BIOBRIDGE: BRIDGING BRAIN AND VISION REPRESENTATIONS WITH BIOLOGICAL INDUCTIVE BIASES

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG based visual neural decoding seeks to infer visual content from neural responses and supports applications such as image retrieval. However, progress is limited by the scarcity of paired EEG and image data and by the substantial mismatch between high fidelity digital images and biological visual representations shaped by retinotopic organization and subject dependent neuroanatomy. To address these challenges, we propose **BioBridge**, a biologically guided brain and vision alignment framework that incorporates structured physiological inductive biases into representation learning. Specifically, Adaptive Blur with Visual Priors uses retinotopic weighting to transform visual inputs into perceptually structured representations, thereby reducing the discrepancy between digital stimuli and biological perception.Biomimetic Visual Feature Extraction then learns complementary visual representations at multiple levels that reflect hierarchical cortical processing and improve robustness to subject variability. Finally, Multi Level Bidirectional Contrastive Learning jointly aligns EEG and visual representations at different semantic levels through symmetric contrastive objectives in a shared embedding space. Experiments demonstrate that BioBridge achieves **80.5%** Top 1 accuracy and **97.6%** Top 5 accuracy in zero shot retrieval from EEG to images, substantially outperforming previous methods and exhibiting strong generalization across subjects and experimental settings. Code is available at `https://anonymous.4open.science/r/BioBridge-227/`.

1 INTRODUCTION

In the human perceptual system, vision is widely regarded as the dominant sensory modality (Hutmacher, 2019; Jiayu et al., 2025), serving as a primary entry point for investigating human perception, cognition, and behavior. This makes understanding the neural processing and response mechanisms underlying visual perception has long been a central focus of neuroscience research. To achieve this, a range of brain imaging techniques have been employed to study visual neural signal decoding, including electroencephalography (EEG) (Zheng & Chen, 2021; Singh et al., 2023), magnetoencephalography (MEG) (Song et al., 2023; Benchetrit et al., 2023), and functional magnetic resonance imaging (fMRI) (Liu et al., 2022; Fang et al., 2023). Among these modalities, EEG has emerged as a particularly prevalent approach in recent years, owing to its millisecond-level temporal resolution, relative ease of acquisition, and cost-effectiveness, making it especially suitable for capturing the rapid dynamics of human visual processing.

EEG visual neural signal decoding aims to establish semantic alignment between the two modalities of visual stimuli and neural signals, enabling retrieval and reconstruction of visual stimuli. ***However, most existing methods do not adequately account for the inherent modality-specific differences between EEG and visual signals, which manifest in two aspects: information structure differences and individual organizational differences.*** Specifically, structural discrepancies arise from the non-uniform nature of retinotopic mapping. As shown in the left image of Figure 1, compared to peripheral vision, a larger cortical area is dedicated to central vision Ribeiro et al. (2025). This results in a foveal blur structure that radiates outward from the fovea, causing information redundancy in the original visual stimuli, particularly in the peripheral regions, relative to neural signals. Furthermore, inter-subject variability in retinal topography and visual cortex organization introduces significant heterogeneity (Ribeiro et al., 2025). Despite a shared functional architecture, these subject-specific anatomical nuances cause identical visual stimuli to elicit divergent neural responses across indi-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

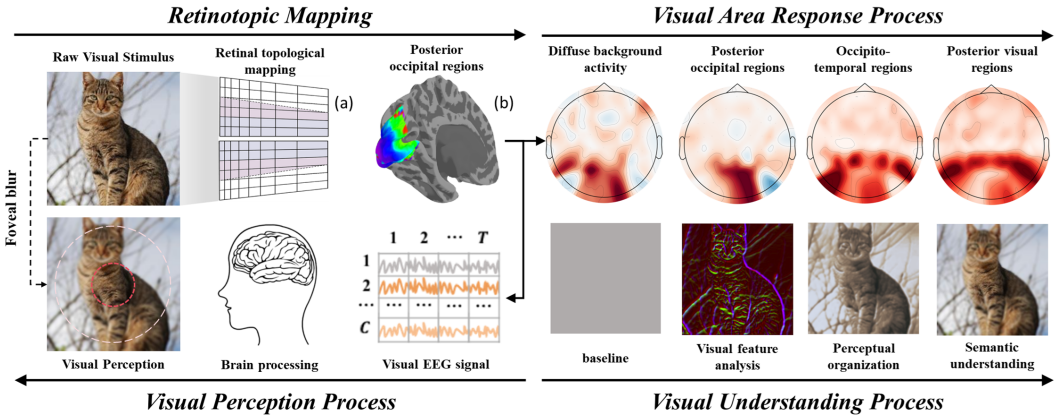

Figure 1: Schematic of visual processing and neural responses. Left: topographic mapping of visual stimuli in the retina and visual cortex, with panels (a) and (b) reproducing classic studies (Ribeiro et al., 2025). Due to greater cortical resources allocated to foveal input, perception exhibits a fovea-centered blur, differing from the original stimulus. Right: staged neural processing of visual understanding, sequentially involving background-related diffuse activity, low-level feature extraction in the posterior occipital cortex, perceptual organization in the occipitotemporal pathway, and high-level semantic representation in occipital regions, with each stage corresponding to distinct brain activations and visual representations.

