

# CognitionCapturer: Decoding Visual Stimuli from Human EEG Signal with Multimodal Information

Kaifan Zhang<sup>1</sup>, Lihuo He<sup>1\*</sup>, Xin Jiang<sup>1</sup>, Wen Lu<sup>1</sup>, Di Wang<sup>2</sup>, Xinbo Gao<sup>1,3</sup>

<sup>1</sup>School of Electronic Engineering, Xidian University, Xi'an, China

<sup>2</sup>School of Computer Science and Technology, Xidian University, Xi'an, China

<sup>3</sup>Chongqing University of Posts and Telecommunications, Chongqing, China  
zhangkf@stu.xidian.edu.cn, lhhe@mail.xidian.edu.cn

## Abstract

Electroencephalogram (EEG) signals have attracted significant attention from researchers due to their non-invasive nature and high temporal sensitivity in decoding visual stimuli. However, most recent studies have focused solely on the relationship between EEG and image data pairs, neglecting the valuable “beyond-image-modality” information embedded in EEG signals. This results in the loss of critical multimodal information in EEG. To address the limitation, this paper proposes a unified framework that fully leverages multimodal data to represent EEG signals, named CognitionCapturer. Specifically, CognitionCapturer trains modality expert encoders for each modality to extract cross-modal information from the EEG modality. Then, it introduces a diffusion prior to map the EEG embedding space to the CLIP embedding space, followed by using a pretrained generative model, the proposed framework can reconstruct visual stimuli with high semantic and structural fidelity. Notably, the framework does not require any fine-tuning of the generative models and can be extended to incorporate more modalities. Through extensive experiments, we demonstrate that CognitionCapturer outperforms state-of-the-art methods both qualitatively and quantitatively.

Code —

<https://github.com/XiaoZhangYES/CognitionCapturer>

## Introduction

Since its inception, a fundamental challenge in brain decoding is optimally expressing the meaningful information within brain signals. Reconstructing visual stimuli from brain signals is one of the interesting tasks with exciting application prospects. Initially, pioneering work using fMRI data (Kay et al. 2008; Miyawaki et al. 2008; Naselaris et al. 2009) validated the possibility of reconstructing visual stimuli from fMRI data and successfully decoded simple textures and shapes. More recently, with the rapid development of deep learning methods, the use of deep learning models to decode fMRI brain signals has produced significant advancements (Ren et al. 2021; Takagi and Nishimoto 2023; Scotti et al. 2024).

\*Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

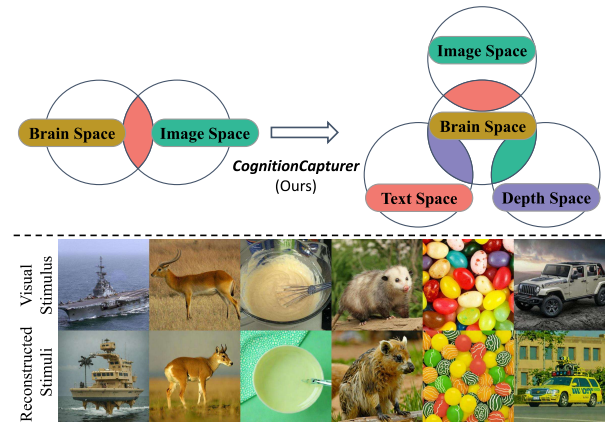


Figure 1: We believe that for image-EEG pairs, relying solely on the mutual information between images and EEG signals can lead to underutilization of EEG information. To address this issue, we utilize multimodal information to capture meaningful information in the EEG signals. The dashed lines in the figure below illustrate some of our successful reconstruction results.

However, brain signals exhibit diverse forms, among which EEG and MEG data offer high temporal resolution and portability, making them particularly suitable for real-time decoding compared to fMRI. This versatility has led to a broader range of downstream applications. Recent works (Benchetrit, Banville, and King 2024; Song et al. 2024; Li et al. 2024) have attempted to align the brain-image modalities using EEG and MEG signals through contrastive learning. These approaches have achieved notable accuracy in decoding related visual stimuli.

However, the internal mechanisms of brain function are diverse and complex. Human perception of visual stimuli is influenced by both the characteristics of the visual stimuli and individual past experiences (Lupyan et al. 2020; Du et al. 2023). Recent works (Benchetrit, Banville, and King 2024; Song et al. 2024; Li et al. 2024) have primarily relied on the image modality as a reference for alignment, enabling the decoding of meaningful visual stimuli. Nonetheless, the objective of contrastive learning may lead to mod-

els that predominantly focus on the shared information between modalities, potentially overlooking the more diverse and complex “beyond-image-modality” information present in the brain signals.

