What Do EEG Representations Reveal? Multiview Neural Representations to Demystify the EEG-Visual Alignment

Salini Yadav[1]* **Taveena Lotey**[1] **Pravendra Singh**[1] **Partha Pratim Roy**[2]
[1]Department of CSE, IIT Roorkee, India [2]Department of CSE, IIT (ISM) Dhanbad, India

Abstract

Zero-shot visual decoding from electroencephalography (EEG) aims to infer visual semantics from non-invasive neural recordings, but remains challenging due to the low signal-to-noise ratio, non-stationarity, and limited spatial resolution of EEG. Existing EEG–vision alignment methods typically model the temporal, spectral, and spatial structure of EEG separately or only partially, leaving their complementary roles in visual perception underexploited. We introduce a unified multiview EEG representation learning framework for aligning brain responses with visual semantic embeddings. Our method builds an EEG encoder that jointly models three complementary views: input-conditioned state-space temporal dynamics, learnable wavelet-based spectral decomposition for sample-adaptive frequency modeling, and attention-modulated graph learning for structured electrode interactions. The resulting multiview EEG embeddings are fused and aligned with pretrained visual representations in a shared semantic space using contrastive learning with EEG-specific regularization, enabling 200-way zero-shot visual classification. Experiments on THINGS-EEG benchmark show that our method achieves state-of-the-art performance, with **54.8% Top-1 and 85.6% Top-5 accuracy** in the within-subject setting and **15.3% Top-1 and 45.4% Top-5 accuracy** in the cross-subject setting. We further present the systematic evaluation of cross-session zero-shot EEG–image semantic decoding/retrieval, achieving **40.8% Top-1 and 78.0% Top-5 accuracy**. These results suggest that explicitly modeling multiview neural structure improves both semantic alignment and generalization in EEG-based visual decoding.

1 Introduction

Humans recognize objects in complex visual scenes rapidly and reliably, transforming sensory input into semantic representations that support perception, decision-making, and interaction [1–3]. Decoding such visual representations from brain activity is a central goal of cognitive neuroscience and brain–computer interface (BCI) research [4]. While fMRI has enabled high-resolution visual reconstruction and categorization [5, 6], its slow hemodynamic response limits temporal precision. MEG provides richer temporal information [7, 8], but remains costly and difficult to deploy. Electroencephalography (EEG), in contrast, is inexpensive, portable, and provides millisecond-level temporal resolution, making it attractive for scalable visual decoding [9]. However, EEG signals are noisy, non-stationary, and spatially coarse, making semantic visual decoding substantially more challenging [10, 11].

Recent large-scale datasets such as THINGS-EEG have expanded EEG-based visual decoding to thousands of natural images and diverse object concepts [12, 13]. Meanwhile, self-supervised and multimodal vision models, especially CLIP, provide semantic embedding spaces that support open-

*Corresponding author. Email: salini_y1@cs.iitr.ac.in

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

vocabulary and zero-shot recognition [14–16]. These developments have motivated EEG–vision alignment methods that map neural signals into pretrained visual-semantic spaces for zero-shot classification or image reconstruction [17]. Nevertheless, their performance is often limited by the quality and stability of the EEG representations being aligned.

Recent EEG–visual decoding studies have explored mapping EEG-derived representations to visual features and semantic spaces [10, 18–24]. However, effective alignment depends on how well these representations capture the heterogeneous structure of visual EEG responses, which comprise temporal dynamics reflecting evolving perceptual processing, spectral patterns associated with frequency-specific neural activity, and spatial dependencies induced by electrode topology and cortical organization. Previous work has modeled these dimensions separately or through partial interactions [25–29], but a unified framework that jointly captures long-range temporal dependencies, adaptive frequency structure, and input-dependent spatial interactions remains underexplored. This fragmentation can yield unstable EEG embeddings that align weakly with visual semantics, particularly under cross-subject and cross-session evaluation.

In this paper, we propose a unified multiview EEG representation learning framework for zero-shot visual decoding. Rather than modifying the pretrained vision encoder, we focus on learning a stronger EEG encoder that produces stable and semantically aligned neural embeddings. Our framework models three complementary views of EEG: an input-conditioned state-space temporal module for long-range and non-stationary dynamics, a learnable wavelet-based spectral module for sample-adaptive frequency decomposition, and an attention-modulated graph module for structured spatial interactions among electrodes. The resulting temporal, spectral, and spatial features are fused into compact EEG embeddings and aligned with pretrained visual representations using contrastive learning and EEG-specific regularization.

We evaluate our method on THINGS-EEG benchmark under within-subject, cross-subject, and cross-session settings. In addition to standard within-subject and cross-subject evaluation, we present the systematic evaluation of cross-session zero-shot EEG–image semantic decoding/retrieval to assess whether semantic alignment remains stable across recording sessions. Across these settings, our approach improves 200-way zero-shot visual classification, indicating that explicit multiview modeling provides a more reliable basis for EEG–visual semantic alignment.

Our main contributions are summarized as follows:

- We introduce a unified multiview EEG representation learning framework for zero-shot visual decoding, jointly modeling temporal dynamics, adaptive spectral structure, and spatial electrode interactions.
- We design a complementary EEG encoder with an input-conditioned state-space temporal module, a learnable wavelet-based spectral module, and an attention-modulated graph spatial module.
- We incorporate EEG-specific regularization to improve representation stability through trial-level consistency and class-level semantic alignment.
- We achieve state-of-the-art performance on THINGS-EEG for 200-way zero-shot visual classification and present a systematic cross-session EEG–image decoding evaluation.

2 Related work

Recent progress in visual neural decoding has been accelerated by large-scale datasets such as THINGS-EEG [13] and by pretrained multimodal and generative models. Existing work mainly focuses on zero-shot semantic decoding, generative reconstruction, and reducing the gap between visual stimuli, neural responses, and semantic representations.

2.1 Zero-shot visual decoding frameworks

Zero-shot visual decoding aims to infer visual semantics beyond categories observed during training. Recent methods commonly align neural signals with pretrained semantic spaces, especially CLIP embeddings. NICE-EEG [23] introduced hierarchical contrastive learning for EEG–CLIP alignment on THINGS-EEG, while NICE-LLM [30] used LLM-generated captions to strengthen semantic supervision. BraVL [22] modeled brain, vision, and language jointly, showing the value of linguistic

priors for semantic generalization. UMind [31] further unified retrieval, classification, and reconstruction using coarse- and fine-grained textual descriptions. Although these methods improve semantic alignment, they often rely on relatively simple EEG encoders and leave the temporal, spectral, and spatial structure of EEG representations insufficiently explored.

In parallel, EEG representation-learning models have increasingly explored structured spatiotemporal modeling. S-JEPA employs spatial block masking to learn structured EEG representations [32], while EEGPT explores hierarchical spatial–temporal modeling [33] and electrode-wise modeling [34]. EEG-DINO further employs hierarchical self-distillation with multi-view EEG representations, channel-aware sampling, and decoupled spatial–temporal positional encoding [35], and Brain-JEPA extends structured spatiotemporal representation learning to fMRI through brain-gradient positioning and spatiotemporal masking [36]. These studies show that representation-learning approaches already capture important spatial and temporal dependencies in neural signals, but they are designed for general-purpose pretraining rather than EEG–visual semantic alignment. Our work builds on these developments by explicitly modeling and jointly fusing temporal dynamics, adaptive spectral structure, and input-dependent spatial interactions, an integration that remains comparatively underexplored for zero-shot EEG–visual decoding.

2.2 Generative reconstruction with diffusion models

Beyond retrieval and classification, neural decoding has also been formulated as a conditional image generation problem. Early reconstruction methods based on GANs were limited by instability during training and high data requirements, motivating the use of pretrained diffusion models. Takagi and Nishimoto [37] showed that latent diffusion models can reconstruct high-resolution images from fMRI signals. This direction was later extended to EEG, where DreamDiffusion [21] generated plausible images from noisy EEG inputs, and ATM-S [24] improved reconstruction by mapping EEG features into CLIP's latent space. While diffusion-based approaches improve visual quality, their success still depends on the stability and semantic fidelity of the EEG embeddings used as conditioning signals.

2.3 Bridging the semantic and modality gap

Several recent works address the mismatch between visual stimuli, perceptual representations, and neural recordings. UBP [38] introduced an uncertainty-aware blur prior to reduce perceptual information loss and measurement noise. Human-Aligned Image Models [39] showed that vision encoders aligned with human perceptual judgments improve decoding across EEG and fMRI. CognitionCapturer [40] incorporated multimodal priors, including depth and textual representations, to better capture cognitive states underlying visual perception.