viduals, thereby severely posing the challenge of cross-modal alignment. Meanwhile, due to the scarcity of paired EEG–image data, relying solely on contrastive learning strategies often struggles to fully exploit the deep semantic associations between the two modalities, thus limiting overall alignment performance. Therefore, it is necessary to introduce physiological prior knowledge during the modeling process and impose physiologically consistent structural constraints, in order to construct a visual neural decoding framework that better aligns with the true mechanisms of human visual processing.

Based on this perspective, we propose **BioBridge**, a biologically guided brain–vision alignment framework that models individual-specific neural responses by incorporating physiological priors and contrastive learning for efficient and accurate visual decoding. Unlike prior methods that focus on empirical cross-modal alignment, BioBridge is the first to jointly model retinotopic structure, cortical hierarchy, and bidirectional cross-modal alignment under a unified biologically constrained framework. Specifically, to address inherent information structure differences between visual stimuli and noisy neural signals, we introduce Adaptive Blur with Visual Priors, which leverages the fovea-to-periphery blur in human perception, incorporates a learnable functional prior, and captures individual variability in retinal topography through adaptive optimization, effectively mitigating peripheral redundancy. Inspired by the hierarchical visual cortex (right panel of Figure 1), we propose the Biomimetic Visual Feature Extraction module to extract low- and high-level visual features in two stages, guiding channel attention weights according to cortical functional divisions to better align feature learning with real human visual processing Csikor et al. (2025). Furthermore, to connect these core modules and reduce modality heterogeneity, we adopt Multi-level Bidirectional Contrastive Learning to project features from both modalities into a shared semantic space, further enhancing cross-modal interaction.

In summary, BioBridge presents a physiologically plausible and efficient visual neural decoding framework. Our main contributions are summarized as follows:

- We propose an ***adaptive blur module with visual priors***, leveraging shared features of retinal topography to incorporate physiologically grounded priors. It reduces modality discrepancies between visual stimuli and EEG signals while capturing individual-specific retinal structures, enhancing cross-modal alignment.

- We introduce a ***biomimetic visual feature extraction module***, simulating hierarchical processing of the human visual system under physiologically plausible constraints. By incorporating biologically informed channel-attention priors, it efficiently extracts low- and high-level visual features while reflecting structured cortical processing.
- We propose a ***multi-level bidirectional contrastive learning strategy***, projecting multi-level EEG and visual features into a shared semantic space to effectively alleviate modality heterogeneity and further strengthen cross-modal correspondence.
- Extensive experiments consistently demonstrate the superior performance and generalization of our proposed framework. For zero-shot brain-to-image retrieval, it achieves an **80.5%** Top-1 and **97.6%** Top-5 accuracy. Furthermore, when adopting Gaussian noise as the image processing strategy, our method achieves a Top-1 accuracy of **86.2%**.

2 Related Work

2.1 Visual Neural Decoding

Visual neural decoding (Miyawaki et al., 2008; Spampinato et al., 2017; Scotti et al., 2024; Fu et al., 2025) aims to uncover the mapping between visual stimuli and neural responses, enabling interpretable modeling and reconstruction of visual information. With its millisecond-level temporal resolution, EEG has become one of the most commonly used neural signal modalities for studying and decoding dynamic visual processing (Zheng & Chen, 2021; Singh et al., 2023).

In recent years, various methods have been proposed for EEG-based visual decoding, most of which are built upon unidirectional contrastive learning frameworks (Chen et al., 2024; Du et al., 2023; Wu et al., 2025). These approaches typically align EEG-derived representations with the embedding space of large-scale vision–language models (e.g., CLIP Radford et al. (2021)) to achieve cross-modal semantic correspondence.Although recent studies have begun to explore structural aspects of the data, such as the information density gap between EEG and images Wu et al. (2025) or hierarchical structures in visual representations Liu et al. (2025), the internal structural modeling of EEG signals remains relatively underexplored. In particular, the joint modeling of individual perceptual differences (e.g., visual perceptual characteristics) and the hierarchical organization of EEG channels still leaves room for further investigation.