To address this issue, we introduce a novel brain decoding model named CognitionCapturer, as illustrated in Fig. 1. It can be trained jointly with brain signals and multiple modalities, effectively capturing the shared information between brain signals and a broader spectrum of modalities. Specifically, based on the understanding that brain data contains information “beyond-image-modality”, we first extend image data using depth estimation models and image captioning models to construct a Image-Text-Depth multimodal aligned dataset. Then introduce modality expert encoders, which focus on different EEG - single modality data. The embeddings obtained in this stage can be directly used for downstream tasks such as classification and retrieval. Subsequently, in the generation phase, we map the EEG embeddings to the CLIP image space via a diffusion prior and feed EEG embeddings associated with different modalities into a pre-trained image generation model, thus decoding fine-grained visual stimuli.

In contrast to previous methods, CognitionCapturer’s training strategy enables models for different modalities to focus on capturing the relationships between information in EEG signals and modality-specific characteristics. This allows the model to capture fine-grained low-level visual information and abstract high-level semantic information. Furthermore, our proposed approach inherently possesses scalability, enabling the modality expert encoder to be extended infinitely to any modality. Another advantage of the proposed method is that the constructed dataset effectively decouples certain image features, allowing different modality expert encoders to focus on structural and semantic features during training, thereby preventing fine-grained information from being overshadowed by coarse-grained information. The main contributions are as follows:

- We propose CognitionCapturer, a contrastive learning-based model that effectively decodes brain signals from multiple modalities.
- Using an alignment module and a pre-trained image generation model without any fine-tuning, we achieve fine-grained reconstruction of images with performance surpassing that of any single modality.
- Through experiments, we validate the effectiveness and rationality of incorporating more modal information for brain signal decoding, providing new insights for subsequent research in neuroscience.

## Related Work

### Decode Visual Stimuli From Brain Signal

Decoding visual stimuli from fMRI brain signals has been widely studied and yielded successful results (Gu et al. 2024; Takagi and Nishimoto 2023; Scotti et al. 2024; Miyawaki et al. 2008; Kay et al. 2008). However, the difficulty of acquiring fMRI data and its low temporal resolution pose challenges for practical applications. In contrast,

EEG signals offer higher temporal resolution and lower acquisition costs, leading researchers to attempt decoding visual stimuli from EEG. Early EEG decoding work typically relied on supervised learning methods and was limited to a finite set of image categories, overlooking the intrinsic relationship between visual stimuli and brain responses (Li et al. 2020). Recently, (Song et al. 2024; Scotti et al. 2024) successfully constructed an image decoding framework using a contrastive learning approach, achieving zero-shot recognition. (Li et al. 2024) built upon song’s work (Song et al. 2024) by further reconstructing decoded visual information into high-quality images using a diffusion model. However, these works only considered EEG-image modality pairs, neglecting the diversity of brain data. Compared with their approaches, the proposed method successfully leverages multiple modalities of data to decode visual stimuli, resulting in superior performance.

### Contrastive Learning for Brain Decoding

Contrastive learning, as an effective cross-modal learning approach, has achieved significant success in CLIP and Moco (Radford et al. 2021; He et al. 2020). However, its effectiveness is closely related to the quality and scale of the data, and the selection of high-quality samples is crucial for improving model performance (Cherti et al. 2023). Works that use contrastive learning to decode brain signals have also shown promising results. For instance, as a representative work, (Défossez et al. 2023) utilizes a pre-trained encoder to decode speech from MEG signals through contrastive learning, and subsequently, (Benchetrit, Banville, and King 2024) adopts a similar idea to decode images from MEG. A series of similar methods emerged subsequently (Song et al. 2024; Liu et al. 2023; Li et al. 2024). However, during the process of using brain data for contrastive learning, the limited amount of brain signal data may lead the model to focus only on the most discriminative features. After transforming image modality into other modalities, since these modalities are less information-rich compared to image modality, this forces our model to attend to finer-grained features, thereby better representing EEG signals.

## Method

CognitionCapturer aims to address the loss of “beyond-image-modality” information in brain decoding. The method overview is depicted in Fig 2, where EEG-Modality pairs<sup>1</sup> are processed by dedicated modality expert encoders to decouple the effective information from different modalities in the EEG signal. In our experiments, we observed that binding the brain modality with different modalities improves classification and reconstruction performance. Subsequently, through a diffusion prior, the EEG embedding space is mapped to the CLIP space and fed into assembled SDXL-turbo and IP-Adapters to reconstruct visual stimuli.

<sup>1</sup>Specifically, the same EEG signals are divided into three pairs: EEG-Image, EEG-Text, and EEG-Depth. For consistency, we will refer to these collectively as EEG-Modality pairs.