These works highlight the importance of semantic and perceptual alignment, but many still assume fixed or weakly structured EEG encoders. In contrast, our method focuses on improving the EEG representation itself by jointly modeling long-range temporal dynamics, adaptive spectral structure, and spatial electrode interactions, enabling more stable alignment with pretrained visual representations.

3 Method

We present a multiview EEG–image alignment framework for zero-shot visual decoding. The following sections define the problem formulation, describe the proposed multiview EEG–image decoding framework, and introduce the training objective.

3.1 Problem formulation

Let $\mathbf{X} \in \mathbb{R}^{C\times T}$ denote a multi-channel EEG signal with C electrodes and T time points, recorded while a subject views a visual stimulus I. Our objective is to learn an EEG encoder $f_\theta(\cdot)$ that maps neural signals to a semantic embedding $\mathbf{z}_{\text{eeg}} \in \mathbb{R}^d$ aligned with a pretrained vision embedding $\mathbf{z}_{\text{img}} \in \mathbb{R}^d$, where d is the embedding dimension. By learning this alignment, visual decoding can be formulated as zero-shot semantic classification in a shared embedding space, enabling generalization beyond closed-set categories. Given a dataset of paired EEG recordings and visual stimuli, the

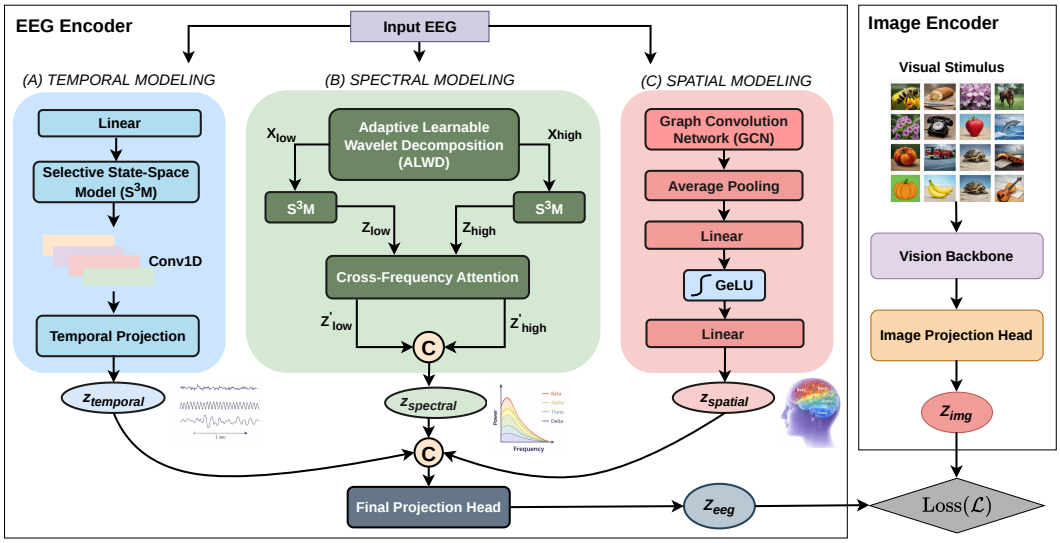


Figure 1: Overview of the proposed multiview EEG–Image decoding framework. Given a multichannel EEG trial, the EEG encoder constructs complementary temporal, spectral, and spatial representations using selective state-space modeling (S^3M), adaptive learnable wavelet-inspired decomposition (ALWD), and an attention-modulated graph convolutional network (GCN). These representations are fused into a compact EEG embedding, which is aligned with pretrained visual representations for zero-shot semantic decoding.

encoder is trained to produce discriminative and semantically meaningful EEG representations that reflect the underlying perceptual content.

3.2 Multiview EEG–Image decoding framework

As illustrated in Figure 1, the proposed framework is a multimodal architecture that aligns EEG signals with visual stimuli. It consists of a dual-stream setup: an EEG encoder that factorizes input signals into complementary temporal, spectral, and spatial components, and a pretrained image encoder that extracts visual representations. The EEG components are processed in parallel and then fused into a unified embedding. Finally, both the EEG and image features are projected into a shared latent space and aligned using a contrastive alignment loss, augmented with EEG-specific regularization terms.

3.2.1 EEG ENCODER

The key blocks of the EEG encoder are (a) Temporal Modeling, (b) Spectral Modeling, and (c) Spatial Modeling. These blocks are discussed in detail as follows.

(A) TEMPORAL MODELING

To capture long-range and input-adaptive temporal dependencies in EEG signals, we employ a Selective State-Space Model (S^{3}M). Given an input EEG sequence $X \in \mathbb{R}^{C\times T}$, the model maintains a latent state $h_t \in \mathbb{R}^{d_s}$ that evolves over time according to

$$h_t = \bar{A}_t \odot h_{t-1} + \bar{B}_t x_t, \qquad \bar{A}_t = \exp(\Delta_t A), \tag{1}$$

where x_t is the embedded EEG token at time step t, Δ_t is an input-dependent step size, A is a learnable state transition parameter, and $\bar{B}_t$ is an input-dependent input projection. The output at each time step is computed as

$$y_t = \sum_{k=1}^{d_s} C_{t,k}\, h_{t,k} \tag{2}$$

where $C_t \in \mathbb{R}^{d_s}$ is an input-dependent readout vector. The sequence of outputs from the S^3M is subsequently processed with a 1D convolution (Conv1d) along the temporal axis to capture local temporal patterns and refine short-range dependencies. The convolved features are then passed through a temporal projection layer that maps the representations to the desired feature dimensionality.

The resulting representation (z_{temporal}) serves as the final temporal feature embedding for subsequent processing.

(B) SPECTRAL MODELING

EEG signals exhibit complex temporal-spectral characteristics, in which discriminative information may lie in frequency bands that vary across subjects and tasks. Conventional EEG analysis pipelines rely on predefined frequency bands (e.g., delta, theta, alpha), which impose rigid and task-agnostic assumptions. To overcome these limitations, we introduce an *adaptive learnable wavelet decomposition* (ALWD) that dynamically selects frequency representations conditioned on the input EEG signal.

Let $\{\phi_k, \psi_k\}_{k=1}^{K}$ denote a bank of learnable low-pass and high-pass filters. We compute sample-specific selection weights $\mathbf{w} = \text{Softmax}(\mathcal{M}(X)) \in \mathbb{R}^K$ via a projection network $\mathcal{M}$. The input-adapted filters (Φ, Ψ) are synthesized as weighted combinations, and the spectral components are obtained via strided convolution ($\circledast_{\downarrow 2}$):

$$\begin{aligned} \Phi = \textstyle\sum_{k=1}^{K} w_k \phi_k, \qquad \Psi = \textstyle\sum_{k=1}^{K} w_k \psi_k, \\ X_{low} = X \circledast_{\downarrow 2} \Phi, \qquad X_{high} = X \circledast_{\downarrow 2} \Psi. \end{aligned} \tag{3}$$

This process produces complementary low- and high-frequency EEG representations with distinct temporal characteristics. Specifically, X_{low} captures slow, stable neural oscillations, while X_{high} emphasizes transient, rapidly changing activity. To account for these differences, we apply separate Selective State-Space Models (S^3M) to each spectral component, as defined in Eq. (4).

$$\begin{aligned} z_{low} &= \mathrm{S}^3\mathrm{M}_{low}(X_{low}), \\ z_{high} &= \mathrm{S}^3\mathrm{M}_{high}(X_{high}). \end{aligned} \tag{4}$$

By conditioning state transitions on the input signal, the S^3M naturally adapts its effective temporal memory to the spectral content of each component, enabling slow state evolution for low-frequency dynamics and faster adaptation for high-frequency transients.

To capture interactions across frequency scales, we further employ cross-frequency attention, allowing information exchange between low- and high-frequency temporal representations. Let $\mathrm{Attn}(Q, K, V)$ denote the standard multi-head attention operation. The cross-attended features are computed as:

$$\begin{aligned} z'_{low} &= \mathrm{Attn}(z_{low}, z_{high}, z_{high}), \\ z'_{high} &= \mathrm{Attn}(z_{high}, z_{low}, z_{low}). \end{aligned} \tag{5}$$

In this formulation, z'_{low} integrates high-frequency context into the stable low-frequency features, while z'_{high} incorporates low-frequency trends into the transient high-frequency features. The attended features $z^{'}_{low}$ and $z^{'}_{high}$ are then concatenated, temporally upsampled to the original resolution, and projected to obtain the spectral embedding $z_{spectral} \in \mathbb{R}^{B \times d}$.