2.2 Multi-modal Contrastive Learning

Multimodal contrastive learning has been a long-standing research focus and has achieved substantial progress across domains. Its core goal is to map heterogeneous data, such as images and text, into a shared latent space using large-scale paired semantics and contrastive objectives Schuhmann et al. (2022). Models such as CLIP Radford et al. (2021), BLIP Li et al. (2022), and ALIGN Jia et al. (2021) leverage massive training data and designed encoders to achieve strong zero-shot retrieval and classification performance, while maintaining generalization to downstream tasks. These successes have further fueled interest in multimodal contrastive learning.

However, specific modality combinations such as video–text Bain et al. (2021), 3D–language Xue et al. (2023), and EEG–image still face challenges in real-world scenarios. These tasks often lack large-scale, high-quality paired data and exhibit pronounced cross-modal heterogeneity. For EEG–image alignment, the complexity and noise of neural signals further hinder effective alignment, limiting model generalization, particularly in zero-shot settings. Therefore, alleviating data scarcity and improving accuracy and consistency of cross-modal alignment remain key open problems in multimodal representation learning.

2.3 Visual Prior and Robust Perception Modeling

Visual perception is not a direct mapping of the original visual stimuli but rather a robust information integration process that relies on intrinsic priors in complex and uncertain environments (Moreno-Bote et al., 2011; Bowers et al., 2023). In natural visual scenes, degradation factors such as blur, noise, and resolution variations are ubiquitous. However, the human visual system is still

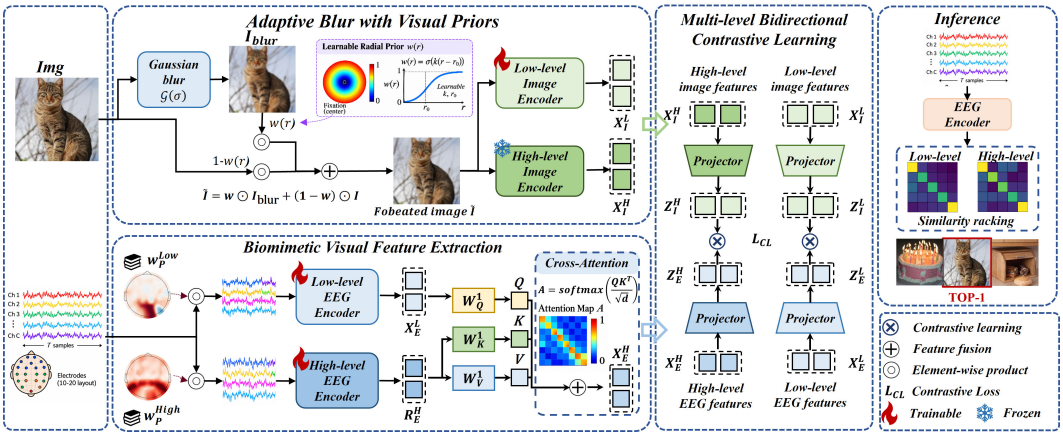

Figure 2: Overview of BioBridge. ABVP adapts visual inputs using a learnable fovea centered radial prior, while BVFE extracts hierarchical EEG representations through physiological channel priors and cross attention. MBCL aligns low and high level EEG and visual features using bidirectional contrastive objectives, enabling zero shot EEG to image retrieval.

able to maintain stable perception, indicating the crucial role of multi-scale feature integration and perceptual priors in visual processing.

Existing studies (Ribeiro et al., 2025; Jang & Tong, 2024) have shown that different forms of visual degradation selectively modulate feature representations, with visual cortical areas exhibiting distinct sensitivity to spatial frequency and clarity: low-level regions emphasize local structures, whereas high-level regions exhibit semantic invariance. This suggests that visual degradation reflects hierarchical perceptual influences rather than mere information loss. However, most EEG-based visual decoding and cross-modal alignment approaches (Wu et al., 2025; Zhang et al., 2025b) treat visual degradation primarily as a data augmentation strategy, with limited investigation into how perceptual priors shape neural representations. Effectively integrating visual priors into EEG–image alignment to address perceptual inconsistencies remains an open challenge.

3 METHOD

We propose BioBridge, a biologically guided brain–vision alignment framework composed of three main modules: Adaptive Blur with Visual Priors (ABVP), Biomimetic Visual Feature Extraction (BVFE), and Multi-level Bidirectional Contrastive Learning (MBCL). The overall architecture of the proposed framework is illustrated in the figure 2.

During training, paired EEG–image data are used to learn modality-specific representations, while contrastive learning explicitly aligns matched pairs across modalities and removes mismatched pairs in the shared embedding space. During inference, the trained encoder and mapper efficiently perform zero-shot classification and image retrieval based on EEG signals alone, demonstrating the effectiveness of the proposed framework for EEG-based visual decoding.