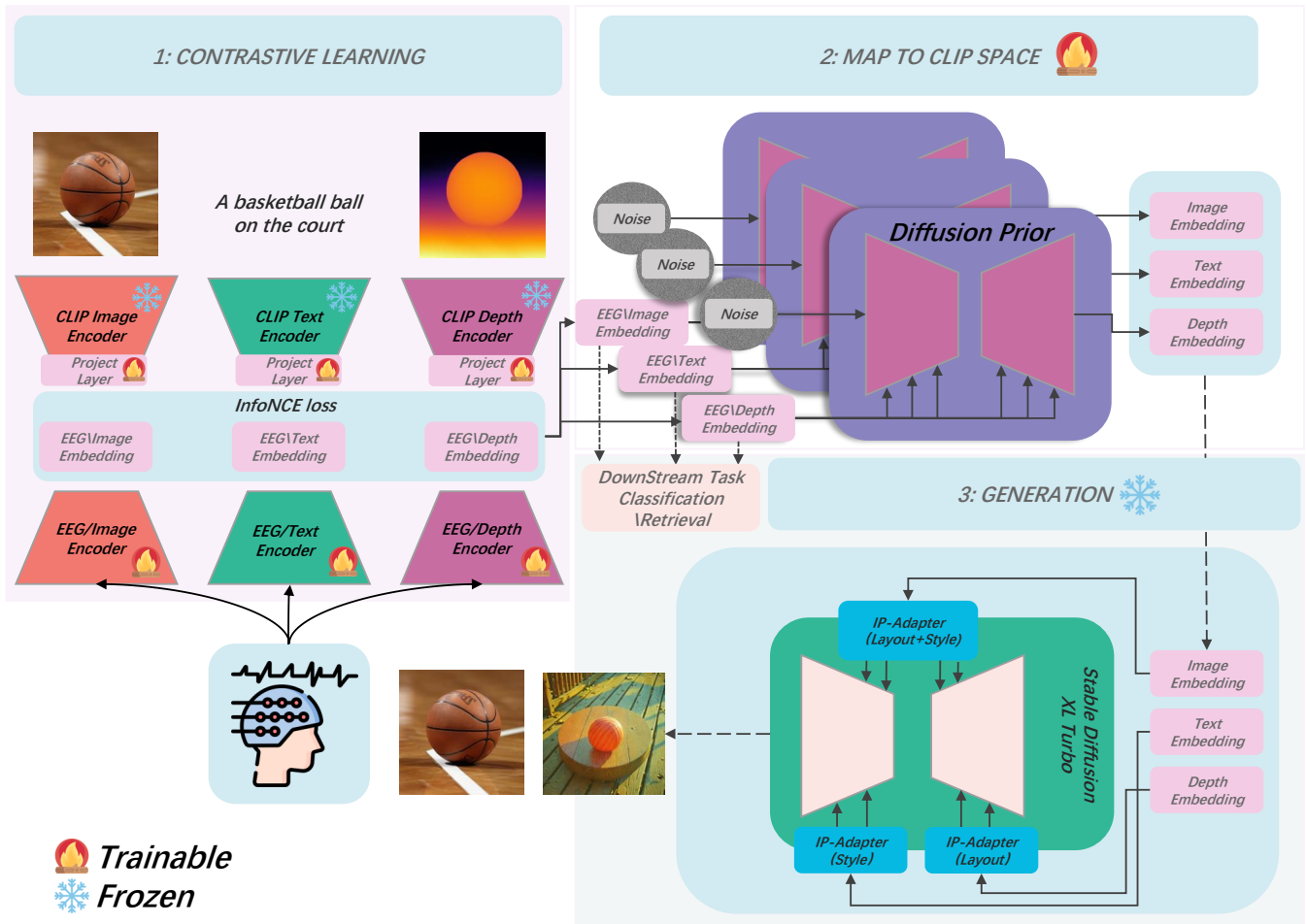


Figure 2: Overall framework of CognitionCapturer. 1: In the contrastive learning stage, different EEG-Modality data pairs are fed into different modality expert encoders for processing. The embeddings obtained from the contrastive learning stage can be used for various downstream tasks. 2: To use pre-trained image generation models, we apply a Diffusion Prior model to map the EEG embeddings into CLIP space while retaining their original information. 3: Using pre-trained SDXL and IP-Adapters with different structures, we integrate the EEG embeddings from different modalities to reconstruct visual stimuli.

### Modality Expert Encoder

CognitionCapturer uses modality pairs  $(E, M)$ , where  $E$  represents the EEG signal and  $M$  represents other modalities. For each modality pair  $(E, M_i)$ , where  $i$  represents the index of different modalities. We construct a dedicated network  $f_i$  and  $g_i$ , named modality expert encoders. This way, each modality pair  $(E, M_i)$  is mapped to the same dimension by its corresponding modality expert encoder for subsequent constraints. In the encoding of the EEG data, raw EEG signals are typically represented as matrices  $C \times T$ , where  $C$  denotes the number of electrode channels and  $T$  denotes the number of time samples. Analysis of EEG signals primarily occurs along these two dimensions.

The proposed EEG encoder, based on a lightweight Transformer and STConv architecture (Li et al. 2024; Vaswani et al. 2017), effectively extracts topological and spatiotemporal information from EEG channels. The network structure is shown in Table 1. Specifically, we first process the

raw EEG data  $e \in R^{C \times T}$  through a layer of Transformer encoder and a linear transformation to organize the topological information, then feed it into a feature extraction module based on STConv to extract spatiotemporal features. Finally, a residual linear layer maps the features output by STConv to the same dimension as the target modality features. Detailed network descriptions are provided in the appendix.

