(1,1)	0	1

We train these two networks on the training dataset and choose the epoch according to the validation dataset as the final encoders.

3.3.1 Semantic Features

As a language model, when LLaMa-2 is given the input text, the model will be activated and the hidden states can reflect the semantic features of the input text. Previous studies have shown that the best semantic features can be extracted from the middle layers of the language models[40]. Among the 33 layers in the LLaMa-2 7b model, we select the twentieth layer and record the hidden states of that layer. For each stimuli image, we put the description from the ImageNet dataset to the model and take the mean value of the hidden states on the first dimension as the semantic representation of the image stimuli.

The width of the hidden layer in the LLaMa model is 4096. To improve the efficiency of the learned semantic representation, we map the average of the hidden layer output by a fully connected network into a 160-dimension embedding space. Accordingly, we also set the output dimension of the EEG semantic encoder to 160. Suppose the textual description of an image t, the final semantic feature of the image is denoted as $SE(t)$, and the EEG signal is denoted as x_i. Then we optimize the weights of the semantic encoder network and the fully connected network by the following loss function:

$$\begin{aligned} Loss_S(\theta_S) =& \lambda_1 L_C(SE(t), E_S(x_i)) \\ &+ (1-\lambda_1) MSE(SE(t), E_S(x_i)), \end{aligned} \tag{5}$$

where x_i denotes an EEG signal sample.

The loss function consists of two parts, the CLIP loss $L_C(\cdot,\cdot)$ and the MSE loss $MSE(\cdot,\cdot)$. The CLIP loss is used to make the cosine distance between the semantic features of the image labels within a batch and the encoding result of the EEG data collected under the corresponding visual stimulus as small as possible, and to make the cosine distance of the mismatched features as large as possible, while the MSE loss is used to make the absolute distance between the encoding result of the EEG data and the semantic features as close as possible. In details, we set the hyperparameter $\lambda_1 = 0.8$. We use the

AdamW[26] optimizer to train the networks and the weight decay is set to 0.0001. The batch size is 16, the learning rate is 0.001, and the number of training epochs is set to 150.

3.3.2 Visual Features

The output dimension of the VGG-19 pre-trained model is 1000. By feeding all the images of the training set into the network, we obtain the visual features of the training images. Before training the visual encoder network, We implement principal component analysis(PCA) and extract the top 20 features, which contribute to 80.01% variance of the original features.

The structure of the visual encoder model and the loss function are similar to the semantic encoder. Suppose an image i, the latent features of the image after PCA are denoted as $IE(i)$. The loss function is as follows:

$$\begin{aligned} Loss_V(\theta_V) =& \lambda_2 L_C(IE(i), E_V(x_i)) \\ &+ (1-\lambda_2) MSE(IE(i), E_V(x_i)). \end{aligned} \tag{6}$$

And we also set the hyperparameter $\lambda_2 = 0.8$. We use the Adam[19] optimizer to train the network. The batch size is 32, the learning rate is 0.001 and the number of training epochs is set to 200.

3.3.3 Feature Prediction from EEG

After training two encoder networks, we can predict the semantic features and the visual features from the EEG signals.

For semantic features, the EEG signals are fed into the semantic encoder network to obtain the corresponding semantic features for two purposes. On the one hand, mapping to the glide embedding space serves as textual guidance information for the diffusion process, allowing the diffusion model to generate images that match the semantic input. On the other hand, by training a linear mapping layer as a classifier, we can predict the category of the original stimulus image from semantic features.

At the same time, EEG signals are input to the visual encoder network to obtain the corresponding visual features for the subsequent selection of images as the initial condition for the diffusion process.

3.4 Generating Images by Diffusion Models

To perform generalized visual decoding tasks, we use diffusion models to generate images, which is a step further than just decoding the categories. Based on the feature extraction and prediction from the EEG signal, a pre-trained glide model[30] is used for image generation. The number of time steps is 100 and the guidance scale is 7.0.

The predicted image category is obtained from the semantic features of EEG, and all images in the training set corresponding to this category are used as candidate images. The visual encoder network of EEG has been optimized by contrastive learning, so the learned features can be used to measure the similarity with the target features. The vector cosine similarity between the visual features and all the candidate image features is computed and the one with the largest result is selected as the initial image. We add noise to the initial image for 80 time steps as the initial noise for the diffusion process and subsequently execute the reverse diffusion process for 80 time steps.