3.1 ADAPTIVE BLUR WITH VISUAL PRIORS

Due to the constraints of retinal topography, visual perception exhibits increasing blur from the fovea to the periphery, leading to peripheral information loss and suggesting that raw visual stimuli contain redundant data relative to brain signals. To address this structural mismatch, we propose an adaptive blur method that incorporates visual priors and learns the blurring effect in a data-driven manner, thereby enhancing the efficiency of cross-modal alignment.

Based on the characteristics of the dataset acquisition paradigm, in which participants were instructed to fixate on a red dot at the center of each image, we treat the image center as the visual center. Guided by visual priors, we construct a blur function whose parameters are learned to achieve adaptive and spatially varying blurring. Following prior studies Wu et al. (2025), Gaussian blur is adopted as the image information attenuation method. Specifically, we first apply Gaussian blurring to the original image to obtain a uniformly blurred image:

$$I_{blur}(i,j) = \sum_{m=-k}^{k} \sum_{n=-k}^{k} I(i-m, j-n) G_\sigma(m,n)$$

where $k = \lfloor r_{kernel}/2 \rfloor$, r_{kernel} represents the radius of the Gaussian kernel, I represents the original image, and I_{blur} represents the generated blurred image.

$$G_\sigma(m,n) = \frac{1}{Z} exp(-\frac{m^2+n^2}{2\sigma^2})$$

where Z denotes the normalization constant and σ the standard deviation, which controls the overall degree of blurriness. The procedure for generating blurred images is as follows:

$$\hat{I}_{visual} = wI_{blur} + (1-w)I$$

where $w(i,j) \in [0,1]$ represents the fusion weight. Physiologically, the visual resolution is the highest in the fovea region of the retina; as the eccentricity increases, beyond the radius of the fovea, the visual resolution rapidly decreases, and the corresponding perceptual blurriness increases rapidly; in the peripheral areas far from the center, this blurriness increase tends to be stable:

$$w(r) = \sigma_{act}(k(r - r_0))$$

where σ_{act} represents the activation function, k is the blur coefficient, r_0 represents the radius of the fovea, and $r(i,j)$ represents the Euclidean distance between the pixel (i,j) and the visual center, which is the origin of the fovea. In our setup, k and r_0 are learnable parameters, and the optimal blur scheme is found through contrast learning between the two modalities.

3.2 Biomimetic Visual Feature Extraction

Human visual processing is hierarchical: the retina converts light into neural signals, relayed via the lateral geniculate nucleus to early visual areas, where low-level features are extracted and progressively processed by higher cortical regions such as the inferior temporal lobe, generating high-level features for object and scene recognition. Inspired by this, we propose a progressive visual feature extraction framework incorporating channel priors.

We decouple low- and high-level visual features from EEG signals using channel attention, with weights initialized by physiological priors derived from electrode locations and cortical functions:

$$E^H = w^H E^R, E^L = w^L E^R$$

where E^H represents the high-level visual bias signal, E^L represents the low-level visual bias signal, E^R represents the original signal, w^H and w^L represent the corresponding learnable weights, and they are initialized according to physiological reality.

Then, we followed the processing flow of visual neural signals and extracted the low-level visual features Z_E^L from E^L. Using this as a guide, we employed the cross-attention mechanism to extract the high-level visual features Z_E^H from E^H. The specific process can be formulated as:

$$Z_E^L = f^L(E^L), Z_E^H = CrossAttn(Z_E^L, f^H(E^H))$$

$$CrossAttn(X,Y) = softmax(\frac{(XW_Q)(YW_K)^T}{\sqrt{d}})(YW_V)$$

Among them, W_Q, W_K, and W_V are learnable weight matrices used to map the input into "query vector", "key vector" and "value vector", and $\sqrt{d}$ is the scaling factor.

This biomimetic visual feature extraction framework is inspired by the progressive processing mechanism of the human visual system and incorporates physiological channel priors. It not only follows the hierarchical principles of biological visual cognition but also enhances the specificity and semantic consistency of EEG visual features, providing a more cognitively coherent and logical feature representation basis for subsequent EEG-based visual information recognition and understanding.

3.3 Multi-level Bidirectional Contrastive Learning

In EEG-based visual decoding, neural signals reflect hierarchical visual processing from early sensory encoding to high-level semantic abstraction. Mapping EEG representations to a single image embedding space overlooks this hierarchy and assumes direct alignment between EEG and image representations at different semantic levels.