When extracting features for the target modality  $M_i$  paired with EEG data  $E$ , there are many successful pre-trained encoders that can effectively extract img, text, and depth features. Recent work (Zhang et al. 2022) and our experiments indicate that CLIP image embeddings contain depth information. To be compatible with generative models and maintain distribution consistency initially, we used the Open CLIP ViT-H/14 (Radford et al. 2021) as both the visual and text encoder, and added a residual linear layer with the same dimension as the original features to ensure stability during training.

Layer	Input Shape	Output Shape
Transformer Block	$(N, C, T)$	$(N, C, T)$
Linear	$(N, C, T)$	$(N, C, T)$
STConv	$(N, C, T)$	$(N, C_1, T_1)$
Project Layer	$(N, C_1, T_1)$	$(N, D)$

Table 1: Architecture of modality expert encoder

## Align EEG-Modality Pairs by Contrastive Learning

After the modality pairs  $(E, M_i)$  are processed by their respective modality expert encoder  $f_i$  and  $g_i$ , they are encoded into the same dimension, resulting in embedding pairs  $(e_i, m_i)$ . Here,  $(e_i, m_i)$  represents a set consisting of  $n$  samples, i.e.,  $e_i = \{q_1^i, q_2^i, \dots, q_n^i\}$ ,  $m_i = \{k_1^i, k_2^i, \dots, k_n^i\}$ .

Subsequently, for different  $(e_i, m_i)$  embedding pairs, we adopted an improved version of the infNCE loss (van den Oord, Li, and Vinyals 2019) as the loss function:

$$L_{E, M_i} = -\log \frac{L_+}{L_+ + L_-} \quad (1)$$

$$L_+ = \sum_{P(idx)=1} \exp(q_{idx}^T k_i / \tau) \quad (2)$$

$$L_- = \sum_{P(idx)=0} \exp(q_{idx}^T k_i / \tau) \quad (3)$$

$$P(idx) = \begin{cases} 1 & \text{when idx is the same as image label} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

In equation (2) and (3),  $\tau$  is a scalar temperature parameter that controls the smoothness of the softmax distribution. Given that the same image is repeatedly viewed in EEG experiments (Gifford et al. 2022), multiple EEG data may correspond to the same image. This can create a contradictory phenomenon where the same data pairs are both pulled closer and pushed apart by the loss function. To address this, we utilize image index as supervisory information. Specifically, when  $idx$  is the same in multiple EEG data, we choose to pull together all the EEG data and the corresponding modality data, thereby avoiding the contradictory phenomenon. In practice, we employ a symmetric loss  $L_{E, M_i} + L_{M_i, E}$ .

## Map EEG Embedding Into CLIP Image Space

After obtaining the aligned embeddings  $e_i$  for EEG and  $m_i$  for other modalities, due to the existence of the modality gap and differences in distribution spaces (Scotti et al. 2024), directly using the EEG embedding  $e_i$  would make it difficult for pre-trained generative models to identify effective information. Following the works of (Scotti et al. 2024; Li et al. 2024; Ramesh et al. 2022), we use a diffusion prior model to map the EEG embeddings  $e_i$  to the CLIP space, thereby making the EEG embeddings recognizable by pre-trained generative models. In practice, we used the MSE loss to train our diffusion prior from scratch.

$$L_{\text{prior}} = E_{t \sim [1, T], m_i^{(t)} \sim q_t} \left[ \|f_\theta(m_i^{(t)}, t, e_i) - m_i\|^2 \right] \quad (5)$$

In equation (5),  $m_i^{(t)}$  represents the CLIP embedding disturbed after a given diffusion timestep  $t$ , and  $f_\theta$  denotes the diffusion prior network. The specific training details are provided in the Implementation Details section and supplementary material.

## Generate Visual Stimulus With Multi-Modal Associated EEG Embeddings

After the EEG embeddings pass through the diffusion prior, they can be used like the original CLIP embeddings. Specifically, to reconstruct high-fidelity visual stimuli and effectively utilize information from three modalities, we employ Multi IP-Adapters (Ye et al. 2023) and SDXL-turbo (Sauer et al. 2023) to simultaneously leverage embeddings from different modalities. As shown in Fig 2’s generation phase, for the image modality, which contains the richest information, we use a full IP-Adapter to process the image embedding. For text and depth modalities, which focus on semantic and structural information respectively, we use modified versions of IP-Adapter, namely IP-Adapter-Style and IP-Adapter-Layout, to process the text and depth embeddings. This approach enables CognitionCapturer to reconstruct semantic information while preserving underlying visual details.

## Experimental Setup

### Datasets and Preprocessing

We utilized Thing-EEG Dataset for our experiments. The Thing-EEG dataset (Gifford et al. 2022) contains EEG data collected from 10 subjects under an RSVP paradigm. The training set comprises 1654 concepts, each associated with 10 images presented four times, resulting in a total of 66,160 EEG recordings. The test set includes 200 unique concepts, each represented by a single image repeated 80 times, totaling 16,000 EEG recordings. Both the training and test images are presented in a pseudorandom order to minimize habituation effects. Each image is displayed for 100 milliseconds followed by a blank screen for another 100 milliseconds to reduce blink-related and other artifacts. The raw EEG data were filtered between 0.1 Hz and 100 Hz, sampled at 1,000 Hz, and recorded using 63 channels.

