

NeuroBridge: Bio-Inspired Self-Supervised EEG-to-Image Decoding via Cognitive Priors and Bidirectional Semantic Alignment

Wenjiang Zhang¹, Sifeng Wang¹, Yuwei Su¹, Xinyu Li¹, Chen Zhang¹, Suyu Zhong^{1,2*}

¹School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

²Queen Mary School Hainan, Beijing University of Posts and Telecommunications, Hainan 570100, China

wjzhang@bupt.edu.cn, zhongsuyu@bupt.edu.cn

Abstract

Visual neural decoding seeks to reconstruct or infer perceived visual stimuli from brain activity patterns, providing critical insights into human cognition and enabling transformative applications in brain-computer interfaces and artificial intelligence. Current approaches, however, remain constrained by the scarcity of high-quality stimulus-brain response pairs and the inherent semantic mismatch between neural representations and visual content. Inspired by perceptual variability and co-adaptive strategy of the biological systems, we propose a novel self-supervised architecture, named NeuroBridge, which integrates Cognitive Prior Augmentation (CPA) with Shared Semantic Projector (SSP) to promote effective cross-modality alignment. Specifically, CPA simulates perceptual variability by applying asymmetric, modality-specific transformations to both EEG signals and images, enhancing semantic diversity. Unlike previous approaches, SSP establishes a bidirectional alignment process through a co-adaptive strategy, which mutually aligns features from two modalities into a shared semantic space for effective cross-modal learning. NeuroBridge surpasses previous state-of-the-art methods under both intra-subject and inter-subject settings. In the intra-subject scenario, it achieves the improvements of **12.3%** in top-1 accuracy and **10.2%** in top-5 accuracy, reaching **63.2%** and **89.9%** respectively on a 200-way zero-shot retrieval task. Extensive experiments demonstrate the effectiveness, robustness, and scalability of the proposed framework for neural visual decoding.

Code — <https://github.com/ferooooo/NeuroBridge>

Extended version — <https://arxiv.org/abs/2511.06836>

Introduction

Understanding how the human brain processes visual information is a fundamental challenge in both cognitive neuroscience and artificial intelligence (Posner, Snyder, and Davidson 1980; Kamitani and Tong 2005; Gao et al. 2021). Visual neural decoding, which interprets neural signals evoked by visual stimuli (e.g., images) to infer perceptual and cognitive states, holds significant promise for unraveling the mechanisms of visual processing and advancing brain-inspired AI. While a variety of non-invasive neu-

roimaging techniques, such as functional magnetic resonance imaging (fMRI) (Scotti et al. 2023, 2024; Chen et al. 2024b,c; Dai et al. 2025) and magnetoencephalography (MEG) (Benchetrit, Banville, and King 2024) enable the capture of dynamic brain activity in response to visual inputs, electroencephalography (EEG) stands out due to its excellent temporal resolution, high cost-effectiveness and high portability (Du et al. 2023).

Cross-modal contrastive learning between EEG signals and visual images has emerged as the predominant approach for EEG-based visual neural decoding by establishing semantic alignment between modalities (Du et al. 2023; Song et al. 2024; Wu et al. 2025). The effectiveness of cross-modal learning is predominantly governed by the degree of achieved semantic alignment between modalities. However, an inherent modality gap exists between EEG signals and images, which manifests in two fundamental levels: dynamic variability gap and static intrinsic gap. More specifically, the dynamic gap encompasses intra- and inter-subject EEG variability to identical images, attributable to probabilistic neuro-cognitive factors (e.g., attentional fluctuations, mental states and physiological noise) alongside inherent biological noise (Boynton 2005). In short, this dynamic fluctuation reflects inter- and intra-individual perceptual variability inherent to EEG neural dynamics. The static intrinsic EEG-image gap originates from the irreducible divergence between temporal, low-dimensional, noise-prone EEG signals and spatially structured, high-dimensional, semantically dense images (Piastra et al. 2021). These intrinsic mismatches hinder effective cross-modal alignment. Although EEG datasets have grown in size, they remain limited compared to the large, diverse datasets in computer vision (Deng et al. 2009; Grootswagers et al. 2022), further widening the gap and necessitating stronger architectural priors for robust, generalizable decoding.

To address this challenge, we propose a novel framework, called NeuroBridge, inspired by perceptual variability and co-adaptive strategy of the biological systems, to bridge EEG-based neural and visual modalities. NeuroBridge integrates Cognitive Prior Augmentation (CPA) and Shared Semantic Projector (SSP). CPA enriches the learning process by simulating the human cognitive response during visual perception. As mentioned above, EEG signals are characterized by perceptual-state-dependent properties (Fig. 1).

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

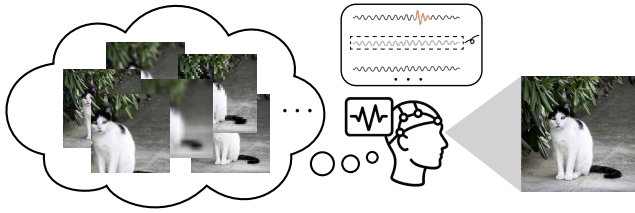


Figure 1: Illustration of human visual cognition and EEG variability. When viewing a visual stimulus, individuals focus on different semantic regions (e.g., parts of a cat), leading to varied EEG responses influenced by attention, mental state, and individual differences.