To address this, we propose MBCL, which uses a trainable shallow ResNet and a frozen pre-trained visual model to extract low- and high-level image features corresponding to hierarchical EEG representations. These multi-level features are then projected into separate trainable latent spaces, enabling effective cross-modal alignment between EEG and image representations at different semantic levels:

$$Z_E^L = p_E^L(X_E^L), \quad Z_E^H = p_E^H(X_E^H), \quad Z_I^L = p_I^L(X_I^L), \quad Z_I^H = p_I^H(X_I^H).$$

Here, p denotes the projection network and Z the latent embedding; subscripts E and I denote EEG and image modalities, while superscripts L and H denote low- and high-level features, respectively.

We use bidirectional contrastive learning to synchronously train the EEG encoder, the shallow Resnet, and four projection networks. We optimize the entire network by bringing the matching pairs closer and pushing the non-matching pairs apart. The loss settings are as follows:

$$L_{total} = \alpha^L L(Z_I^L, Z_E^L) + \alpha^H L(Z_I^H, Z_E^H)$$

$$\mathcal{L}(Z_I, Z_E) = \frac{1}{2N} \sum_{j=1}^{N} \left(\mathcal{L}_{\mathrm{I}\to\mathrm{E}}(j) + \mathcal{L}_{\mathrm{E}\to\mathrm{I}}(j)\right)$$

$$\mathcal{L}_{\mathrm{I}\to\mathrm{E}}(j) = -\log \frac{\exp\left(\mathrm{sim}(z_{I,j}, z_{E,j})/\tau\right)}{\sum_{z_{E,j}^- \in \bar{\mathcal{Z}}_{E,j}} \exp\left(\mathrm{sim}(z_{I,j}, z_{E,j}^-)/\tau\right)}$$

$$\mathcal{L}_{\mathrm{E}\to\mathrm{I}}(j) = -\log \frac{\exp\left(\mathrm{sim}(z_{E,j}, z_{I,j})/\tau\right)}{\sum_{z_{I,j}^- \in \bar{\mathcal{Z}}_{I,j}} \exp\left(\mathrm{sim}(z_{E,j}, z_{I,j}^-)/\tau\right)}$$

where α^L and α^H represent the loss weights, $Z_{I,j} \in Z_I$ and $Z_{E,j} \in Z_E$ represent image and EEG representations, while $z_{I,j}^-$ and $Z_{E,j}^-$ are their negative sets. N is the number of data pairs, τ is the temperature coefficient, and $sim(\cdot,\cdot)$ uses cosine similarity.

Unlike traditional learning based on single and fixed image representations, we consider and simulate the hierarchical structure of visual processing and construct a learnable shared latent space, which effectively alleviates heterogeneous representation issues between the two modalities and further improves the robustness and accuracy of visual decoding performance.

4 Experiments and Results

4.1 Datasets and Implementation Details

THINGS-EEG Gifford et al. (2022) contains EEG data from 10 participants collected using the RSVP paradigm (Grootswagers et al., 2019; Intraub, 1981; Keysers et al., 2001). The training set contains 1,654 concepts with 10 images and 4 trials each, while the test set contains 200 concepts with 1 image and 80 trials each. Following prior work (Song et al., 2023; Wu et al., 2025), repeated trials are averaged, yielding 16,540 training and 200 test samples per participant.

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

Method	Subj. 1		Subj. 2		Subj. 3		Subj. 4		Subj. 5		Subj. 6		Subj. 7		Subj. 8		Subj. 9		Subj. 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
Intra-subject : train and test on one subject																						
BraVL	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE	13.2	39.5	13.5	40.3	14.5	42.7	20.6	52.7	10.1	31.5	16.5	44.0	17.0	42.1	22.9	56.1	15.4	41.6	17.4	45.8	16.1	43.6
NICE-SA	13.3	40.2	12.1	36.1	15.3	39.6	15.9	49.0	9.8	34.4	14.2	42.4	17.9	43.6	18.2	50.2	14.4	38.7	16.0	42.8	14.7	41.7
NICE-GA	15.2	40.1	13.9	40.1	14.7	42.7	17.6	48.9	9.0	29.7	16.4	44.4	14.9	43.1	20.3	52.1	14.1	39.7	19.6	46.7	15.6	42.8
ATM-S	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
Neural-MCRL	27.5	64.0	28.5	61.5	37.0	69.0	35.0	66.0	22.5	51.5	31.5	61.0	31.5	62.5	42.0	74.5	30.5	59.5	37.5	71.0	32.4	64.1
VE-SDN	32.6	63.7	34.4	69.9	38.7	73.5	39.8	72.0	29.4	58.6	34.5	68.8	34.5	68.3	49.3	79.8	39.0	69.6	39.8	75.3	37.2	69.9
CognitionCapturer	27.2	59.5	28.7	57.0	37.2	66.1	37.7	63.2	21.8	47.8	31.6	58.1	32.8	59.6	47.6	73.5	33.4	57.7	35.1	63.6	33.3	60.6
UBP	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
BioBridge(Ours)	**82.5**	**99.0**	**85.0**	**99.0**	**80.0**	**98.0**	**67.0**	**95.0**	**74.5**	**95.5**	**86.5**	**98.5**	**78.5**	**96.0**	**86.5**	**99.5**	**77.5**	**95.5**	**86.5**	**99.5**	**80.5**	**97.6**
Inter-subject : leave one subject out for test																						
BraVL	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	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	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	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	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
Neural-MCRL	13.0	31.5	12.0	30.5	14.5	**35.5**	12.5	35.5	11.5	29.0	13.5	35.5	14.0	36.0	**18.5**	38.5	13.5	32.5	17.0	39.0	14.0	34.3
UBP	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
BioBridge(Ours)	**22.5**	**62.5**	**30.0**	**57.0**	**16.5**	**35.5**	**20.5**	**43.5**	**12.0**	32.0	**16.5**	**41.5**	**21.0**	**43.5**	15.5	35.0	**17.0**	**42.0**	**26.5**	**54.5**	**19.8**	**44.7**