For EEG preprocessing, we follow the methodology outlined in (Song et al. 2024; Li et al. 2024). We segment the EEG data into trials ranging from 0 to 1000 ms post-stimulus onset and perform baseline correction using the average value over the 200 ms period preceding the stimulus. All electrodes are retained, and the data are downsampled to 250 Hz. Multivariate noise normalization is applied to the training data, and the EEG repetitions for each image in the test set are averaged to improve the signal-to-noise ratio. Subsequently, to obtain a multimodally aligned dataset, we use BLIP2 (Li et al. 2023) for textual descriptions of the images and DepthAnything (Yang et al. 2024) for depth estimation, resulting in an aligned text and depth dataset.

Method	sub-01	sub-02	sub-03	sub-04	sub-05	sub-06	sub-07	sub-08	sub-09	sub-10	Ave
<i>CognitionCapturer (all)</i>	<i>31.41</i>	<i>31.44</i>	<i>38.19</i>	<i>40.37</i>	<i>24.44</i>	<i>34.84</i>	<i>34.65</i>	<i>48.10</i>	<i>37.42</i>	<i>35.57</i>	<i>35.64</i>
	<i>79.65</i>	<i>77.80</i>	<i>85.65</i>	<i>85.80</i>	<i>66.34</i>	<i>78.75</i>	<i>80.95</i>	<i>88.60</i>	<i>79.36</i>	<i>79.29</i>	<i>80.22</i>
<b>CognitionCapturer (image)</b>	<b>27.22</b>	<b>28.72</b>	<b>37.19</b>	<b>37.69</b>	<b>21.84</b>	<b>31.55</b>	<b>32.80</b>	<b>47.60</b>	<b>33.36</b>	<b>35.07</b>	<b>33.30</b>
	59.50	<b>56.95</b>	<b>66.10</b>	<b>63.20</b>	<b>47.75</b>	<b>58.05</b>	59.55	<b>73.50</b>	<b>57.64</b>	<b>63.57</b>	<b>60.58</b>
CognitionCapturer (text)	17.97	16.16	20.19	26.75	13.12	19.90	22.10	29.40	21.93	21.29	20.88
	35.45	33.85	38.10	46.30	29.90	36.45	37.90	48.60	37.86	40.64	38.51
CognitionCapturer (depth)	23.10	21.85	29.65	34.40	15.75	27.50	30.90	36.90	27.14	26.86	27.41
	57.40	53.25	61.65	65.50	40.25	50.20	54.55	60.20	49.00	49.14	54.11
BraVL (Du et al. 2023)	6.1	4.9	5.6	5.0	4.0	6.0	6.5	8.8	4.3	7.0	5.8
	17.9	14.9	17.4	15.1	13.4	18.2	20.4	23.7	14.0	19.7	17.5
NICE (Song et al. 2024)	12.3	10.4	13.1	16.4	8.0	15.1	15.2	20.0	13.1	14.9	13.8
	36.6	33.9	39.0	47.0	26.9	40.6	42.1	49.9	37.1	41.9	39.5
ATM (Li et al. 2024)	25.6	22.0	25.0	31.4	12.9	21.3	30.5	38.8	24.4	29.1	26.1
	<b>60.4</b>	54.5	62.4	60.9	43.0	51.1	<b>61.5</b>	72.0	51.5	63.5	58.1

Table 2: Overall accuracy (acc±std) of 200-way zero-shot classification: Top-1 and Top-5. The first line in each cell represents the Top-1 accuracy, and the second line represents the Top-5 accuracy. (In the calculation of CognitionCapturer (all)’s classification accuracy, if any modality expert encoder correctly classifies a sample, the sample is considered correctly classified.)

Method (Averaged across subject)	Low-level			High-level			
	PixCorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
CognitionCapturer (all)	0.150	0.347	0.754	0.623	0.669	0.715	0.590
CognitionCapturer (image)	0.132	0.321	0.813	0.671	0.664	0.705	0.599
CognitionCapturer (text)	0.102	0.288	0.727	0.582	0.586	0.598	0.673
CognitionCapturer (depth)	0.104	0.370	0.796	0.638	0.565	0.579	0.686
META-MEG Benchetrit, Banville, and King	0.090	0.341	0.774	0.876	0.703	0.811	0.567
MindEye-fMRI Scotti et al.	0.309	0.323	0.947	0.978	0.938	0.941	0.367

Table 3: Quantitative comparison results on Things-EEG (Gifford et al. 2022) (compared to MEG data on Things-MEG (Hebart et al. 2023) and fMRI data on NSD (Allen et al. 2022)). We report 7 different metrics to quantify the model’s performance in reconstructing images at both low-level and high-level aspects.