By leveraging the perceptual variability of the EEG, CPA transforms both EEG signals and images. Besides, considering the static inherent gap between EEG and image modalities (i.e. data structure, feature scale, and semantic richness), CPA also employs an asymmetric augmentation strategy tailored to each modality. This enables more targeted enhancement and alignment, helping to bridge the modality gap more effectively. Considering that conventional EEG-based visual decoding employs unidirectional alignment, which prioritizes computational efficiency yet is prone to semantic incongruence, we propose SSP with the biologically inspired co-adaptation mechanism (Song et al. 2024). Specifically, SSP complements cross-modal semantic alignment by projecting modality-specific features into a unified semantic space. This shared semantic projection not only aligns multimodal features but also enhances cross-modal interaction. While the image encoder is typically pretrained and frozen, the optimization of SSP allows EEG features to be more effectively aligned with the rich semantic representations derived from large-scale visual datasets.

Overall, NeuroBridge provides a principled and efficient approach to bridging the modality gap between EEG and visual data, demonstrating strong decoding performance and good scalability. Our main contributions are summarized as follows:

1. We present **NeuroBridge**, a unified and efficient framework that achieves strong performance and robust generalization across diverse settings by effectively bridging the modality gap between EEG signals and visual representations.
2. We propose **Cognitive Prior Augmentation (CPA)**, a powerful strategy that simulates cognitive variations, enhancing the diversity of both EEG and image features.
3. We design a **Shared Semantic Projector (SSP)** that efficiently maps multimodal features into a common semantic space, enabling more effective cross-modal learning.
4. Extensive experiments demonstrate the strong performance and generalization of our method. On the zero-shot brain-to-image retrieval task, our framework achieves **63.2%** top-1 and **89.9%** top-5 accuracy, surpassing previous state-of-the-art results by large margins (**+12.3%** and **+10.2%**, respectively).

Related Work

Multimodal Contrastive Representation Learning

Multimodal contrastive learning focuses on aligning representations from different modalities, such as vision and language, within a shared embedding space.

Multimodal contrastive representation learning is fundamentally predicated on cross-modal alignment, the process of projecting heterogeneous modalities (e.g., vision, language) into a shared embedding space where semantically congruent instances are jointly optimized through contrastive objectives. In recent years, numerous multimodal contrastive learning approaches have achieved impressive results on large-scale datasets (Schuhmann et al. 2021). Representative works such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), and BLIP (Li et al. 2022) leverage hundreds of millions of image-text pairs to train dual-encoder architectures that align modalities in a shared embedding space. These models exhibit strong zero-shot performance across a wide range of downstream tasks, including image-text retrieval, classification, and visual question answering, demonstrating the effectiveness of large-scale contrastive pretraining for learning generalizable multimodal representations when sufficient training pairs exist. Notably, this success hinges critically on the (1) sufficient quantities of paired samples to cover the semantic space, and (2) rigorous semantic alignment that preserves conceptual relationships across modalities.

However, certain modality pairs still face significant challenges. For instance, audio-image (Guzhov et al. 2022), video-text (Bain et al. 2021), 3D-language (Xue et al. 2023), and EEG-image all lack large-scale, high-quality aligned datasets. These modalities often exhibit greater heterogeneity or weaker natural alignment. EEG-image, in particular, suffers from the complexity of neural signals, which hinders effective cross-modal correspondence. As a result, models trained in these settings often struggle to generalize, especially in zero-shot scenarios. Addressing data scarcity and improving alignment quality remain key open problems in contrastive multimodal representation learning.

Neural Visual Decoding

Neural visual decoding aims to reconstruct or interpret visual content perceived by humans from brain activity, such as fMRI, MEG or EEG signals. EEG has become pivotal in neural decoding for capturing dynamic visual processing, owing to its millisecond temporal resolution. In the realm of EEG-based decoding, one pioneering work (Spampinato et al. 2017) demonstrated significant potential for visual recognition tasks, sparking a wave of subsequent research. However, the block-design paradigm used in their dataset, which grouped images of the same class together, has been criticized for enabling models to exploit low-level statistical cues instead of genuine neural representations (Li et al. 2021; Ahmed et al. 2021). Consequently, the performance of methods based on the dataset using the block-design paradigm remains subject to further scrutiny (Kavassidis et al. 2017; Tirupattur et al. 2018; Bai et al. 2025). In contrast, the Rapid Serial Visual Presentation (RSVP) paradigm miti-