(C) SPATIAL MODELING

EEG electrodes are spatially distributed over the scalp, capturing neural activity from interacting brain regions. Consequently, meaningful EEG representations must account for spatial dependencies rather than treating electrodes independently. To this end, we employ a Graph Convolutional Network (GCN) in which electrodes serve as nodes. We define a dynamic adjacency matrix $\mathcal{A}$ that combines a learnable global structure A_{global} with input-dependent edge attention $A_{att}(X)$. The spatial embedding ($z_{spatial}$) is computed as:

$$\begin{aligned} \mathcal{A} &= A_{global} \odot A_{att}(X), \\ z_{spatial} &= \sigma(\mathcal{A} X W_{GCN}). \end{aligned} \tag{6}$$

where W_{GCN} is the weight matrix and σ denotes the non-linear activation. This formulation captures both shared spatial structure and sample-specific functional interactions.

Finally, the EEG encoder integrates complementary representations that capture temporal, spectral, and spatial information. These embeddings are concatenated and mapped through a final projection W_{proj} to yield the aligned EEG representation z_{eeg}:

$$z_{eeg} = W_{proj}[z_{temporal} \parallel z_{spectral} \parallel z_{spatial}] \tag{7}$$

where $\parallel$ denotes the concatenation operation.

3.2.2 IMAGE ENCODER

In this work, we employ CLIP-ViT-H-14 [41] as the image encoder, pretrained on large-scale image datasets. The encoder extracts semantically discriminative visual embeddings that serve as fixed targets for contrastive alignment with EEG representations. To this end, all images are resized to 224×224 to ensure consistent and stable feature representations. All stimulus images are passed through the pretrained CLIP-ViT-H-14 once prior to training the EEG model, and the resulting visual embeddings are stored and reused throughout training. The image encoder parameters are frozen during training, which reduces computational cost and ensures a stable visual embedding space for learning EEG–image correspondence. The image encoder produces an image representation z_{img}, which is then used for cross-modal alignment.

3.2.3 Training objective

We aim to learn a shared embedding space that aligns neural responses with visual stimuli. Let $\mathbf{z}_{\text{eeg}}, \mathbf{z}_{\text{img}} \in \mathbb{R}^d$ denote the ℓ_2-normalized EEG and image embeddings, respectively. We optimize the symmetric contrastive alignment loss as follows:

$$\mathcal{L}_{\text{align}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\langle \mathbf{z}_{\text{eeg}}^{(i)}, \mathbf{z}_{\text{img}}^{(i)} \rangle / \tau)}{\sum_{j=1}^{B} \exp(\langle \mathbf{z}_{\text{eeg}}^{(i)}, \mathbf{z}_{\text{img}}^{(j)} \rangle / \tau)}, \tag{8}$$

where τ is a temperature parameter. To account for EEG variability and class-level structure, we introduce repetition consistency regularization ($\mathcal{L}_{\text{RCR}}$) and prototype-level contrastive regularization ($\mathcal{L}_{\text{CPCA}}$). These regularizers introduce EEG-specific inductive biases that improve robustness to trial-level variability and enhance class-level semantic structure, detailed in Appendix A. The total training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda_{\text{RCR}}\mathcal{L}_{\text{RCR}} + \lambda_{\text{CPCA}}\mathcal{L}_{\text{CPCA}}, \tag{9}$$

where λ_{RCR} and λ_{CPCA} control the regularization strengths.

4 Experimental setup

We evaluate the proposed framework on THINGS-EEG benchmark under within-subject, cross-subject, and cross-session zero-shot decoding settings. This section summarizes the dataset, evaluation protocols, and implementation details.

4.1 Dataset and preprocessing

We evaluate our method on **THINGS-EEG** dataset [13], a large-scale benchmark for EEG-based visual object recognition. The dataset contains EEG recordings collected while participants viewed natural images drawn from THINGS-EEG database, which includes 1,854 object concepts. EEG data were recorded from 10 subjects, each completing four sessions. The training set consists of 16,540 image conditions (1,654 concepts $\times$ 10 images), each repeated four times, while the test set contains 200 held-out object concepts, each presented 80 times, ensuring no overlap in object concepts between training and testing. Additional dataset characteristics and cross-subject variability analyses are provided in Appendix B.

4.2 Task definitions and evaluation protocols

We evaluate EEG-image decoding under three zero-shot settings: **within-subject**, **cross-subject**, and **cross-session**, where test classes are disjoint from training classes.

In the within-subject setting, models are trained and evaluated separately for each subject using subject-specific splits. In the cross-subject setting, models are trained on multiple subjects and evaluated on a held-out subject to assess cross-subject generalization. In the cross-session setting, for each subject, models are trained on EEG trials from three recording sessions and evaluated on the held-out fourth session. In both cross-subject and cross-session settings, the training and test EEG data are strictly disjoint, and no EEG trials or temporally adjacent segments from the same recording are shared across the training and test splits. This protocol tests robustness to session-dependent distribution shifts caused by changes in recording noise, electrode contact, and subject state. Although cross-session evaluation has been studied in other EEG tasks, it remains largely unexplored in the context of EEG-image semantic decoding.

Table 1: Overall accuracy (%) for 200-way zero-shot classification under the **within-subject** setting across 10 subjects. Our method achieves the best average performance.

Method	Sub-01		Sub-02		Sub-03		Sub-04		Sub-05		Sub-06		Sub-07		Sub-08		Sub-09		Sub-10		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
BraVL [22]	6.1	17.9	4.9	14.9	5.6	15.1	6.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	7.0	19.7	5.8	17.5
NICE [23]	12.3	36.6	13.1	39.0	16.4	47.0	8.0	26.9	14.1	40.6	15.2	42.1	20.0	49.9	13.3	37.1	14.9	41.9	14.9	41.9	13.8	39.5
NICE-SA [23]	17.2	44.1	14.9	52.0	12.6	38.3	11.2	34.7	16.3	52.5	10.1	32.2	15.4	49.5	12.2	39.9	10.3	30.1	10.3	30.1	14.7	40.7
NICE-GA [23]	18.5	45.0	15.5	52.7	13.2	38.8	11.7	35.6	17.1	53.3	10.5	33.0	16.0	50.2	12.8	40.3	11.1	30.8	11.1	30.8	15.6	41.2
ATM-S [24]	25.6	60.4	22.0	54.5	25.0	62.4	31.4	60.9	12.9	43.0	21.3	51.1	30.5	61.5	38.8	72.0	34.4	51.5	29.1	63.5	28.5	60.4
UMind [31]	27.0	56.0	32.0	70.0	34.0	70.0	36.0	70.5	23.0	50.5	32.5	68.5	28.0	59.0	46.5	80.0	37.5	66.5	42.0	76.0	33.9	66.7
CC [40]	31.4	79.7	31.4	77.8	38.2	85.6	40.4	85.8	24.4	66.3	34.8	78.8	34.7	81.0	48.1	88.6	37.4	79.4	35.6	79.3	35.6	80.2
UBP [38]	41.2	70.5	51.2	80.9	51.2	82.0	51.1	76.9	42.2	72.8	57.5	83.5	49.0	79.9	58.6	85.8	45.1	76.2	61.5	88.2	50.9	79.7
Ours	**53.0**	**84.5**	**51.0**	**84.0**	**58.0**	**88.0**	**55.0**	**88.0**	**38.0**	**69.5**	**54.5**	**87.0**	**51.0**	**84.5**	**67.5**	**89.0**	**52.5**	**87.5**	**67.5**	**94.0**	**54.8**	**85.6**

Table 2: Overall accuracy (%) for 200-way zero-shot classification under the **cross-subject** setting (leave-one-subject-out). Our method achieves the best overall performance.