THINGS-MEG Hebart et al. (2023) contains 271-channel MEG data from four participants. The training set contains 1,654 concepts with 12 images presented once each, while the test set contains 200 concepts with 1 image repeated 12 times. Following prior work Song et al. (2023); Wu et al. (2025), repeated trials are averaged to improve the signal-to-noise ratio.

EEG-Encoder. We adopt EEGProject from UBP Wu et al. (2025) as the default encoder, and compare it with Shallownet Schirrmeister et al. (2017), EEGNet Lawhern et al. (2018), and TSConv Song et al. (2023) to evaluate generalization.

Image-Encoder. We use ResNet-18/34/50/101/152 He et al. (2016) to extract hierarchical visual representations and assess robustness across network depths.

Additional dataset details are provided in Appendix A.1.

4.2 Comparison with Baselines

Baseline. We compare our method with several recent and representative EEG-based visual decoding approaches, including BraVL Du et al. (2023), NICE Song et al. (2023), ATM Li et al. (2024a), Neural-MCRL Li et al. (2024b), CognitionCapturer Zhang et al. (2025a), VE-SDN Chen et al. (2024), and UBP Wu et al. (2025). Detailed descriptions and implementation details of these baseline methods are provided in Appendix A.3.

Table 2: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-MEG

Method	S1		S2		S3		S4		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
Intra-subject										
NICE	9.6	27.8	18.5	47.8	14.2	41.6	9.0	26.6	12.8	36.0
NICE-SA	9.8	27.8	18.6	46.4	10.5	38.4	11.7	27.2	12.7	35.0
NICE-GA	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.3	42.3
UBP	15.0	38.0	46.0	80.5	27.3	59.0	18.5	43.5	26.7	55.2
BioBridge	**19.0**	**42.0**	**66.5**	**88.0**	**38.0**	**72.5**	**21.0**	**50.5**	**36.1**	**69.5**
Inter-subject										
UBP	2.0	5.7	1.5	17.2	2.7	10.5	2.5	8.0	2.2	10.4
BioBridge	**4.5**	**12.5**	**6.0**	**17.0**	**7.5**	**18.0**	**3.0**	**12.5**	**5.3**	**15.0**

Comparison. Tables 1 and 2 compare our method with baselines on EEG and MEG under the same settings (see Appendix A.2). Our method consistently achieves state-of-the-art performance in both within- and across-subject settings. Specifically, it achieves Top-1/Top-5 accuracies of 80.5%/97.6% on THINGS-EEG and 36.1%/69.5% on THINGS-MEG for zero-shot brain-to-image retrieval, demonstrating robust generalization across subjects and recording modalities.

4.3 Ablation Study on Core Framework Components

To evaluate each component, ablation experiments were conducted by removing ABVP (eliminating adaptive blurring and using original visual stimuli), BVFE (removing the channel prior and cross-attention module), and MBCL (excluding trainable modules on the image side); partial experiments were omitted due to ABVP's reliance on MBCL. As shown in Table 3, all components positively affect performance. While BVFE alone causes a slight decline, its combination with

Table 3: Ablation study on the core components of BioBridge(%)

ABVP	BVFE	MBCL	Top-1	Top-5
✗	✗	✗	27.2	59.4
✗	✓	✗	25.3	54.5
✗	✗	✓	65.4	92.1
✓	✗	✓	75.9	96.6
✗	✓	✓	69.7	94.0
✓	✓	✓	**80.5**	**97.6**

other modules consistently improves results, as MBCL mitigates modality heterogeneity and enables BVFE to contribute to feature learning. Detailed ablation results and further analyses are provided in Appendix A.4.