Images in the dataset are all photos of objects. In order to make the generated images closer to the images in the dataset, we set the prompts with the prefix "a photo of" to input the diffusion model. Then the prompts will be transformed into glide embedding spaces and generate the text-conditioned images. The prefix "a photo of " corresponds to a tensor of dimension [3,512] in the text embedding space. The largest corresponding dimension of all labels in the dataset is [6,512], so we convert all labels to [6,512] shape by adding padding tokens. Then we train a fully connected network to map the EEG semantic representations into the label embeddings and then concatenate them with the tensors of prefixes to obtain the glide text embeddings. They are fed into the glide model as the condition during the reverse process to perform classifier-free guidance.

4 Experiments

4.1 Dataset and Implementation

The EEG dataset was collected from six subjects when they are watching visual stimuli images.[39] The images used in the experiments include 2000 images from ImageNet dataset[21] (40 classes of objects and 50 images for each class). Images were shown in sequence for 0.5 seconds each. Each class of images lasted for 25 seconds and a black image was presented for 10 seconds between every two classes. The EEG signals contain 128 channels. Following the previous works, we choose the high frequency of gamma band (from 55Hz to 95 Hz), which is related to perceptual processes in visual tasks and achieved the highest performance[32].

During the 500ms of the recording EEG signals, we dropped out the first 40ms and the last 20ms parts. We intercepted consecutive 160ms records from the remained 440ms as the input data for the training and decoding process.

Data from all channels are used in semantic extraction. To extract the visual feature more efficiently, we select 14 channels in the occipital region (PO7, PO3, POz, PO4, PO8, POO9, POO1, POO2, POO10, O1, Oz, O2, Ol1h, Ol2h). These channels are close to the V1 visual cortex, which is associated with low-level visual features. The distribution of electrodes on the EEG cap and the name and location of selected channels are shown in Figure 3.

The dataset is split into three sets by images, 80% for training, 10% for validation and 10% for testing, ensuring that images from different sets do not overlap each other across subjects.

4.2 Experiment Details

Our architecture takes 2 encoders, where the visual encoder has 19K parameters, and the semantic encoder has 523K parameters. The mapping model has 1.6M parameters. The diffusion model has 385M parameters, with total time steps of 160.

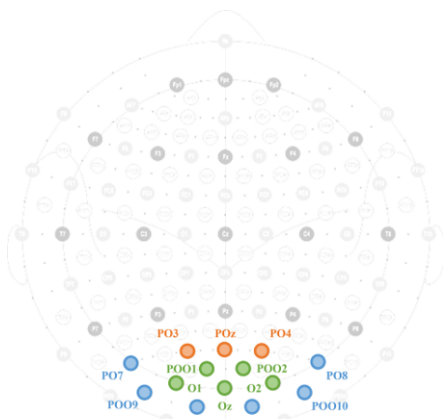

Figure 3. Among the 128 channels, 14 of them are selected and they are highlighted, which are close to the V1 visual cortex.