## Implementation Details

The proposed method is conducted on a single GeForce RTX 2080 Ti GPU. following the training strategy described in (Song et al. 2024). The model was evaluated on the test set at the end of each epoch, with both training and testing conducted on separate subjects. For the training of the modality expert encoder phase, we used the AdamW optimizer with a learning rate of 0.0003, a batch size of 1024, and trained for 20 epochs. Training for one subject took approximately 30 minutes. For the classification accuracy, we ran 10 experiments and took the average.

Images were resized to 224×224 pixels and normalized before being processed by the modality expert encoder. During the training of the diffusion prior, we used a batch size of 512, trained for 100 epochs, and set the number of inference steps to 50. The guidance scale was set to 7.5. In each batch, 10% of the image embeddings were randomly replaced with noise. The embedding dimension was 1024.

In the generation process, we utilized SDXL-Turbo and IP-Adapter from Hugging Face. We set the inference steps for SDXL-Turbo to 5. When configuring the IP-Adapter, for the image modality, we used the full IP-Adapter with the scale set to 1. For the text and depth modalities, we set the scale of their respective IP-Adapter’s Layout block and Style

block to 0, ensuring a focus on structural and semantic control in the reconstruction results.

## Results and Discussion

### Classification Performance

The classification results of CognitionCapturer are shown in Table 2. We evaluated CognitionCapturer’s ability to decode EEG embeddings based on different baseline modalities. To verify whether CognitionCapturer extracts complementary information across multiple modalities, we combined the top-5 results from three modalities, as shown in the upper bound row of Table 2. The results indicate that compared to previous work (Li et al. 2024; Du et al. 2023), CognitionCapturer achieves state-of-the-art performance on the image modality. With the introduction of the text and depth modalities, the model gains access to more complementary information<sup>2</sup>, leading to a significant increase in the potential amount of effective information. This suggests that complementary information across different modalities is indeed effective.

<sup>2</sup>Note: This does not represent the actual accuracy that can be achieved in practice but rather serves to demonstrate the effectiveness of the complementary information.

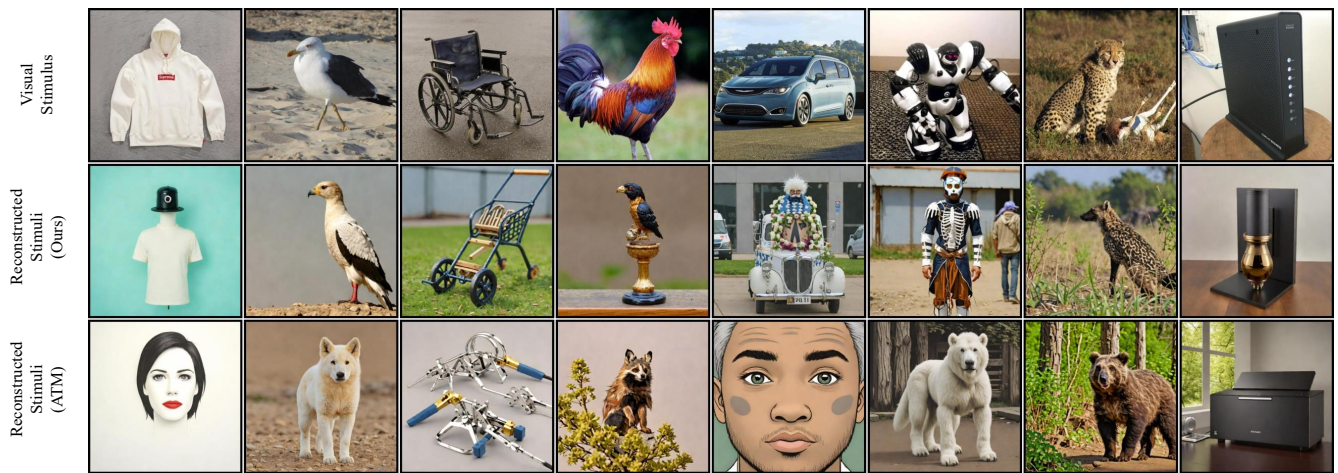


Figure 3: Visual Comparison. Selected reconstruction results from subject-08 show that our reconstructed visual stimuli exhibit finer-grained features.

### Visual Stimuli Reconstruction Performance

Since subject-08 showed the highest classification results in both our model and ATM, we chose subject-08 for the comparison. Some of the visual stimuli reconstructed by CognitionCapturer are shown in Fig 3.

The results show that CognitionCapturer outperforms previous work (Li et al. 2024) in the fine-grained alignment of reconstructed visual stimuli. To further qualitatively analyze the effectiveness of CognitionCapturer’s reconstruction, we recovered visual stimuli for each individual modality and compared them with the complete CognitionCapturer. As shown in Fig. 4, there are differences in reconstruction performance when using single modalities: stimuli recovered only using the text modality tend to be more abstract, while the depth modality can better reconstruct structural information but performs poorly on semantic information. Notably, the image modality, which contains the richest information, sometimes loses certain details in its reconstructions. However, with the assistance of the text and depth modalities, CognitionCapturer recovers more reasonable visual stimuli. For instance, in Fig. 4, when the visual stimulus is a basketball, the image modality misses the “circular” feature, whereas the depth modality retains this information well.