Method	Sub-01		Sub-02		Sub-03		Sub-04		Sub-05		Sub-06		Sub-07		Sub-08		Sub-09		Sub-10		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
BraVL [22]	2.3	8.0	1.5	6.3	1.4	5.9	1.7	6.7	1.5	5.6	1.8	7.2	2.1	8.1	2.2	7.6	1.6	6.4	2.3	8.5	1.8	7.0
NICE [23]	7.6	22.8	5.9	20.5	6.0	22.3	6.3	20.7	4.4	18.3	5.6	22.2	5.6	19.7	6.3	22.0	5.7	17.6	8.4	28.3	6.2	21.4
NICE-SA [23]	7.0	22.6	6.6	23.2	7.5	23.7	5.4	21.4	6.4	22.2	7.5	22.5	3.8	19.1	8.5	24.4	7.4	22.3	9.8	29.6	7.0	23.1
NICE-GA [23]	5.9	21.4	6.4	22.7	5.5	20.1	6.1	21.0	4.7	19.5	6.2	22.5	5.9	19.1	7.3	25.3	4.8	18.3	6.2	26.3	5.9	21.6
ATM-S [24]	10.5	26.8	7.1	24.8	11.9	33.8	14.7	**39.4**	7.0	23.9	11.1	35.8	**16.1**	43.5	15.0	40.3	4.9	22.7	**20.5**	46.5	11.8	33.7
UMind [31]	7.5	31.0	7.5	32.0	**15.5**	**37.5**	**17.0**	38.5	8.0	22.5	10.5	37.5	11.5	30.0	16.0	46.0	10.0	32.5	16.5	36.5	12.0	34.4
UBP [38]	11.5	29.7	15.5	40.0	9.8	27.0	13.0	32.3	8.8	33.8	11.7	31.0	10.2	23.8	12.2	32.2	15.5	40.5	16.0	43.5	12.4	33.4
Ours	**17.5**	**48.5**	**15.5**	**41.5**	9.0	34.0	12.5	37.5	**14.5**	**42.0**	**17.0**	**52.0**	14.0	**45.0**	**16.0**	**47.5**	**17.0**	**49.5**	19.5	**56.0**	**15.3**	**45.4**

Zero-shot classification is performed by aligning EEG embeddings with image-based class prototypes in a shared semantic space. Performance is evaluated using Top-1 and Top-5 accuracy, where Top-1 measures exact prediction correctness and Top-5 checks whether the correct class appears among the five most confident predictions.

4.3 Implementation details

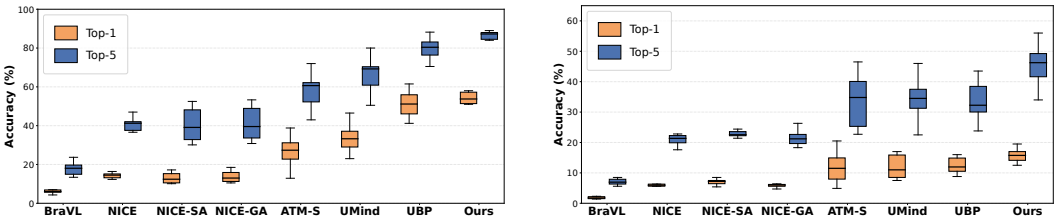
All models are trained for 50 epochs using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. Following common practice in THINGS-EEG evaluation, we report results from a fixed random seed to ensure direct comparability with prior work. We use a batch size of 64 for training, validation, and testing. All experiments are conducted on a single NVIDIA RTX 4500 Ada GPU with 24 GB memory. Additional implementation details, hyperparameter settings, and training configurations are provided in Appendix C.

5 Results

This section presents a comprehensive evaluation of the proposed method under three zero-shot EEG decoding settings: *within-subject*, *cross-subject*, and *cross-session* classification. We compare against recent representative methods, including BraVL [22], NICE, NICE-SA, NICE-GA [23], ATM-S [24], UMind [31], CognitionCapturer (CC) [40], and UBP [38]. Across all settings, the proposed method consistently achieves superior performance, demonstrating stronger robustness and semantic alignment of learned representations. Additional subject-wise analyses of representation stability are provided in Appendix D.

5.1 Within-Subject zero-shot classification

In the within-subject setting, the proposed method achieves the best overall performance across subjects. As shown in Table 1, our method obtains an average Top-1 accuracy of **54.8%** and Top-5 accuracy of **85.6%**, consistently outperforming all prior methods across subjects. Compared with the strongest baseline, UBP, our method improves Top-1 accuracy by **+3.9%** and Top-5 accuracy by **+5.9%**. These results indicate that the proposed framework effectively captures subject-specific neural dynamics while preserving strong semantic alignment with pretrained visual representations. As illustrated in Figure 2a, the proposed method also demonstrates higher median accuracy and lower variability across subjects, reflecting more stable and robust performance in the within-subject setting.



(a) Within-subject (b) Cross-subject

Figure 2: Zero-shot decoding performance across within-subject and cross-subject settings. The proposed method achieves higher median accuracy and lower performance variability compared to prior methods, indicating improved robustness and generalization.

Table 3: Overall accuracy (%) for 200-way zero-shot classification under the cross-session setting. Our method achieves the highest average Top-1 and Top-5 accuracy among the evaluated methods.

Method	Sub-01		Sub-02		Sub-03		Sub-04		Sub-05		Sub-06		Sub-07		Sub-08		Sub-09		Sub-10		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
NICE	14.0	38.5	15.5	34.5	21.0	43.5	21.0	47.0	12.0	28.0	22.0	45.0	12.0	36.0	29.0	59.0	18.0	46.0	21.5	49.5	18.6	42.7
NICE-SA	18.5	43.0	18.5	44.5	28.5	51.0	26.0	53.5	14.0	32.5	21.5	51.0	19.0	40.0	31.0	63.5	22.5	44.5	21.0	53.5	22.1	47.7
ATM-S	20.0	41.0	16.5	38.5	27.0	58.5	25.5	51.5	17.5	33.5	25.5	57.5	25.0	54.0	33.0	68.0	25.5	53.0	28.0	59.0	24.4	51.5
UBP	28.5	59.5	**39.0**	67.0	42.5	73.5	37.0	71.5	36.0	65.5	**44.0**	76.5	36.0	69.0	47.0	79.5	37.5	67.0	49.5	80.5	39.7	71.0
Ours	**31.0**	**67.5**	38.5	**76.4**	**44.0**	**81.1**	**38.6**	**75.7**	**37.0**	**75.0**	40.0	**78.2**	**39.0**	**77.0**	**50.2**	**86.6**	**39.3**	**77.5**	**50.6**	**85.2**	**40.8**	**78.0**

Table 4: Ablation study of temporal, spectral, and spatial EEG representation modules under the within-subject setting on THINGS-EEG benchmark. Results show average Top-1 and Top-5 classification accuracy (%) across all ten subjects.

Metric	Full Model	w/o Temporal	w/o Spatial	w/o Spectral	Temporal Only	Spatial Only	Spectral Only
Top-1 ↑	**54.8**	33.1	44.7	46.0	44.7	33.7	37.5
Top-5 ↑	**85.6**	64.7	79.8	80.0	78.7	66.2	71.8

Table 5: Ablation study evaluating the contribution of CPCA and RCR regularization terms under a within-subject setting.

Configuration	Sub-01		Sub-02		Sub-03		Sub-04		Sub-05		Sub-06		Sub-07		Sub-08		Sub-09		Sub-10		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Full model	53.0	84.5	51.0	84.0	58.0	88.0	55.0	88.0	38.0	69.5	54.5	87.0	51.0	84.5	67.5	89.0	52.5	87.5	67.5	94.0	54.8	85.6
w/o CPCA	53.5	82.5	49.5	81.5	59.5	89.5	54.5	87.0	36.5	70.0	53.5	86.0	50.5	86.0	66.0	89.0	52.0	86.0	70.5	93.0	54.6	85.1
w/o RCR	55.0	80.5	49.0	82.0	40.0	78.5	52.0	87.5	43.0	71.0	50.0	87.0	53.5	88.0	66.5	91.5	52.0	85.5	67.5	95.0	52.9	84.7
w/o CPCA & RCR	39.0	73.5	33.0	67.5	46.0	82.5	44.5	78.5	22.5	53.5	40.5	75.5	43.5	81.5	58.5	88.5	37.0	66.0	54.5	85.0	41.9	75.2

5.2 Cross-Subject zero-shot classification

In the cross-subject setting, we evaluate whether the proposed method can generalize to unseen subjects despite substantial inter-subject variability in EEG signals. The model is trained on all but one subject and tested on the held-out subject. As shown in Table 2, the proposed method achieves an average Top-1 accuracy of **15.3%** and Top-5 accuracy of **45.4%**. Compared with UBP, the strongest baseline in this setting, our method improves Top-1 accuracy by **+2.9%** and Top-5 accuracy by **+12.0%**. These results demonstrate stronger generalization capability under subject-level distribution shifts. Figure 2b further shows higher median accuracy and lower performance variability across subjects, indicating more stable cross-subject decoding performance. Overall, these results suggest that the proposed method learns robust and semantically meaningful subject-invariant representations.

5.3 Cross-Session zero-shot classification

The cross-session setting evaluates whether the proposed method remains effective when training and testing are conducted on different recording sessions of the same subject. This setting is more challenging because EEG signals can exhibit session-dependent distribution shifts caused by recording noise, fatigue, electrode variability, and changes in subject state. We further present a systematic evaluation of cross-session zero-shot EEG–image semantic decoding and retrieval. We compare our method against four established baselines: NICE, NICE-SA, ATM-S, and UBP. As shown in Table 3, our method achieves the highest average performance, with 40.8% Top-1 and 78.0% Top-5 accuracy, outperforming the strongest baseline, UBP (39.7% Top-1 and 71.0% Top-5). These results indicate that the proposed multiview representation learning framework improves cross-session Top-5 retrieval performance, although Top-1 generalization remains challenging.