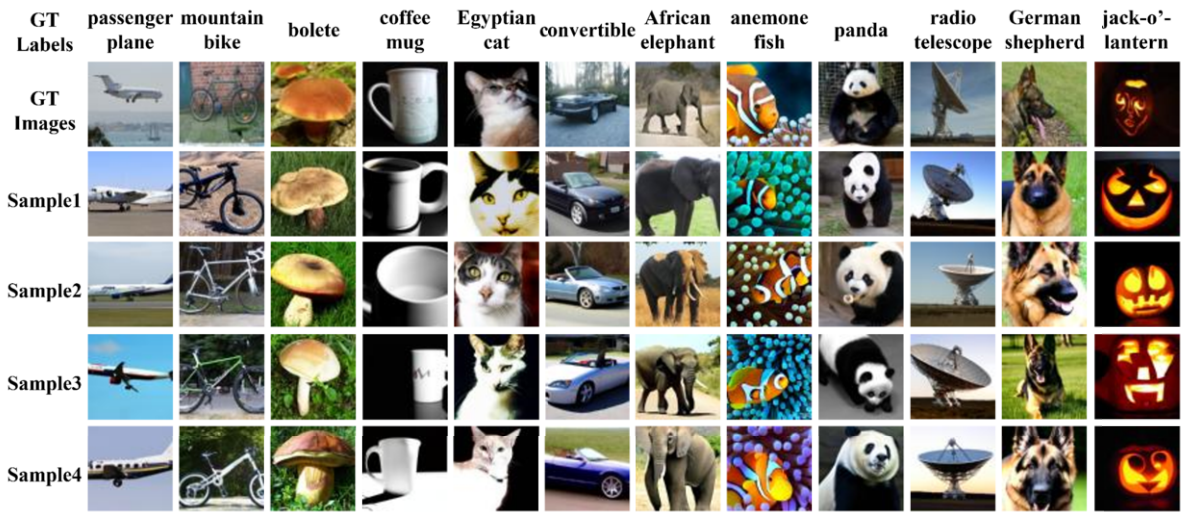

Figure 4. Examples for image reconstruction results. Ground-truth labels are at the first row. Ground-truth images are at the second row. And the remaining rows are different generated image samples.

The recommended hardware environment is the one we used for training and testing, with Intel Xeon Gold 5115 as CPU, NVIDIA GeForce RTX 2080Ti as GPU, and 32GB DDR4 as RAM.

4.3 Image Reconstruction Visualization

Through our framework, the end-to-end reconstruction from EEG signals to images can be accomplished. Some examples for the generated images are shown in Figure 4. It can be seen that, by extracting the semantic features and the visual features from EEG signals, the images reconstructed with our method reproduce the stimulus images in both semantic and visual features. For the same ground-truth image, different samples are generated, showing that images generated by our framework are of high diversity rather than simply repeating the same result.

Certainly, there are some failure results. Besides extracting the wrong features from EEG, limited pre-trained generative models can also lead to failure reconstruction. Figure 5 shows some examples. Since we use a pre-trained diffusion model as a generative model, it sometimes fails to generate right images even with the direct input of ground-truth labels. These failure results demonstrate the shortcomings of pre-trained diffusion models, which may influence the accuracy of image reconstruction.

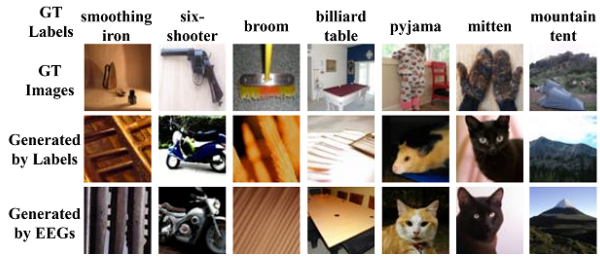

Figure 5. Some failure reconstruction results caused by the pre-trained diffusion model. The four rows from top to bottom are: ground-truth labels from the ImageNet description, ground-truth images of the visual stimuli, images generated by inputting the ground-truth labels directly, and images generated by inputting features predicted from EEG signals.

4.4 Comparison

4.4.1 Semantic Classification

In order to evaluate the effectiveness of representation learning, we calculate the accuracy of the category classification of semantic features. The results for classification compared with other methods are shown in Table 2.

Table 2. Result comparison with other methods on the classification accuracy and the length of the time window used for decoding.

Methods	Accuracy	Time window
EEGChannelNet[32]	60.40%	480ms
DreamDiffusion[4]	45.80%	500ms
Ours	**70.31%**	**160ms**

The results show that our method has a great improvement in classification accuracy compared to other methods while reducing the length of the time window used to decode EEG signals. By introducing hidden states of a large language model, the efficiency of semantic representation learning can be facilitated. And by intercepting continuous EEG signals at different time points, the amount of training data is expanded and it also helps the model to learn the temporal invariance of the signal under the same condition. So our method is more effective and more responsive to time for the real-time EEG classification tasks.

4.4.2 Image Reconstruction

The following three metrics were used to evaluate the quality of reconstructed images.

Inception Score. The Inception Score is a metric used to evaluate the quality and the diversity of images generated by generative models[36]. The higher score indicates the more specific object and the higher diversity of images from different classes[5].

SSIM. The structure similarity index measure (SSIM) is a metric used to quantify the similarity between two images. It compares the structural information of the images by considering local patterns of pixel intensities and their relationships.