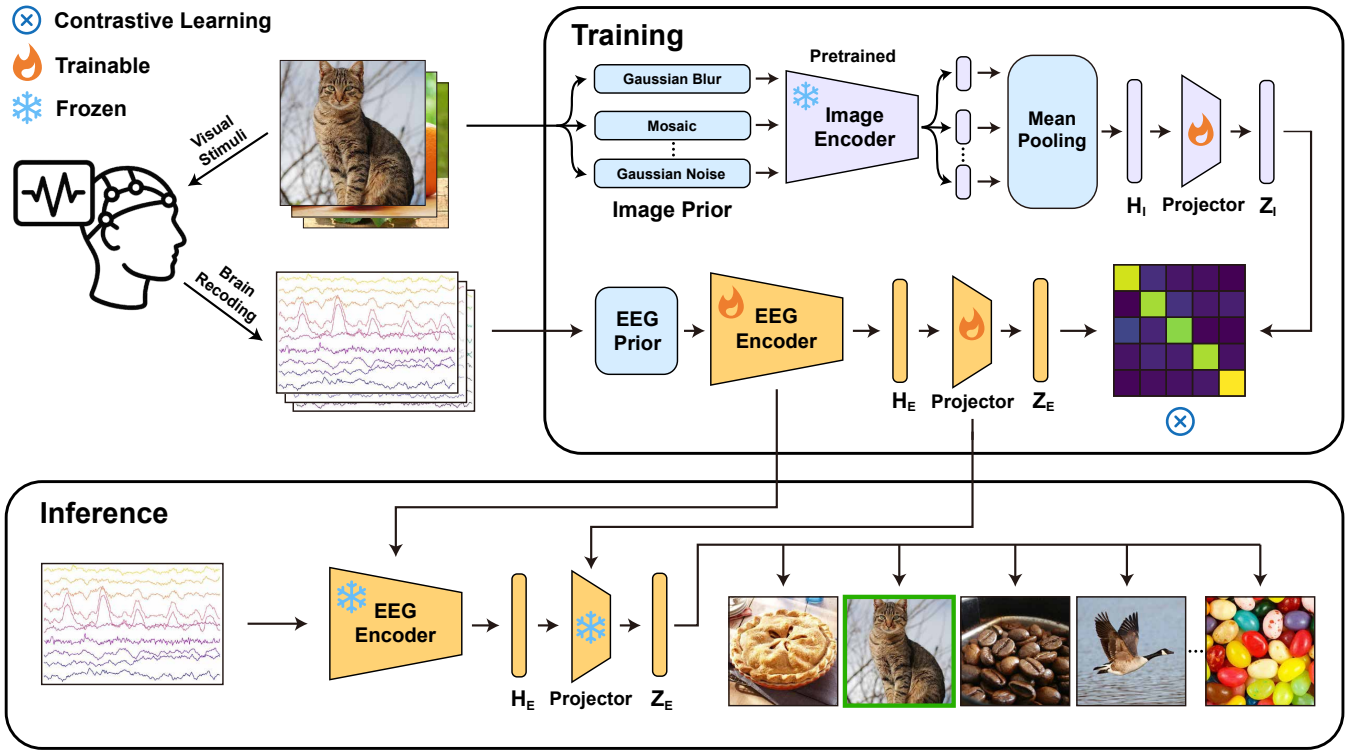


Figure 2: Overall framework of NeuroBridge. During training, cognitively augmented image-EEG pairs are encoded and projected into a shared semantic space using contrastive learning. At inference, EEG signals are decoded by matching their embeddings to visual concepts.

gates these issues by presenting stimuli in a randomized and fast-paced manner, reducing experimental bias (Grootswagers, Robinson, and Carlson 2019). Large-scale datasets based on RSVP, such as Things-EEG (Grootswagers et al. 2022), have since become essential benchmarks for EEG-based neural visual decoding research (Du et al. 2023; Song et al. 2024; Wu et al. 2025).

For the model, although many studies have explored joint modeling of image and EEG modalities, few have explicitly addressed the semantic and representational discrepancies between them. Besides, UBP (Wu et al. 2025) introduced a blur prior to capture early-stage perceptual information in the visual modality, while Neural-MCRL (Li et al. 2024b) proposed intra-modal semantic completion to enhance semantic representation within the EEG modality. Although these methods partially reflect underlying cognitive processes, they predominantly focus on unimodal enhancements such as blur priors for visual refinement and semantic completion for EEG, lacking a comprehensive bimodal framework for synergistic cross-modal modeling.

Method

We propose **NeuroBridge**, a simple yet effective self-supervised EEG-to-image decoding framework with strong extensibility. It consists of two key modules, that is CPA and SSP. The overall architecture is illustrated in Figure 2.

During training, paired EEG-image data are used to ex-

Algorithm 1: Training Algorithm of NeuroBridge

Input: Paired data $D_{\text{train}} = \{(x_I, x_E)\}$; image transformations. $\{t_{I,k}\}_{k=1}^K$; EEG transformation. t_E ; encoders f_I , f_E ; projectors p_I, p_E ; temperature τ ; optimizer \mathcal{O}

Output: Trained f_E, p_I, p_E

- 1: **for each** batch $(X_I, X_E) \in D_{\text{train}}$ **do**
- 2: $X_{I',k} = t_{I,k}(X_I)$
- 3: $X_{E'} = t_E(X_E)$
- 4: $H_{I,k} = f_I(X_{I',k})$
- 5: $H_E = f_E(X_{E'})$
- 6: $H_I = \frac{1}{K} \sum_{k=1}^K H_{I,k}$
- 7: $Z_I = \text{norm}(p_I(H_I)), Z_E = p_E(H_E)$
- 8: $L = \text{ContrastiveLoss}(Z_I, Z_E; \tau)$
- 9: Update $\theta_{f_E}, \theta_{p_I}, \theta_{p_E}$ with $\mathcal{O}(\nabla_{\theta} L)$
- 10: **end for**

tract modality-specific representations. A contrastive learning objective aligns matched pairs while pushing apart mismatched ones. After training, the encoders and projectors enable efficient inference. The full algorithmic flow is outlined in Algorithm 1.

Problem Definition

Our task is zero-shot neural visual decoding, following the setup established in prior work (Song et al. 2024). Let the image data be denoted as $x_I \in \mathbb{R}^{C_I \times H \times W}$ and the corresponding EEG data as $x_E \in \mathbb{R}^{C_E \times T}$, where C_I is the number of image channels, H and W are image height and width, C_E is the number of EEG channels, and T denotes the time window length. Denote the complete sets of image and EEG data as X_I and X_E , respectively. The paired dataset is represented as $D = \{(x_I^{(n)}, x_E^{(n)})\}_{n=1}^N$, which is split into training and test sets D_{train} and D_{test} , with $D_{\text{train}} \cap D_{\text{test}} = \emptyset$.