Table 6: Statistical significance of the proposed method against the compared baselines under within-subject and cross-subject settings. Values below 0.05 indicate statistically significant differences.

(a) Within-subject							
Metric	**BraVL**	**NICE**	**NICE-SA**	**NICE-GA**	**ATM-S**	**UMind**	**UBP**
Top-1	3.16×10^{-8}	2.10×10^{-7}	4.37×10^{-7}	5.33×10^{-7}	1.83×10^{-7}	4.11×10^{-8}	0.0410
Top-5	3.32×10^{-10}	7.55×10^{-8}	3.15×10^{-6}	3.65×10^{-6}	1.12×10^{-7}	1.62×10^{-6}	0.00479
(b) Cross-subject							
Metric	**BraVL**	**NICE**	**NICE-SA**	**NICE-GA**	**ATM-S**	**UMind**	**UBP**
Top-1	8.98×10^{-8}	1.76×10^{-6}	3.79×10^{-6}	3.11×10^{-6}	0.0802	0.0931	0.00734
Top-5	9.37×10^{-9}	5.39×10^{-7}	9.35×10^{-7}	4.98×10^{-7}	0.00454	0.00344	3.85×10^{-4}

Table 7: Effect of adding CPCA and RCR losses to DWT under the within-subject setting on THINGS-EEG. Results report average Top-1 and Top-5 classification accuracy (%) across all ten subjects.

Model	**Top-1 ↑**	**Top-5 ↑**
DWT [25]	33.4	61.8
DWT + CPCA + RCR	33.2	62.3
Ours	**54.8**	**85.6**

5.4 Statistical Significance Analysis

To assess statistical significance, we conduct paired two-sided t-tests on subject-level Top-1 and Top-5 accuracies between our method and each baseline ($p < 0.05$); p-values are reported in Table 6. In the within-subject setting, our method significantly outperforms all baselines (BraVL, NICE, NICE-SA, NICE-GA, ATM-S, UMind, and UBP) in both Top-1 and Top-5 accuracy. In the cross-subject setting, Top-5 improvements are significant over all baselines, and Top-1 improvements are significant over all except ATM-S ($p = 0.0802$) and UMind ($p = 0.0931$), where the higher mean accuracy is not large enough relative to inter-subject variability to reach significance. However, for Top-5 accuracy, the improvements are statistically significant over all compared baselines. Overall, these results provide statistical evidence for the effectiveness of the proposed multiview representation learning framework.

5.5 Ablation Study

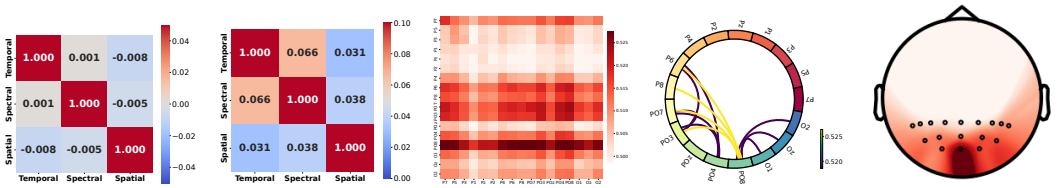
We conduct ablation studies to assess the contribution of the main architectural modules and regularization terms. All ablations follow the same training protocol and evaluation metrics as the full model. Table 4 reports the within-subject results when removing the temporal, spectral, or spatial module individually, showing the contribution of each EEG representation branch.

We also evaluate the effect of the proposed regularization terms by removing CPCA and RCR individually and jointly. Table 5 reports the average Top-1 and Top-5 accuracy across all 10 subjects under within-subject and cross-subject settings. Removing either regularizer degrades performance, while removing both causes the largest drop. The larger degradation from removing RCR suggests that relational consistency regularization is particularly important for preserving discriminative EEG representations. Additional details of RCR and CPCA are provided in Appendix A.

To isolate the contribution of the proposed architecture from the loss formulation, we retrain DWT with the same CPCA and RCR objectives under an identical training protocol. As shown in Table 7, adding these losses to DWT yields comparable performance, whereas our multiview model achieves substantially higher accuracy under the same objectives. This indicates that the observed improvements cannot be attributed to CPCA and RCR alone.

6 Discussion

Beyond the main decoding results, we analyze the learned representations to interpret how the proposed framework supports EEG–visual alignment. Specifically, we examine inter-view similarity among temporal, spectral, and spatial branches, and visualize the spatial connectivity patterns learned



(a) Cosine similarity (b) CKA similarity (c) Connectivity heatmap (d) Top graph connections (e) Mean topography

Figure 3: Analysis of learned EEG representations and spatial connectivity. (a) and (b) show interview similarity among temporal, spectral, and spatial representations using cosine similarity and CKA. Low off-diagonal values indicate complementary multiview representations. (c) shows the learned adaptive GCN adjacency matrix averaged across subjects. (d) visualizes the strongest learned graph connections. (e) shows the averaged electrode-importance topomap. The connectivity and topographic patterns emphasize posterior and occipital regions associated with visual processing.

by the adaptive GCN. Limitations and broader impacts, including generalization beyond controlled EEG settings and neural-data privacy considerations, are discussed in Appendix E.

6.1 Representation disentanglement

We analyze the degree of alignment between the learned temporal, spectral, and spatial representations using cosine similarity and Centered Kernel Alignment (CKA). As shown in Figure 3(a) and (b), cosine similarities remain close to zero ($|\text{mean}| \leq 0.01$), while CKA scores are generally below 0.07. These low similarity scores suggest that the learned branches occupy less-aligned representation subspaces. However, low cosine similarity and CKA do not establish complete branch independence, particularly because the spectral representation is derived from the temporal EEG signal and normalization may influence the measured similarities. To further assess consistency across individuals, we provide per-subject average cosine-similarity and CKA visualizations in Appendix D.1 and Appendix D.2, respectively. Separately, the all-subject ablation results show that removing any individual branch reduces average decoding performance, supporting the contribution of each branch to the overall framework.

6.2 Learned EEG connectivity

We further examine the adaptive GCN by visualizing the learned adjacency matrices averaged across subjects. Figure 3(c)–(e) shows that the learned connectivity is non-uniform and consistently emphasizes posterior and occipital electrodes, including PO7, PO8, O1, Oz, and neighboring parieto-occipital regions. The sparse graph also reveals long-range interactions between parietal and occipital areas. These patterns suggest that the graph-learning module captures spatial dependencies relevant to visual EEG decoding rather than relying on random or uniformly distributed electrode interactions. In addition to the group-level visualization, Appendix D.3 provides per-subject average topographic maps of learned EEG connectivity. These subject-wise visualizations further demonstrate that the spatial patterns learned by the model are broadly consistent across participants while still reflecting individual variability in EEG responses.

7 Conclusion

In this work, we presented a unified framework for EEG-based zero-shot visual decoding through multiview neural representation learning. By jointly modeling temporal dynamics, adaptive spectral structure, and spatial electrode interactions, the proposed method learns compact EEG embeddings that align effectively with pretrained visual representations in a shared semantic space. Experiments on THINGS-EEG benchmark show consistent improvements over prior approaches under within-subject, cross-subject, and cross-session settings. In particular, our systematic cross-session evaluation highlights the importance of structured EEG inductive biases for robust generalization under realistic recording variability. These findings suggest that multiview EEG modeling provides a principled foundation for EEG–visual alignment and offers a scalable path toward more reliable EEG-based neural decoding.

References

[1] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[2] Melvyn A. Goodale and A.David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. ISSN 0166-2236. doi: https://doi.org/10.1016/0166-2236(92)90344-8. URL `https://www.sciencedirect.com/science/article/pii/0166223692903448`.

[3] Keiji Tanaka. Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19(1): 109–139, 1996.

[4] Muhammad Firoz Mridha, Sujoy Chandra Das, Muhammad Mohsin Kabir, Aklima Akter Lima, Md Rashedul Islam, and Yutaka Watanobe. Brain-computer interface: Advancement and challenges. *Sensors*, 21(17):5746, 2021.

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

[6] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

[7] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–462, 2014.

[8] Frank Tong and Michael S Pratte. Decoding patterns of human brain activity. *Annual review of psychology*, 63(1):483–509, 2012.