4.4 Ablation Study on Various Encoders

To verify the generalizability of our framework, we conducted comprehensive experiments by training over one thousand models with four EEG encoders and five image encoders spanning diverse architectural variants and representation capacities. Across all settings, our framework consistently outperforms comparison methods, demonstrating stable and reliable performance gains across different encoder combinations and training configurations. The Top-1 accuracy results on the THINGS-EEG dataset are illustrated in Figure 3, while detailed quantitative results, ablation studies, and additional analyses are provided in Appendix A.5.

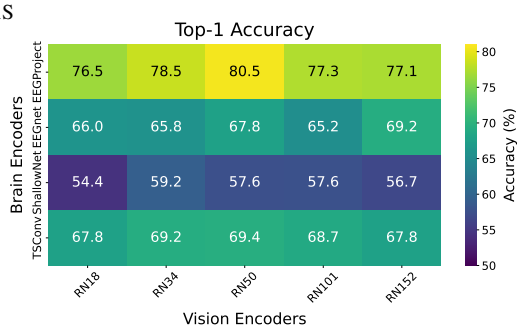

Figure 3: Top-1 accuracy (%) of BioBridge across various brain and vision encoder combinations on the THINGS-EEG.

4.5 Comparative Analysis of Different Image Processing Methods

To investigate the impact of different image processing methods and validate ABVP, we applied five techniques—color jittering, cyanscaling, Gaussian noise, low resolution, and mosaic—in addition to Gaussian blurring, under two conditions: with and without ABVP. Results are shown in Figure 4.

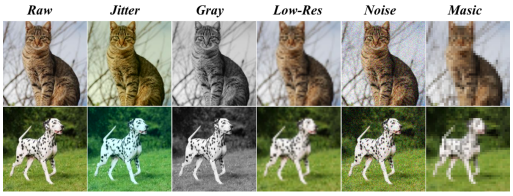

Figure 4: Visualization of representative images processed with different image processing methods used in the experiments.

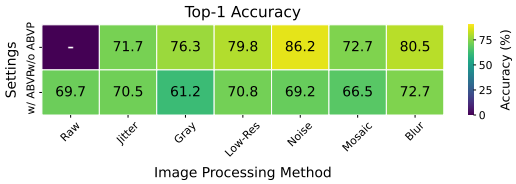

Figure 5: Top-1 accuracy (%) of BioBridge across different image processing methods on the THINGS-EEG dataset.

Figure 5 shows the Top-1 accuracy of various image processing methods on the test set. Vertically, introducing ABVP improves performance for all methods except color jitter, which alters rather than weakens image content. Notably, Gaussian noise achieves a Top-1 accuracy of 86.2%, potentially reflecting human visual perception more closely. Horizontally, without ABVP, performance fluctuates around that of the originals: global weakening may degrade central task-relevant features, while peripheral information contains redundancy, leading to varying accuracy across methods. Detailed results and analyses are provided in Appendix A.6.

We further compare different prior forms for the adaptive blurring strategy. Beyond the logistic gating function, we evaluate exponential and quadratic priors, as well as adaptive learning without explicit priors. As shown in Table 4, the logistic gating function achieves the best performance, likely due to its consistency with retinal topography. Detailed results for each prior and subject are provided in Appendix A.7. To analyze the effect of different visual prior functions, we evaluate four types of curve functions in our adaptive blur module: without prior, exponential, quadratic, and logistic. As shown in Table 4, the logistic function achieves the best performance, surpassing other curves by a clear margin. This demonstrates that the smooth and bounded characteristics of the logistic function better fit the visual prior distribution of human EEG signals.

4.6 Channel Attention Analysis

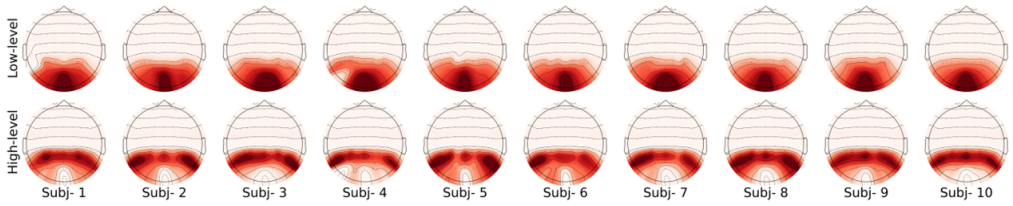

Figure 6: EEG channel attention heatmaps across subjects for low- and high-level features.