During training, the model is optimized using D_{train} . In the testing phase, $X_I^{\text{test}} = \{x_I^{(m)}\}_{m=1}^M$ is used as the visual concept pool, and EEG features from $X_E^{\text{test}} = \{x_E^{(m)}\}_{m=1}^M$ are decoded by computing their similarities to the candidate visual features.

Cognitive Prior Augmentation

We posit that many augmentations or transformations of images and EEG signals can serve as cognitive priors, providing auxiliary cues during visual interpretation. To this end, we introduce CPA, a modality-specific augmentation strategy designed to incorporate cognitively relevant priors into contrastive learning. Unlike conventional augmentations primarily designed for classification tasks, CPA aims to simulate human perceptual invariance to transformations in both visual and neural domains.

Given the rich prior knowledge embedded in pretrained image encoders, multiple augmentation strategies are applied to image data. Conversely, due to the absence of such priors in EEG data, an asymmetric design is adopted, where only a single augmentation is applied to EEG signals.

The augmentation process is defined as:

$$X_{I',k} = t_{I,k}(X_I), \quad X_{E'} = t_E(X_E) \quad (1)$$

Here, $t_{I,k}$ and t_E denote image and EEG transformations, respectively, where k indexes different image transformation strategies. CPA offers a flexible augmentation strategy, where specific transformations can be tailored to the requirements of different tasks. The specific augmentation methods are introduced in the following sections.

Image and EEG Feature Extraction

We employ a frozen pretrained image encoder (i.e., CLIP) to extract visual features, while the EEG encoder is trained from scratch. The architecture of both encoders is modular and can be customized.

Due to the large-scale architecture and rich semantic knowledge of the pretrained image encoder, leveraging multiple augmented views is essential to fully exploit its capabilities.

The encoding process is defined as:

$$H_{I,k} = f_I(X_{I',k}), \quad H_E = f_E(X_{E'}) \quad (2)$$

where f_I and f_E denote the image and EEG encoders, and $H_{I,k}$, H_E are the corresponding features.

Semantic Aggregation of Image Representations

Each image transformation produces a distinct feature. Instead of using all features independently, which would significantly increase computational cost and complicate feature alignment, we perform feature fusion to obtain a single representative embedding.

Given that the image encoder is frozen and pretrained, its output features reside in a stable semantic space. Therefore, we average the features obtained from multiple augmented views to generate a compact and semantically enriched representation:

$$H_I = \frac{1}{K} \sum_{k=1}^K H_{I,k} \quad (3)$$

where K denotes the number of applied image augmentation strategies.

Shared Semantic Projector

Pretrained vision-language models, such as CLIP, learn semantic spaces shaped by large-scale image-text associations and are heavily influenced by linguistic co-occurrence. In contrast, EEG signals reflect perceptual and cognitive processes rooted in neural dynamics, which differ fundamentally from language-driven semantics. This inherent modality gap presents significant challenges for direct alignment between EEG features and pretrained visual representations.

To address this issue, we introduce the SSP, which maps both image and EEG features into a unified, trainable semantic space. This modality-agnostic space enables effective alignment by learning semantic correspondence from data, rather than relying on a fixed prior. The projection networks within SSP can be implemented as either linear or nonlinear transformations, offering flexibility to adapt to different tasks or datasets.

The projection process is formulated as:

$$Z_I = p_I(H_I), \quad Z_E = p_E(H_E) \quad (4)$$

where p_I and p_E are the projection networks for the image and EEG modalities, respectively. Z_I and Z_E denote their corresponding embeddings in the shared semantic space.

Modality-Aware Contrastive Learning

We adopt a modality-aware contrastive learning method to align image and EEG features by bringing matched pairs closer and pushing apart mismatched pairs. The loss is defined as:

$$L = \frac{1}{2N} \sum_{j=1}^N \left[-\log \frac{\exp(\text{sim}(z_{I,j}, z_{E,j})/\tau)}{\sum_{z_{E',j}^- \in Z_{E',j}^-} \exp(\text{sim}(z_{I,j}, z_{E',j}^-)/\tau)} - \log \frac{\exp(\text{sim}(z_{E,j}, z_{I,j})/\tau)}{\sum_{z_{I',j}^- \in Z_{I',j}^-} \exp(\text{sim}(z_{E,j}, z_{I',j}^-)/\tau)} \right] \quad (5)$$