[9] Hongqi Li, Xiaoya Li, and José R del Millán. Noninvasive eeg-based intelligent mobile robots: a systematic review. *IEEE Transactions on Automation Science and Engineering*, 22:6291–6315, 2024.

[10] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4503–4511. IEEE, 2017.

[11] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, 2015.

[12] Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792, 2019.

[13] Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

[14] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

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

[17] Matteo Ferrante, Tommaso Boccato, Stefano Bargione, and Nicola Toschi. Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models. *Computers in Biology and Medicine*, 178:108701, 2024.

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

[19] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak 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, 2020.

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

[21] Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023.

[22] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

[23] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

[24] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 102822–102864, 2024.

[25] Enshang Zhang, Zhicheng Zhang, and Takashi Hanakawa. Category-aware eeg image generation based on wavelet transform and contrast semantic loss. *arXiv preprint arXiv:2505.24301*, 2025.

[26] Yueming Sun and Long Yang. Spatial-functional awareness transformer-based graph archetype contrastive learning for decoding visual neural representations from eeg. *arXiv preprint arXiv:2509.24761*, 2025.

[27] Weijian Mai, Fengjie Wu, and Xiaoting Mai. Learning spatial-spectral-temporal eeg representations with dual-stream neural networks for motor imagery. *Biomedical Signal Processing and Control*, 92:106003, 2024.

[28] Ashwin Kamble, Pradnya H. Ghare, Vinay Kumar, Ashwin Kothari, and Avinash G. Keskar. Spectral analysis of eeg signals for automatic imagined speech recognition. *IEEE Transactions on Instrumentation and Measurement*, 72:1–9, 2023. doi: 10.1109/TIM.2023.3300473.

[29] Adel Hameed, Rahma Fourati, Boudour Ammar, Amel Ksibi, Ala Saleh Alluhaidan, Mounir Ben Ayed, and Hussain Kareem Khleaf. Temporal–spatial transformer based motor imagery classification for bci using independent component analysis. *Biomedical Signal Processing and Control*, 87:105359, 2024.

[30] Yonghao Song, Yijun Wang, Huiguang He, and Xiaorong Gao. Recognizing natural images from eeg with language-guided contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[31] Chengjian Xu, Yonghao Song, Zelin Liao, Haochuan Zhang, Qiong Wang, and Qingqing Zheng. Umind: A unified multitask network for zero-shot m/eeg visual decoding. *arXiv preprint arXiv:2509.14772*, 2025.

[32] Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-jepa: towards seamless cross-dataset transfer through dynamic spatial attention. *arXiv preprint arXiv:2403.11772*, 2024.

[33] Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

[34] Tongtian Yue, Shuning Xue, Xuange Gao, Yepeng Tang, Longteng Guo, Jie Jiang, and Jing Liu. Eegpt: Unleashing the potential of eeg generalist foundation model by autoregressive pre-training. 2024.

[35] Xujia Wang, Xuhui Liu, Xi Liu, Qian Si, Zhaoliang Xu, Yang Li, and Xiantong Zhen. Eeg-dino: Learning eeg foundation models via hierarchical self-distillation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 196–205. Springer, 2025.

[36] Zijian Dong, Ruilin Li, Yilei Wu, Thuan T Nguyen, Joanna S Chong, Fang Ji, Nathanael R Tong, Christopher L Chen, and Juan H Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

[37] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14453–14463, 2023.

[38] Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2246–2257, 2025.

[39] Nona Rajabi, Antônio H Ribeiro, Miguel Vasco, Farzaneh Taleb, Mårten Björkman, and Danica Kragic. Human-aligned image models improve visual decoding from the brain. In *Forty-second International Conference on Machine Learning*, 2025.

[40] Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14486–14493, 2025.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Supplementary material

This supplementary material provides additional implementation details, dataset analysis, regularization formulations, representation analysis, and subject-wise visualizations supporting the findings presented in the main paper.

The supplementary material is organized as follows.

- Appendix A provides details of EEG-specific regularization.
- Appendix B describes dataset preprocessing and subject variability.
- Appendix C presents the implementation and hyperparameter details.
- Appendix D provides subject-wise disentanglement and connectivity analyses.
- Appendix E discusses limitations and broader impacts.

A EEG-Specific regularization

This section provides additional details on the EEG-specific regularization components used in this work, namely, repetition consistency regularization (RCR) and prototype-level contrastive regularization (CPCA). As described in Section 3.2.3 of the main paper, these regularizers augment the primary EEG–image contrastive objective without modifying the underlying task definition or inference-time architecture.

A.1 Repetition consistency regularization (RCR)

EEG recordings are inherently noisy and exhibit substantial trial-to-trial variability, even for repeated presentations of the same visual stimulus. To encourage stable neural representations, repetition consistency regularization penalizes deviations of single-trial EEG embeddings from a repetition-averaged reference.

Let $\{\mathbf{z}_e^{(r)}\}_{r=1}^{R}$ denote the EEG embeddings corresponding to R repetitions of the same stimulus. The repetition-averaged embedding is defined as:

$$\bar{\mathbf{z}}_e = \frac{1}{R}\sum_{r=1}^{R}\mathbf{z}_e^{(r)}, \tag{10}$$

followed by ℓ_2 normalization. For a given trial embedding $\mathbf{z}_e$, the repetition consistency loss is:

$$\mathcal{L}_{\text{RCR}} = \left\|\hat{\mathbf{z}}_e - \hat{\bar{\mathbf{z}}}_e\right\|_2^2, \tag{11}$$

where $\hat{\cdot}$ denotes ℓ_2 normalization.

Similarity-conditioned weighting. Not all EEG trials contribute equally to representation learning due to variability in signal quality and cross-modal alignment strength. We therefore modulate the contribution of repetition consistency using a similarity-conditioned weighting scheme derived from cross-modal similarity statistics. EEG trials exhibiting weaker alignment with their paired visual embeddings are regularized more strongly, while well-aligned trials are regularized more weakly, improving robustness to noisy or ambiguous neural responses.

A.2 Prototype-Level contrastive regularization (CPCA)

While instance-level contrastive learning aligns EEG embeddings with individual image features, it does not explicitly enforce class-level semantic structure. Prototype-level contrastive regularization addresses this limitation by anchoring EEG representations to class prototypes derived from visual embeddings.

For each class c, a prototype $\boldsymbol{\mu}_c$ is computed as the normalized mean of visual embeddings belonging to that class:

$$\boldsymbol{\mu}_c = \frac{\sum_{i\in\mathcal{D}_c}\mathbf{z}_v^{(i)}}{\left\|\sum_{i\in\mathcal{D}_c}\mathbf{z}_v^{(i)}\right\|_2}, \tag{12}$$

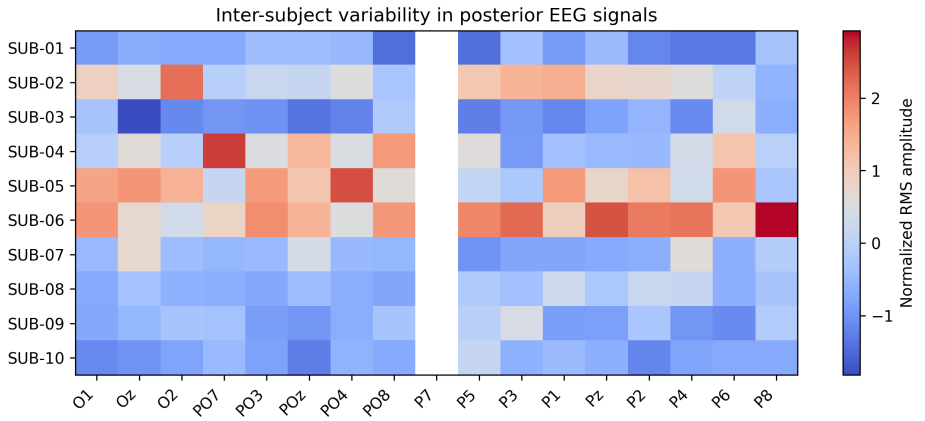


Figure 4: Cross-subject variability of posterior EEG signals in THINGS-EEG dataset. Z-scored RMS amplitudes are shown for each subject and posterior electrode, computed across all trials. The pronounced variability across subjects and electrodes highlights the challenges of subject-generalized EEG decoding.

where $\mathcal{D}_c$ denotes the set of training images labeled as class c.