CLIP Score. CLIP Score is a metric to measure the semantic similarity between images and text, which can also be used to evaluating

the semantic correlation between images by computing the cosine similarity of the CLIP embeddings of two images[14].

We compute the CLIP scores and the SSIM between generated images and ground-truth images, and compute the average inception score of each category. Table 3 shows the metrics of our method and the quantitative results compared with other methods.

Table 3. Results comparison with other methods. IS represents the inception score, CS represents the CLIP score, and SSIM represents the structure similarity index measure.

Metrics	IS	CS	SSIM
Brain2Image[17]	5.07	/	/
NeuroVision[18]	5.15	/	/
Ours	**7.20**	**0.6230**	**0.1849**

To further demonstrate our quality of image reconstruction, we selected images from the reconstruction results in the same categories as the results presented in other methods. The visualization results are shown in Figure 6. It can be seen that the clarity, realism and diversity of the reconstructed images are greater by our method.

As the results show, our method outperforms other methods on both low-level and high-level metrics.

4.5 Ablation Study

For the EEG representation learning and the image reconstruction task, we further conduct several ablation studies. We evaluate the effectiveness of the modules in our framework by calculating the inception scores, CLIP scores, and structure similarity index measures for different models. Furthermore, as shown in Figure 5, some semantic reconstruction results differed significantly from the original images. To better measure the semantic similarity, we compute the CLIP scores between images generated from EEG signals and ground-truth labels.

Visual features. By selecting the images from the training dataset, we fuse the visual features into the image reconstruction stage, which can help to improve the quality of reconstructed images at low levels. To validate the effectiveness of the visual features, we conduct ablations by inputting the pure noise as initial images into the diffusion model and performing the denoising processing.

Semantic features. The semantic features are used in the image reconstruction stage to generate images that are similar to stimuli images on semantic levels. To validate the effectiveness of semantic features, we also conduct ablation studies. First, we perform the ablation on the semantic features used in the diffusion process. We take the unconditional generation by inputting the initial images. For the ablation of the two modules, the diffusion model will generate completely stochastic images. So we replace the contrastive learning of the semantic encoder network by optimizing it with the MSE loss, instead. The output features are also mapped into the glide embedding spaces by training a fully-connected network and used to generate images as the guidance information.

Figure 6. Qualitative results compared with other methods. The three rows from top to bottom are images reconstructed by Brain2Image, DreamDiffusion and our method, respectively.

We perform ablation studies on visual features, semantic features and both of them, respectively. Table 4 shows the results for ablation studies. IS, CS and SSIM represent the same meaning as Table 3. CLIP scores with images generated by ground-truth labels are denoted by CSwLabels

Table 4. Results for ablation studies. SEM represents incorporating semantic features as guidance information. CL represents optimizing the semantic encoder network through contrastive learning. VIS represents generating images by incorporating visual features.

Model	SEM	CL	VIS	IS	CS	CSwLabels	SSIM
1	✓	✗	✗	6.16	0.5547	0.7352	0.1576
2	✗	✗	✓	**8.26**	0.5540	0.7213	0.1847
3	✓	✓	✗	7.10	0.6098	0.7975	0.1586
4	✓	✓	✓	7.20	**0.6230**	**0.7992**	**0.1849**

It can be seen that the inclusion of semantic features for guidance leads to a lower IS and reduces the diversity of the generated images. Contrastive learning of semantics enables the semantic encoder to learn a more effective representation of the semantic information of the image so that the semantic similarity between the reconstructed image and the original image increases. Introducing visual features by retrieving images in the training set will lead to an increase in image similarity at the low level, and can also increase the similarity metrics with original images at the high level.

5 Conclusion

In this work, we propose a method for EEG representation learning and visual decoding and complete end-to-end image reconstruction tasks from EEG signals with high performance. To alleviate the difficulty faced in EEG visual decoding tasks such as the complex patterns and low signal-to-noise, we explore the possible application of LLMs in these tasks. We incorporate the hidden states in pre-trained LLaMa-2 as extra knowledge to enhance the performance of EEG representation learning and visual decoding. We align the representation of EEG signals with the stimuli images on both semantic features and visual features by contrastive learning, extract information from brain activity efficiently, and finally reconstruct the corresponding stimuli images. To generate images similar to real images both on low levels and high levels, we fuse the semantic features and visual features at the image generation stage by applying a pre-trained diffusion model, selecting images from the training dataset according to the extracted features as initial input, and performing diffusion processes. Our method achieves the state-of-the-art result of EEG semantic classification and image reconstruction on the ImageNet-EEG dataset.