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
Intra-Subject: train and test on one subject												
BraVL	Top-1	6.1	4.9	5.6	5.0	4.0	6.0	6.5	8.8	4.3	7.0	5.8
	Top-5	17.9	14.9	17.4	15.1	13.4	18.2	20.4	23.7	14.0	19.7	17.5
NICE	Top-1	13.2	13.5	14.5	20.6	10.1	16.5	17.0	22.9	15.4	17.4	16.1
	Top-5	39.5	40.3	42.7	52.7	31.5	44.0	42.1	56.1	41.6	45.8	43.6
ATM	Top-1	25.6	22.0	25.0	31.4	12.9	21.3	30.5	38.8	34.4	29.1	27.1
	Top-5	60.4	54.5	62.4	60.9	43.0	51.1	61.5	72.0	51.5	63.5	58.1
CognitionCapturer	Top-1	27.2	28.7	37.2	37.7	21.8	31.6	32.8	47.6	33.4	35.1	33.3
	Top-5	59.5	57.0	66.1	63.2	47.8	58.1	59.6	73.5	57.7	63.6	60.6
Neural-MCRL	Top-1	27.5	28.5	37.0	35.0	22.5	31.5	31.5	42.0	30.5	37.5	32.4
	Top-5	64.0	61.5	69.0	66.0	51.5	61.0	62.5	74.5	59.5	71.0	64.1
VE-SDN	Top-1	32.6	34.4	38.7	39.8	29.4	34.5	34.5	49.3	39.0	39.8	37.2
	Top-5	63.7	69.9	73.5	72.0	58.6	68.8	68.3	79.8	69.6	75.3	70.0
UBP	Top-1	41.2	51.2	51.2	51.1	42.2	57.5	49.0	58.6	45.1	61.5	50.9
	Top-5	70.5	80.9	82.0	76.9	72.8	83.5	79.9	85.8	76.2	88.2	79.7
NeuroBridge (Ours)	Top-1	50.0	63.2	61.6	61.4	54.8	69.7	62.7	71.2	64.0	73.6	63.2
	Top-5	77.6	90.6	91.1	90.0	85.0	92.9	88.8	95.1	91.0	97.1	89.9
Inter-Subject: leave one subject out for test												
BraVL	Top-1	2.3	1.5	1.4	1.7	1.5	1.8	2.1	2.2	1.6	2.3	1.8
	Top-5	8.0	6.3	5.9	6.7	5.6	7.2	8.1	7.6	6.4	8.5	7.0
NICE	Top-1	7.6	5.9	6.0	6.3	4.4	5.6	5.6	6.3	5.7	8.4	6.2
	Top-5	22.8	20.5	22.3	20.7	18.3	22.2	19.7	22.0	17.6	28.3	21.4
ATM	Top-1	10.5	7.1	11.9	14.7	7.0	11.1	16.1	15.0	4.9	20.5	11.9
	Top-5	26.8	24.8	33.8	39.4	23.9	35.8	43.5	40.3	22.7	46.5	33.8
UBP	Top-1	11.5	15.5	9.8	13.0	8.8	11.7	10.2	12.2	15.5	16.0	12.4
	Top-5	29.7	40.0	27.0	32.3	33.8	31.0	23.8	32.2	40.5	43.5	33.4
Neural-MCRL	Top-1	13.0	12.0	14.5	12.5	11.5	13.5	14.0	18.5	13.5	17.0	14.0
	Top-5	31.5	30.5	35.5	35.5	29.0	35.5	36.0	38.5	32.5	39.0	34.3
NeuroBridge (Ours)	Top-1	23.2	21.2	13.2	17.0	14.5	25.0	15.3	20.1	13.7	27.2	19.0
	Top-5	52.4	49.3	36.5	45.3	37.7	55.0	45.1	44.9	36.5	56.3	45.9

Table 1: Overall accuracy (%) of 200-way zero-shot retrieval on THINGS-EEG: Top-1 and Top-5.

Here, $z_{I,j} \in Z_I$ and $z_{E,j} \in Z_E$ are paired image and EEG embeddings; $Z_{E,j}^-$ and $Z_{I,j}^-$ are their negative sets; N is the number of pairs; and τ is the temperature. The similarity function $sim(\cdot, \cdot)$ uses cosine similarity, with only image features ℓ_2 -normalized.

Unlike symmetric contrastive learning, our method normalize image features to the unit hypersphere while allowing EEG feature magnitudes to vary. This leverages feature direction for semantic alignment and magnitude for learnable confidence, preserving the pretrained image structure while improving EEG robustness and training stability.

Experiments and Results

Datasets and Preprocessing

We conducted our experiments on the THINGS-EEG dataset (Grootswagers et al. 2022), which contains recordings from 10 subjects using the RSVP paradigm. The training set comprises 1,654 concepts, each associated with 10

images, with each image repeated four times per subject. The test set consists of 200 concepts, each represented by one image, with each image repeated 80 times per subject.

For data preprocessing, we follow the methodology described in previous work (Song et al. 2024; Li et al. 2024a; Wu et al. 2025). EEG data are segmented into 0-1,000 ms trials following stimulus onset, with baseline correction using the average signal from the 200 ms preceding the stimulus. The data are then downsampled to 250 Hz. Multivariate noise normalization (MVNN) is applied for normalization. Repetitions are averaged to enhance SNR, resulting in a total of 16,540 training samples and 200 test samples per subject.

See the extended version (Zhang et al. 2025b) for additional evaluation on THINGS-MEG (Hebart et al. 2023).

Implementation Details

Our method is implemented using PyTorch and trained on two NVIDIA GeForce RTX 3090 GPUs. We employ a batch size of 1,024 and train the model for 50 epochs. The learning

rate is set to $1e-4$. Gradient updates are performed using the AdamW optimizer with a weight decay of $1e-4$. The temperature parameter τ is set to 0.07. All reported results represent the average over five independent training runs.