Given an EEG embedding $\mathbf{z}_e$ with label y, its similarity to all class prototypes is computed and the Top-K most similar prototypes are selected. The prototype-level contrastive loss is defined as:

$$\mathcal{L}_{\text{CPCA}} = -\log \frac{\exp(\langle \mathbf{z}_e, \boldsymbol{\mu}_y \rangle / \tau_p)}{\sum_{c \in \mathcal{K}} \exp(\langle \mathbf{z}_e, \boldsymbol{\mu}_c \rangle / \tau_p)}, \tag{13}$$

where $\mathcal{K}$ denotes the set of selected prototypes and τ_p is a temperature parameter.

Restricting the contrastive competition to a semantically relevant subset of prototypes improves class-level discrimination while reducing noise introduced by global negatives.

Together, RCR and CPCA inject EEG-specific inductive biases into the training process. Repetition consistency improves robustness to trial-level variability, while prototype-level alignment enhances semantic structure in the learned representations.

B Dataset preprocessing, characteristics, and subject variability

EEG signals were acquired using a 64-channel system at a sampling rate of 1000 Hz. EEG preprocessing follows the protocol introduced in prior work on THINGS-EEG dataset [38]. Specifically, signals were bandpass filtered, epoched from -200 ms to 800 ms relative to stimulus onset, baseline-corrected using the pre-stimulus interval, and downsampled to 250 Hz. Session-wise multivariate noise normalization and whitening were applied to mitigate noise and reduce inter-session variability. Consistent with prior work, we retain posterior electrodes associated with visual processing. No additional artifact rejection is performed. Following the preprocessing protocol of UBP [38], we retain posterior electrodes associated with visual processing and apply standard EEG preprocessing, including bandpass filtering, epoch extraction, baseline correction, downsampling, and session-wise normalization.

EEG signals are segmented into 1-second epochs (250 time points) aligned to stimulus onset. Stimulus images were originally presented at a resolution of 500×500 pixels. For model input, images are resized to 224×224 pixels and normalized using standard preprocessing pipelines compatible with pretrained vision and vision–language encoders. Figure 4 illustrates the substantial inter-subject variability present in posterior EEG responses across subjects and electrodes.

Table 8: Implementation details and hyperparameters used in all experiments.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Batch size	64
Training epochs	50
Early stopping patience	5
Embedding dimension	1024
Image encoder	CLIP-ViT-H-14
Image resolution	224×224
EEG sampling rate	250 Hz
EEG epoch duration	1 second
Contrastive temperature τ	0.07
CPCA temperature τ_p	0.07
CPCA Top-K prototypes	20
λ_{RCR}	0.05
λ_{CPCA}	0.05
GPU	NVIDIA RTX 4500 Ada (24 GB)
Framework	PyTorch Lightning

C Implementation details and hyperparameters

All experiments are implemented in PyTorch and PyTorch Lightning and trained on a single NVIDIA RTX 4500 Ada GPU with 24 GB memory. The proposed framework is trained using the AdamW optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. Models are trained for 50 epochs using a batch size of 64. Early stopping is applied with a patience of 5 epochs. Following common practice in THINGS-EEG evaluation, all experiments use a fixed random seed to ensure direct comparability with prior work. All of the implementation and hyperparameter details are shown in Table 8.

The image encoder is based on pretrained CLIP-ViT-H-14 representations. The image encoder parameters remain frozen during training, while only the EEG encoder and projection layers are optimized. All stimulus images are resized to 224×224 prior to feature extraction. EEG signals are represented using posterior electrodes and segmented into 1-second epochs containing 250 temporal samples.

The temporal branch employs selective state-space modeling (S^3M) followed by temporal convolution and projection layers. The spectral branch uses adaptive learnable wavelet decomposition with learnable low-pass and high-pass filters, followed by cross-frequency attention. The spatial branch employs a graph convolutional network with dynamically learned adjacency matrices combining global graph structure and input-dependent edge attention.

For EEG-image alignment, both EEG and image embeddings are projected into a shared semantic embedding space of dimension 1024. All embeddings are ℓ_2 normalized prior to contrastive learning. The symmetric contrastive alignment objective uses a temperature parameter $\tau = 0.07$.

The repetition consistency regularization (RCR) and prototype-level contrastive regularization (CPCA) are jointly optimized with the alignment objective. Unless otherwise stated, the regularization weights are set to $\lambda_{\mathrm{RCR}} = 0.05$ and $\lambda_{\mathrm{CPCA}} = 0.05$. For CPCA, Top-$K$ prototype selection uses $K = 20$ candidate prototypes with prototype temperature $\tau_p = 0.07$. Similarity-conditioned weighting is enabled during training to adaptively regularize uncertain EEG trials.

D Subject-wise representation analysis

This section presents additional subject-wise analyses of the learned EEG representations under the within-subject evaluation setting. We analyze inter-view representation similarity, centered kernel alignment (CKA), and spatial electrode importance patterns across all ten subjects to further validate the disentanglement and interpretability of the proposed multiview framework.

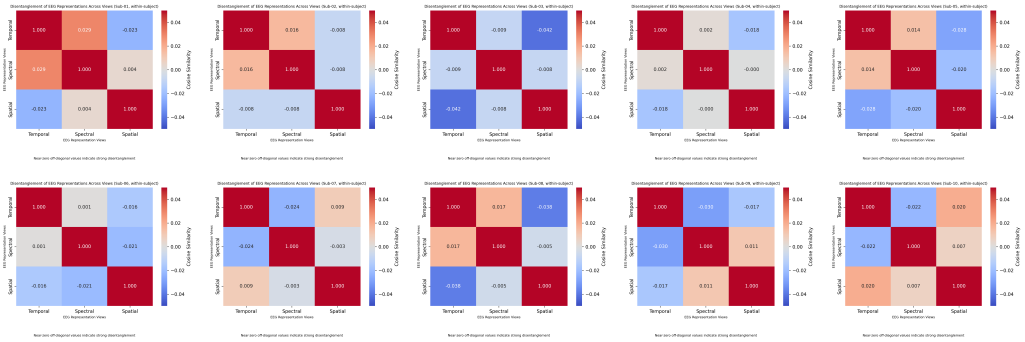


Figure 5: Subject-wise cosine similarity matrices between temporal, spectral, and spatial EEG representations across all ten subjects. The low off-diagonal similarity values indicate relatively low alignment between the learned representations across different branches.

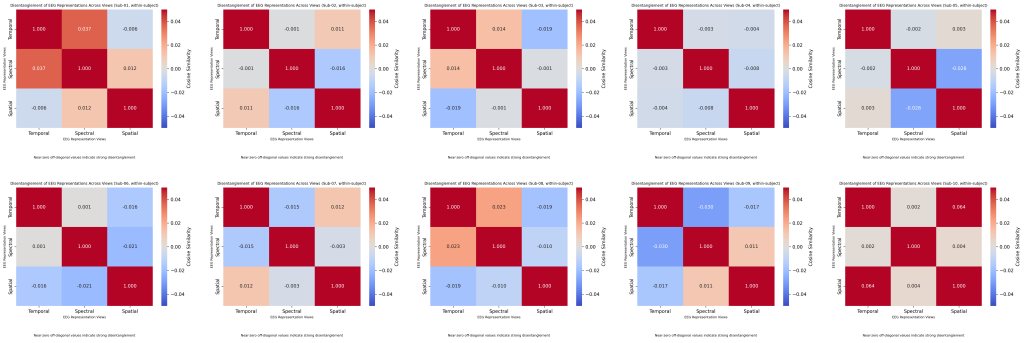


Figure 6: Subject-wise cosine similarity matrices under Top-5 semantic retrieval evaluation.

D.1 Representation Similarity Matrices

To further examine the similarity among the learned multiview representations, we compute pairwise cosine similarities between the temporal, spectral, and spatial EEG embeddings for each subject. The resulting similarity matrices characterize the degree of alignment between the learned representations, with lower off-diagonal values indicating less similarity between different branches. The subject-wise similarity matrices are shown in Figure 5.

Figure 6 further shows the corresponding similarity structure under Top-5 evaluation. Similar trends across subjects indicate that the learned multiview representations preserve complementary semantic structure even when broader semantic neighborhoods are considered.

D.2 Centered kernel alignment (CKA) analysis

To further quantify representation similarity beyond cosine distance, we compute centered kernel alignment (CKA) between the temporal, spectral, and spatial branches. Compared with cosine similarity, CKA provides a more robust measure of representational overlap and functional redundancy.

Figure 7 and 8 present the corresponding CKA analysis under Top1 and Top-5 retrieval evaluation. The consistently low inter-view similarity scores suggest stable disentanglement across subjects and semantic retrieval settings.