To quantitatively compare our approach with the current state-of-the-art methods, we follow the evaluation metrics outlined in (Benchetrit, Banville, and King 2024) and conduct further quantitative comparisons on the reconstructed images. The results in Table 3 show that CognitionCapturer, when using all modality information, outperforms the use of a single modality in both low-level and high-level metrics. In low-level metrics, CognitionCapturer even matches or surpasses work using higher spatial resolution MEG signals. However, in high-level metrics, there remains a significant gap relative to MEG and fMRI signals, indicating that MEG and fMRI signals are easier to decode for meaningful information than EEG signals.

### How Different Modality Expert Encoders Focus on Brain Regions

In the previous section, we analyzed the reconstruction results of CognitionCapturer. To provide evidence for the feasibility and interpretability of CognitionCapturer, we use Grad-CAM (Selvaraju et al. 2017) to visualize the regions of interest for different modality encoders. To mitigate the influence of individual subjects, we conducted an average analysis of the Grad-CAM results across all subjects’ models. As shown in Fig. 5(A), the raw EEG signal is heavily influenced by frontal lobe responses, whereas our modality expert encoder primarily focuses on the occipital and temporal lobes, areas responsible for processing visual information (DiCarlo and Cox 2007). Notably, compared to the image expert encoder, which mainly attends to the occipital region, the text expert encoder and depth expert encoder attend to broader regions including both the occipital and temporal lobes.

Surprisingly, the depth expert encoder exhibits more significant attention to the right inferior temporal lobe, an area primarily involved in object recognition but less sensitive to object shape, size, and orientation (Epstein and Kanwisher 1998). We believe this is because depth information lacks many lower-level visual features such as color and texture, leaving only shape and depth information. Similar to the phenomenon of sensory compensation (Rauschecker 1995), this forces the model to seek higher-level brain information to ensure effective recognition of similar objects. This demonstrates that our modality-specific expert models reasonably focus on different brain regions, aligning with existing neuroscience theories.

### How Different Brain Area Interact With Visual Stimuli

The analysis in the previous section demonstrated exciting results. To provide additional evidence for the effective interaction between EEG and image information, we further



Figure 4: Reconstruction results of CognitionCapturer on different modality and comparison with prior work.

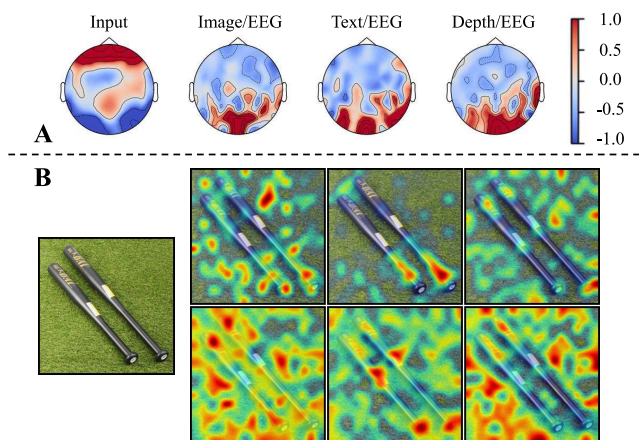


Figure 5: (A) The Grad-CAM results from different modality expert encoders show the activation in the occipital and temporal lobes related to the input EEG signals. (B) The Grad-CAM results from different modality expert encoders on the brain signals corresponding to the example image, visualizing the regions of attention in the images and comparing them with the original CLIP embeddings.

used Grad-CAM to visualize the image regions attended to by the embeddings produced by our modality expert encoders and compared them with the original CLIP embeddings.

As shown in Fig. 5(B), first, in the original CLIP model,

the text embedding focuses more on the object itself, while the image and depth embeddings have broader attention areas. Our modality expert encoders yield EEG embeddings for different modalities that show similar results to those of CLIP. Specifically, the EEG embedding from the text expert encoder focuses more on high-level information in the image, such as the baseball bats. In contrast, the image and depth expert encoders have broader attention over the image. Correspondingly, the brain regions attended to by the image and depth models are also more extensive compared to text. This provides strong evidence for the interpretability of CognitionCapturer.

## Conclusion

In this work, we propose CognitionCapturer to extract multimodal representations from EEG signals and decode visual stimuli from them. Specifically, we introduce multiple modality expert encoders to specialize in aligning EEG embeddings with those of different modalities, enabling the model to capture both semantic and structural information simultaneously. The analysis of brain activity and the interpretability of our model demonstrate that it successfully obtains meaningful representations of brain signals. This provides new insights for subsequent work in brain decoding.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62276203, 62036007, U21A20514, 62072354.