Encoder For the image encoder, we adopt CLIP (Radford et al. 2021) with weights from OpenCLIP (Cherti et al. 2023), utilizing various backbones including RN50, RN101, ViT-B-16, ViT-B-32, ViT-L-14, ViT-H-14, ViT-g-14, and ViT-bigG-14, with RN50 as default. For the brain encoder, we implement EEGNet (Lawhern et al. 2018), TSCov (Song et al. 2024), ATM (Li et al. 2024a), and EEGProject (Wu et al. 2025), where EEGProject is the default.

CPA For image augmentation, we employ gaussian blur, gaussian noise, low resolution, mosaic, color jitter, grayscale, and random crop. EEG augmentation includes channel dropout, noise addition, smoothing, and temporal shifting. By default, gaussian blur, gaussian noise, low resolution and mosaic are used for image prior, while smoothing is used for EEG prior.

SSP We experimented with both linear projection and MLPs across various feature dimensions.

Comparison with Baselines

We compared our method with several recent works: BraVL (Du et al. 2023), NICE (Song et al. 2024), ATM (Li et al. 2024a), CognitionCapturer (Zhang et al. 2025a), Neural-MCRL (Li et al. 2024b), VE-SDN (Chen et al. 2024a), and UBP (Wu et al. 2025). Table 1 present the quantitative results on the THINGS-EEG dataset. Our method surpasses the state-of-the-art in both intra-subject and inter-subject evaluations. Specifically, our approach achieves a Top-1 accuracy of 63.2% and a Top-5 accuracy of 89.9% in the zero-shot brain-image retrieval task.

Effect of Image and EEG Transformations

We evaluated the effectiveness of various image and EEG transformations used in CPA. Figure 3 illustrates the Top-1 accuracy results of different transformation combinations.

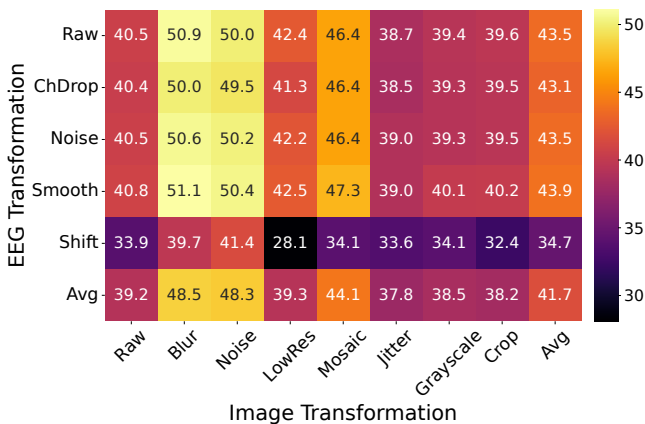


Figure 3: Top-1 accuracy(%) for different transformation combinations.

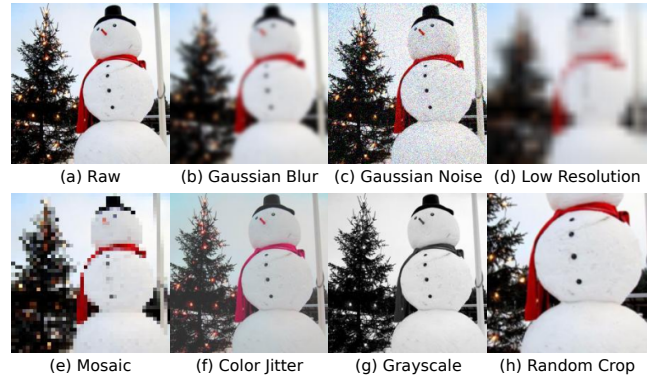


Figure 4: Visual examples of applied image transformations.

Image Transformation Figure 4 displays the visual examples of different image transformation techniques. Gaussian blur, gaussian noise, low resolution, and mosaic contribute positively to retrieval performance. In contrast, color jitter, grayscale, and random cropping lead to a performance decline. We hypothesize that the former group preserves high-level semantic content while mitigating low-level pixel variations. The negative effects of jitter and grayscale indicate that human perception is sensitive to color information, aligning with findings from current neuroscience research (Teichmann et al. 2020). Cropping may remove crucial semantic regions, thereby disrupting the image’s structural or semantic integrity.

EEG Transformation EEG transformations provide limited improvement compared to image transformations. Among them, only smoothing consistently enhances performance. Temporal shifting may disturb the temporal dynamics of EEG signals, impairing semantic representation. The beneficial effect of smoothing likely stems from the inherently low SNR of EEG data, as smoothing can mitigate noise interference.

Effect of Multi-Transformation Feature Fusion

We assessed the impact of fusing image features obtained through different transformations. The number of fused transformations ranges from one to seven, following the order shown in Figure 4. As indicated in Table 2, accuracy improves as the number of fused transformations increases from one to four, then slightly declines from five to seven. This observation is consistent with the earlier results on individual transformation effects. Overall, these findings suggest that fusing diverse image features is beneficial up to a certain point.

# Transforms	1	2	3	4	5	6	7
Top-1 (%)	50.9	54.2	58.4	62.1	58.5	58.1	57.0
Top-5 (%)	81.4	84.5	88.1	89.8	88.1	87.1	86.0

Table 2: Top-1 and Top-5 accuracy (%) under different numbers of fused transformations.

Effect of Different Projector Designs

We evaluated the performance of both linear projection and MLP-based projectors with varying output feature dimensions. As shown in Figure 5, the 512-dimensional linear projection yields the best result. We conjecture this reflects a balance between the model’s expressive capacity and the preservation of information from the pretrained image encoder. Higher-dimensional projectors may lead to overfitting or redundancy, while lower-dimensional ones may be too constrained to capture shared semantics. A larger EEG-image dataset could potentially alleviate this trade-off.