D.3 Subject-wise electrode importance topomaps

To examine the spatial patterns learned by the adaptive EEG encoder, we visualize electrode importance distributions for each subject. These topomaps indicate the relative contribution of different electrodes to the learned EEG representations. Across subjects, the model consistently assigns higher importance to posterior and occipital electrodes, which are closely associated with visual cortical processing, while still showing subject-specific variations due to individual neural responses and recording variability. These results suggest that the proposed framework captures spatial structure

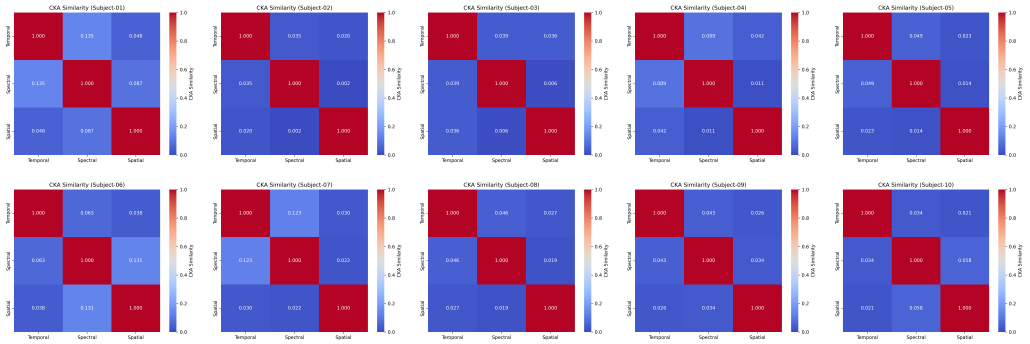


Figure 7: Subject-wise centered kernel alignment (CKA) similarity matrices between temporal, spectral, and spatial EEG representations across all ten subjects. The low inter-view CKA values indicate relatively low representational similarity between the learned branches across subjects.

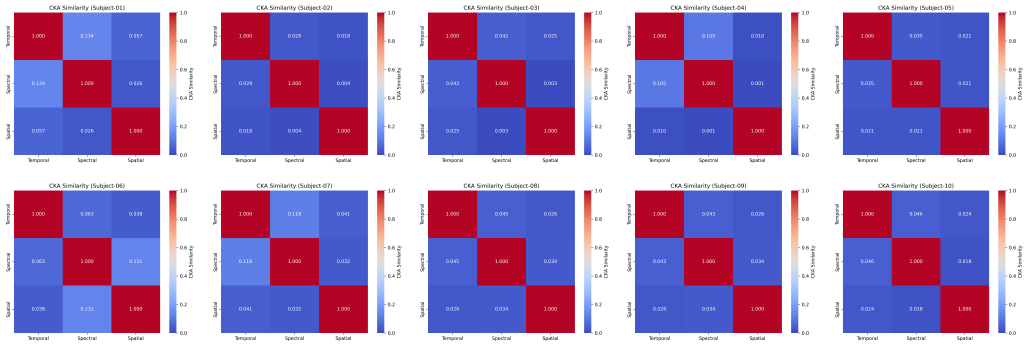


Figure 8: Subject-wise CKA similarity matrices under Top-5 semantic retrieval evaluation. The results demonstrate stable multiview representation disentanglement across subjects.

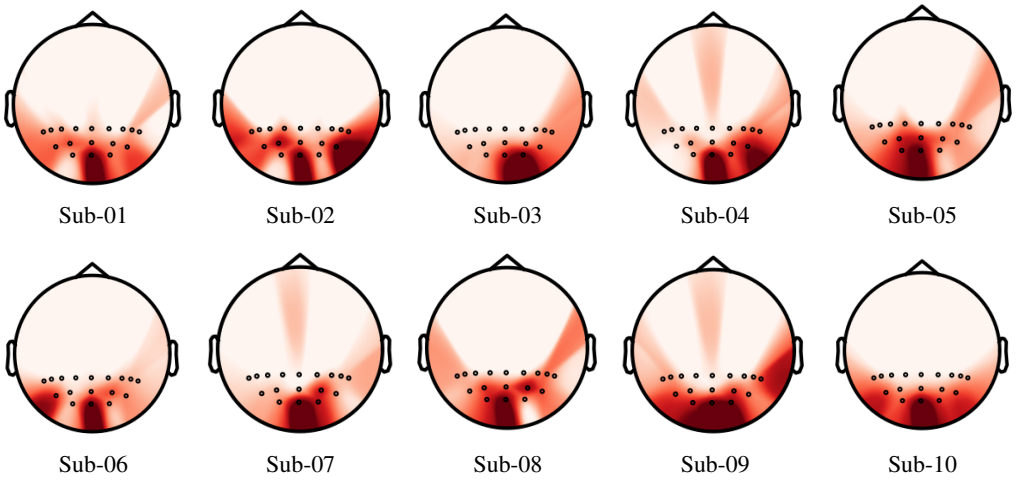


Figure 9: Subject-wise electrode importance topomaps for all participants under the within-subject EEG-image decoding setting. Across subjects, dominant activations consistently emerge over posterior and occipital regions, indicating that visual cortical areas contribute most strongly to EEG-based semantic representation learning.

relevant to visual decoding rather than relying on uniformly distributed electrode information. The complete subject-wise topographic maps are shown in Figure 9.

E Limitations and broader impacts

Despite strong performance across within-subject, cross-subject, and cross-session settings, several limitations remain. First, all experiments are conducted on THINGS-EEG benchmark, and generalization to other EEG paradigms, clinical populations, or real-world environments remains uncertain. Second, although the proposed framework improves robustness to inter-subject and inter-session variability, fully subject-invariant EEG decoding remains an open challenge due to the non-stationary nature of EEG signals. Third, the framework relies on pretrained CLIP representations and focuses primarily on semantic alignment rather than explicit image reconstruction. Our use of the final-layer CLIP representation is intended to provide a fixed semantic target for zero-shot EEG–visual decoding and fair architectural comparison, rather than a neurobiological model of visual processing. The current framework does not test or claim consistency with either the classical hierarchical view of visual processing or predictive-coding/Bayesian accounts. Furthermore, it does not establish a correspondence between CLIP representation layers and cortical processing stages. Investigating these relationships requires dedicated neuroscientific analyses and is beyond the scope of the present study. From a broader impact perspective, advances in EEG-based neural decoding may benefit assistive brain-computer interfaces, adaptive human-computer interaction, and neuroscience research. However, neural decoding technologies also raise important privacy and ethical concerns, since EEG signals may contain sensitive cognitive information. Responsible deployment therefore requires appropriate safeguards for informed consent, neural data security, and ethical use.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe the proposed EEG-image decoding framework, the uncertainty-aware learning strategy, and the reported performance gains on the THINGS-EEG benchmark. All claims are supported by experimental results and ablation studies presented in the main paper and supplementary material.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section discussing the scope of evaluation, assumptions of controlled EEG recording conditions, and reliance on frozen pretrained visual embeddings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper primarily focuses on empirical model design and experimental evaluation rather than formal theoretical analysis.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed descriptions of datasets, preprocessing steps, model architectures, training procedures, optimization settings, evaluation metrics, and ablation protocols necessary to reproduce the reported results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The implementation code, training scripts, and configuration details will be released upon publication, together with instructions for reproducing the experiments. Publicly available benchmark datasets and pretrained models used in this work are properly referenced.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies dataset splits, optimizer settings, learning rates, batch sizes, model hyperparameters, training schedules, evaluation protocols, and implementation details in the methodology and supplementary sections.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are reported across multiple subjects and experimental settings, with averaged performance metrics and variability statistics included in the supplementary material to demonstrate robustness and consistency.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper describes the hardware configuration, including GPU type, software framework, and training environment used for all experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. The work uses publicly available research datasets and focuses on methodological advances in EEG-based visual decoding.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses potential positive applications in assistive neurotechnology and brain-computer interfaces, as well as possible risks related to privacy, misuse of neural data, and ethical considerations surrounding EEG-based decoding systems.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The work does not release high-risk generative models or newly collected sensitive datasets. The study relies on publicly available benchmark datasets and pretrained models distributed under their original licenses.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets, pretrained models, and software frameworks used in this work are properly cited and acknowledged. Their corresponding licenses and usage conditions are respected and referenced where applicable.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Newly introduced implementation components, training configurations, and evaluation scripts are documented in the supplementary material and will be released with usage instructions to facilitate reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not conduct new crowdsourcing experiments or recruit new human participants. It only uses existing publicly available EEG benchmark datasets collected by prior studies.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve new data collection from human subjects. The utilized EEG datasets were previously collected and released by the original authors under their respective ethical approval procedures.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models are not part of the proposed methodology or experimental pipeline. Any use of language models was limited to writing assistance and did not affect the scientific contributions or experimental results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.