Also, there are some limitations in our work, such as the generalization to other datasets is unclear, the method depends heavily on the pre-trained models, and well-annotated datasets are needed. They need to be further explored in EEG visual decoding tasks.

In short, our work shows the potential applicability of hidden state representations from language models, validates the relationship between large language models and human visual perception tasks, and can be further studied on more complex cognition tasks in the future.

Acknowledgements

This research received funding from the National Natural Science Foundation of China through Grants 62088102, STI2030-Major Projects No.2022ZD0208801 and STI2030-Major Projects No. 2021ZD0113604. And special appreciation to Siyang Wang for helping revise this paper.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] J. Agnese, J. Herrera, H. Tao, and X. Zhu. A survey and taxonomy of adversarial neural networks for text-to-image synthesis, 2019.

[3] A. Ajay, Y. Du, A. Gupta, J. Tenenbaum, T. Jaakkola, and P. Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022.

[4] Y. Bai, X. Wang, Y. pei Cao, Y. Ge, C. Yuan, and Y. Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals, 2023.

[5] S. Barratt and R. Sharma. A note on the inception score, 2018.

[6] Y. Benchetrit, H. Banville, and J.-R. King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*, 2023.

[7] C. Caucheteux, A. Gramfort, and J.-R. King. Deep language algorithms predict semantic comprehension from brain activity. *Scientific reports*, 12(1):16327, 2022.

[8] Z. Chen, J. Qing, T. Xiang, W. L. Yue, and J. H. 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.

[9] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

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

[11] Y. Duan, J. Zhou, Z. Wang, Y.-K. Wang, and C.-T. Lin. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *arXiv preprint arXiv:2309.14030*, 2023.

[12] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023.

[13] P. Ghosh, A. Mazumder, S. Bhattacharyya, and D. Tibarewala. An eeg study on working memory and cognition. In *Proceedings of the 2nd international conference on perception and machine intelligence*, pages 21–26, 2015.

[14] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.

[15] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[16] H. Jing, M. Du, Y. Ma, and N. Zheng. Exploring the relationship between visual information and language semantic concept in the human brain. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 394–406. Springer, 2022.

[17] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817, 2017.

[18] S. Khare, R. N. Choubey, L. Amar, and V. Udutalapalli. Neurovision: perceived image regeneration using cprogan. *Neural Computing and Applications*, 34(8):5979–5991, 2022.

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] W. Klimesch. Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, 29(2-3):169–195, 1999.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[22] M. Lamarre, C. Chen, and F. Deniz. Attention weights accurately predict language representations in the brain. *bioRxiv*, pages 2022–12, 2022.

[23] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15 (5):056013, 2018.

[24] J. Li, D. Li, C. Xiong, and S. 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.

[25] Y. Liu, Y. Ma, W. Zhou, G. Zhu, and N. Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding. *arXiv preprint arXiv:2302.12971*, 2023.

[26] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[27] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

[28] M. Mozafari, L. Reddy, and R. VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

[29] B. Murphy, M. Baroni, and M. Poesio. Eeg responds to conceptual stimuli and corpus semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 619–627, 2009.

[30] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[31] F. Ozcelik and R. VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1): 15666, 2023.

[32] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849, 2021. doi: 10.1109/TPAMI.2020.2995909.

[33] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

[36] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[37] K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M. A. Van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018.

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah. Deep learning human mind for automated visual classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4503–4511, 2017. doi: 10.1109/CVPR.2017.479.

[40] J. Tang, A. LeBel, S. Jain, and A. G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

[41] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[43] X. Xu, Z. Wang, E. Zhang, K. Wang, and H. Shi. Versatile diffusion: Text, images and variations all in one diffusion model, 2024.

[44] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[45] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[46] Z. Zhao, H. Jing, J. Wang, W. Wu, and Y. Ma. Images structure reconstruction from fmri by unsupervised learning based on vae. In *International Conference on Artificial Neural Networks*, pages 137–148. Springer, 2022.