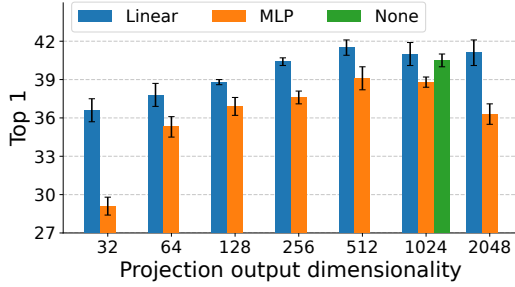


Figure 5: Top-1 accuracy (%) across different projector types and feature dimensions.

Generalization across Encoder Variants

To demonstrate the generalizability of our method across different encoder architectures, we implemented several image and EEG encoders. Figure 6 presents the accuracy improvements achieved with our framework across different encoder combinations.

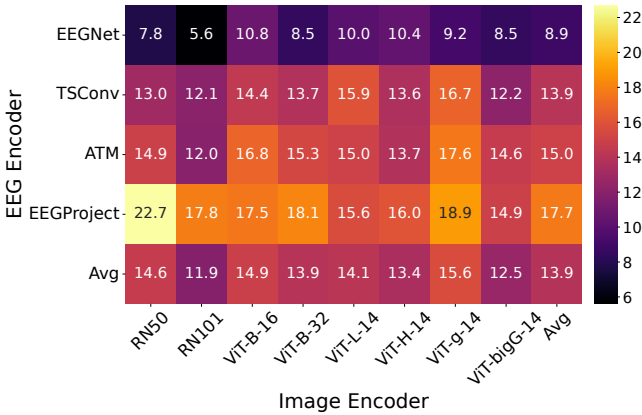


Figure 6: Top-1 accuracy improvement(%) for various encoders architectures.

Ablation Study on Core Framework Components

To assess the contributions of each component, we performed ablation studies by removing the CPA (including image and EEG priors) and SSP modules. As shown in Table 3, all components positively impact overall performance. The image prior has the greatest effect, likely due to the rich visual information and benefits of pretrained image models.

CPA		SSP	Top-1	Top-5
Image Prior	EEG Prior			
✗	✗	✗	40.5	72.2
✓	✗	✗	60.0	89.1
✗	✓	✗	40.8	72.7
✗	✗	✓	41.5	73.5
✗	✓	✓	41.8	73.6
✓	✗	✓	62.1	89.8
✓	✓	✗	60.8	89.8
✓	✓	✓	63.2	89.9

Table 3: Ablation study on the core components of NeuroBridge.

Semantic Alignment Analysis and Visualization

To evaluate the semantic alignment between EEG and visual representations, we computed similarity scores for all 200 test concepts from Subject 8. As illustrated in Figure 7 (a), the similarity matrix reveals strong alignment across modalities. Furthermore, the retrieval examples in Figure 7 (b) demonstrate semantic consistency; for example, when the query image depicts an animal, the top-5 retrieved images are also animals, and a similar trend is observed for food-related images.

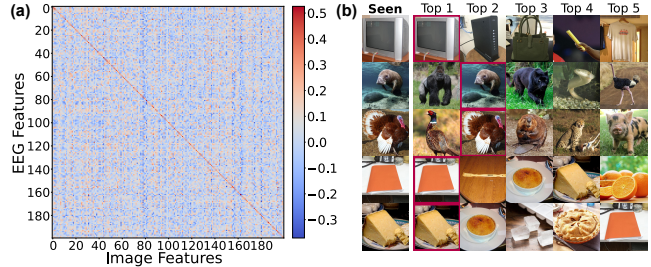


Figure 7: Semantic analysis on subject 8. (a) Similarity matrix. (b) Retrieved samples.

Conclusion

We present NeuroBridge, a unified framework that effectively bridges EEG signals and visual representations for zero-shot neural decoding. By integrating Cognitive Prior Augmentation and Shared Semantic Projector, the method simulates cognitive variability and aligns cross-modal features in a shared semantic space. Experiments demonstrate significant performance gains over existing approaches, confirming the efficacy and generalizability of our framework in both intra-subject and inter-subject settings.

Limitations While NeuroBridge shows strong performance, it has limitations. Manually designed augmentation may insufficiently capture cognitive variability, and reliance on pretrained visual encoders can introduce vision-language biases. These issues suggest future work directions, including adaptive augmentation, data-driven alignment, and scaling up data to improve generalization.