## References

- Allen, E. J.; St-Yves, G.; Wu, Y.; Breedlove, J. L.; Prince, J. S.; Dowdle, L. T.; Nau, M.; Caron, B.; Pestilli, F.; Charest, I.; et al. 2022. A Massive 7T fMRI Dataset to Bridge Cognitive Neuroscience and Artificial Intelligence. *Nature Neuroscience*, 25(1): 116–126.
- Benchetrit, Y.; Banville, H.; and King, J.-R. 2024. Brain Decoding: Toward Real-time Reconstruction of Visual Perception. arXiv:2310.19812.
- Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible Scaling Laws for Contrastive Language-image Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.
- Défossez, A.; Caucheteux, C.; Rapin, J.; Kabeli, O.; and King, J.-R. 2023. Decoding Speech Perception from Non-invasive Brain Recordings. *Nature Machine Intelligence*, 5(10): 1097–1107.
- DiCarlo, J. J.; and Cox, D. D. 2007. Untangling Invariant Object Recognition. *Trends in Cognitive Sciences*, 11(8): 333–341.
- Du, C.; Fu, K.; Li, J.; and He, H. 2023. Decoding Visual Neural Representations by Multimodal Learning of Brain-Visual-Linguistic Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10760–10777.
- Epstein, R.; and Kanwisher, N. 1998. A Cortical Representation of the Local Visual Environment. *Nature*, 392(6676): 598–601.
- Gifford, A. T.; Dwivedi, K.; Roig, G.; and Cichy, R. M. 2022. A Large and Rich EEG Dataset for Modeling Human Visual Object Recognition. *NeuroImage*, 264: 119754.
- Gu, Z.; Jamison, K.; Kuceyeski, A.; and Sabuncu, M. R. 2024. Decoding Natural Image Stimuli from fMRI Data with a Surface-based Convolutional Network. In *Medical Imaging with Deep Learning*, 107–118. PMLR.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.
- Hebart, M. N.; Contier, O.; Teichmann, L.; Rockter, A. H.; Zheng, C. Y.; Kidder, A.; Corriveau, A.; Vaziri-Pashkam, M.; and Baker, C. I. 2023. THINGS-data, a Multimodal Collection of Large-scale Datasets for Investigating Object Representations in Human Brain and Behavior. *Elife*, 12: e82580.
- Kay, K. N.; Naselaris, T.; Prenger, R. J.; and Gallant, J. L. 2008. Identifying Natural Images from Human Brain Activity. *Nature*, 452(7185): 352–355.
- Li, D.; Wei, C.; Li, S.; Zou, J.; and Liu, Q. 2024. Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion. arXiv:2403.07721.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping Language-image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*, 19730–19742. PMLR.
- Li, R.; Johansen, J. S.; Ahmed, H.; Ilyevsky, T. V.; Wilbur, R. B.; Bharadwaj, H. M.; and Siskind, J. M. 2020. The Perils and Pitfalls of Block Design for EEG Classification Experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1): 316–333.
- Liu, Y.; Ma, Y.; Zhou, W.; Zhu, G.; and Zheng, N. 2023. BrainCLIP: Bridging Brain and Visual-Linguistic Representation Via CLIP for Generic Natural Visual Stimulus Decoding. arXiv:2302.12971.
- Lupyan, G.; Rahman, R. A.; Boroditsky, L.; and Clark, A. 2020. Effects of Language on Visual Perception. *Trends in Cognitive Sciences*, 24(11): 930–944.
- Miyawaki, Y.; Uchida, H.; Yamashita, O.; Sato, M.-a.; Morito, Y.; Tanabe, H. C.; Sadato, N.; and Kamitani, Y. 2008. Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders. *Neuron*, 60(5): 915–929.
- Naselaris, T.; Prenger, R. J.; Kay, K. N.; Oliver, M.; and Gallant, J. L. 2009. Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, 63(6): 902–915.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.
- Rauschecker, J. P. 1995. Compensatory Plasticity and Sensory Substitution in the Cerebral Cortex. *Trends in Neurosciences*, 18(1): 36–43.
- Ren, Z.; Li, J.; Xue, X.; Li, X.; Yang, F.; Jiao, Z.; and Gao, X. 2021. Reconstructing Seen Image from Brain Activity by Visually-guided Cognitive Representation and Adversarial Learning. *NeuroImage*, 228: 117602.
- Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2023. Adversarial Diffusion Distillation. arXiv:2311.17042.
- Scotti, P.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Dempster, A.; Verlinde, N.; Yundler, E.; Weisberg, D.; Norman, K.; et al. 2024. Reconstructing the Mind’s Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. *Advances in Neural Information Processing Systems*, 36.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.
- Song, Y.; Liu, B.; Li, X.; Shi, N.; Wang, Y.; and Gao, X. 2024. Decoding Natural Images from EEG for Object Recognition. In *International Conference on Learning Representations*.
- Takagi, Y.; and Nishimoto, S. 2023. 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*, 14453–14463.
- van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30.
- Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth Anything: Unleashing the Power of Large-scale Unlabeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.
- Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv:2308.06721.
- Zhang, R.; Zeng, Z.; Guo, Z.; and Li, Y. 2022. Can Language Understand Depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, 6868–6874.