References

- Ahmed, H.; Wilbur, R. B.; Bharadwaj, H. M.; and Siskind, J. M. 2021. Object Classification From Randomized EEG Trials. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3845–3854.
- Bai, Y.; Wang, X.; Cao, Y.-P.; Ge, Y.; Yuan, C.; and Shan, Y. 2025. DreamDiffusion: High-Quality EEG-to-Image Generation with Temporal Masked Signal Modeling and CLIP Alignment. In *Computer Vision – ECCV 2024*, 472–488. Cham: Springer Nature Switzerland. ISBN 978-3-031-72751-1.
- Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1728–1738.
- Benchetrit, Y.; Banville, H.; and King, J.-R. 2024. Brain decoding: toward real-time reconstruction of visual perception. In *The Twelfth International Conference on Learning Representations*.
- Boynton, G. M. 2005. Attention and visual perception. *Current opinion in neurobiology*, 15(4): 465–469.
- Chen, H.; He, L.; Liu, Y.; and Yang, L. 2024a. Visual Neural Decoding via Improved Visual-EEG Semantic Consistency. *arXiv preprint arXiv:2408.06788*.
- Chen, J.; Qi, Y.; Wang, Y.; and Pan, G. 2024b. Bridging the semantic latent space between brain and machine: Similarity is all you need. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 11302–11310.
- Chen, J.; Qi, Y.; Wang, Y.; and Pan, G. 2024c. Mind artist: Creating artistic snapshots with human thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27207–27217.
- 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.
- Dai, Y.; Yao, Z.; Song, C.; Zheng, Q.; Mai, W.; Peng, K.; Lu, S.; Ouyang, W.; Yang, J.; and Wu, J. 2025. MindAligner: Explicit Brain Functional Alignment for Cross-Subject Visual Decoding from Limited fMRI Data. *arXiv preprint arXiv:2502.05034*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- 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.
- Gao, X.; Wang, Y.; Chen, X.; and Gao, S. 2021. Interface, interaction, and intelligence in generalized brain–computer interfaces. *Trends in Cognitive Sciences*, 25(8): 671–684.
- Grootswagers, T.; Robinson, A. K.; and Carlson, T. A. 2019. The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188: 668–679.
- Grootswagers, T.; Zhou, I.; Robinson, A. K.; Hebart, M. N.; and Carlson, T. A. 2022. Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1): 3.
- Guzhov, A.; Raue, F.; Hees, J.; and Dengel, A. 2022. Audio-clip: Extending Clip to Image, Text and Audio. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 976–980.
- 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.
- Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 4904–4916. PMLR.
- Kamitani, Y.; and Tong, F. 2005. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5): 679–685.
- Kavassidis, I.; Palazzo, S.; Spampinato, C.; Giordano, D.; and Shah, M. 2017. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, 1809–1817.
- Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5).
- Li, D.; Wei, C.; Li, S.; Zou, J.; and Liu, Q. 2024a. Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 102822–102864. Curran Associates, Inc.
- Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.
- Li, R.; Johansen, J. S.; Ahmed, H.; Ilyevsky, T. V.; Wilbur, R. B.; Bharadwaj, H. M.; and Siskind, J. M. 2021. The Perils and Pitfalls of Block Design for EEG Classification Experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1): 316–333.
- Li, Y.; Kang, Z.; Gong, S.; Dong, W.; Zeng, W.; Yan, H.; Siok, W. T.; and Wang, N. 2024b. Neural-MCRL: Neural Multimodal Contrastive Representation Learning for EEG-based Visual Decoding. *arXiv preprint arXiv:2412.17337*.

- Piastra, M. C.; Nüßing, A.; Vorwerk, J.; Clerc, M.; Engwer, C.; and Wolters, C. H. 2021. A comprehensive study on electroencephalography and magnetoencephalography sensitivity to cortical and subcortical sources. *Human Brain Mapping*, 42(4): 978–992.
- Posner, M. I.; Snyder, C. R.; and Davidson, B. J. 1980. Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109(2): 160–174.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.
- Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.
- Scotti, P.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Dempster, A.; Verlinde, N.; Yundler, E.; Weisberg, D.; Norman, K.; et al. 2023. Reconstructing the mind’s eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36: 24705–24728.
- Scotti, P.; Tripathy, M.; Villanueva, C.; Kneeland, R.; Chen, T.; Narang, A.; Santhirasegaran, C.; Xu, J.; Naselaris, T.; Norman, K.; and Abraham, T. 2024. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. *Proceedings of Machine Learning Research*, 235: 44038–44059. Publisher Copyright: Copyright 2024 by the author(s); 41st International Conference on Machine Learning, ICML 2024 ; Conference date: 21-07-2024 Through 27-07-2024.
- 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*.
- Spampinato, C.; Palazzo, S.; Kavasidis, I.; Giordano, D.; Souly, N.; and Shah, M. 2017. Deep Learning Human Mind for Automated Visual Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Teichmann, L.; Quek, G. L.; Robinson, A. K.; Grootswagers, T.; Carlson, T. A.; and Rich, A. N. 2020. The Influence of Object-Color Knowledge on Emerging Object Representations in the Brain. *Journal of Neuroscience*, 40(35): 6779–6789.
- Tirupattur, P.; Rawat, Y. S.; Spampinato, C.; and Shah, M. 2018. ThoughtViz: Visualizing Human Thoughts Using Generative Adversarial Network. In *Proceedings of the 26th ACM International Conference on Multimedia, MM ’18*, 950–958. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356657.
- Wu, H.; Li, Q.; Zhang, C.; He, Z.; and Ying, X. 2025. Bridging the Vision-Brain Gap with an Uncertainty-Aware Blur Prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2246–2257.
- Xue, L.; Gao, M.; Xing, C.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J. C.; and Savarese, S. 2023. ULIP: Learning a Unified Representation of Language, Images, and Point Clouds for 3D Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1179–1189.
- Zhang, K.; He, L.; Jiang, X.; Lu, W.; Wang, D.; and Gao, X. 2025a. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14486–14493.
- Zhang, W.; Wang, S.; Su, Y.; Li, X.; Zhang, C.; and Zhong, S. 2025b. NeuroBridge: Bio-Inspired Self-Supervised EEG-to-Image Decoding via Cognitive Priors and Bidirectional Semantic Alignment. *arXiv:2511.06836*.