As shown in Figure 6, we present channel attention weights after learning low- and high-level visual features. The results indicate that low-level visual features are more focused on neural responses in the occipital lobe and occipito-parietal junction, while high-level visual features are more concentrated in the parietal lobe, occipito-parietal junction, and posterior temporal lobe. This aligns with existing physiological knowledge, further validating the effectiveness of our channel prior.

Table 4: Comparison of different visual priors in the ABVP(%)

Visual Prior	Top-1	Top-5
w/o	72.8	94.8
exp	77.7	96.6
quad	75.6	95.7
logistic	**80.5**	**97.6**

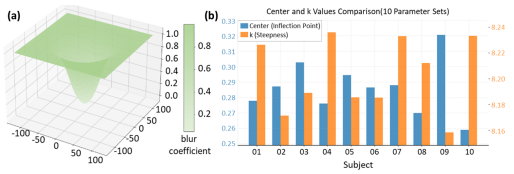

Figure 7: Visualization of adaptive blur function with visual priors: (a) Initialized blur function; (b) Learned blur function.

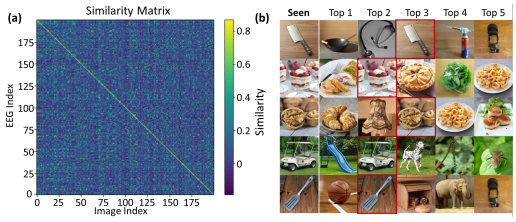

Figure 8: EEG–image alignment visualization: (a) Similarity matrix; (b) Retrieval results.

4.7 Visualization Analysis of Blur Function

To investigate individual differences in visual blur, we visualize learned blur functions. As shown in Figure 7, even under a shared central fovea prior, participants exhibit substantial variations in spatial decay distribution across visual space, reflecting heterogeneous retinal topography and cortical organization, demonstrating the model's ability to capture subject-specific visual characteristics.

4.8 Semantic Alignment Analysis and Visualization

To assess semantic alignment, we compute cross-modal similarities for all 200 test concepts of Subject 8. Figure 8(a) shows strong EEG–image alignment, while Figure 8(b) shows that the top five retrieved images for Top-1 errors remain similar to the correct samples in semantic content and low-level features. Additional results are provided in Appendix A.8 and A.9.

5 Conclusion

We propose BioBridge, a biomimetic framework for EEG-based visual neural decoding that leverages physiological visual priors to improve cross-modal alignment. By modeling foveated perception and hierarchical visual representations, it reduces perceptual redundancy and enhances semantic consistency between EEG and visual stimuli. Experiments on THINGS-EEG and THINGS-MEG show consistent gains over existing methods. However, EEG temporal dynamics are modeled generically, and the fixed central fixation assumption does not account for microsaccades, potentially limiting naturalistic fidelity and leaving room for improvement.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript language polishing, formula typesetting and figure caption refinement. We have not used generative AI tools for core model design, experimental design and result analysis, and data collection are not applicable to this work. Additionally, we used generative AI tools for preliminary literature-search assistance. We have reviewed all AI-assisted work. We manually cross-checked AI-polished text against our original arguments, and all experiment-related descriptions are verified by two authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All experiments in this paper are based on publicly released THINGS-EEG and THINGS-MEG datasets. The original data collection has completed corresponding ethical review and IRB approval in prior publications. We do not collect new human-subject data in this work. All analyses follow the original dataset usage agreement. No privacy-releasing or discriminatory processing is performed on subject data.

REPRODUCIBILITY STATEMENT

To support reproducibility, dataset preprocessing pipelines, model architecture details and full hyper-parameter settings are provided in Appendix A. Anonymous source code will be released as supplementary material upon submission. All baseline implementations and evaluation metrics follow descriptions in existing published papers.

REFERENCES

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 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*, pp. 1728–1738, 2021.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*, 2023.

Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46: e385, 2023.

Hongzhou Chen, Lianghua He, Yihang Liu, and Longzhen Yang. Visual neural decoding via improved visual-eeg semantic consistency. *arXiv preprint arXiv:2408.06788*, 2024.

Ferenc Csikor, Balázs Meszéna, Katalin Ócsai, and Gergő Orbán. Top-down perceptual inference shaping the activity of early visual cortex. *Nature Communications*, 16(1):9998, 2025.

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.

Tao Fang, Qian Zheng, and Gang Pan. Alleviating the semantic gap for generalized fmri-to-image reconstruction. *Advances in Neural Information Processing Systems*, 36:15096–15107, 2023.

Honghao Fu, Hao Wang, Jing Jih Chin, and Zhiqi Shen. Brainvis: Exploring the bridge between brain and visual signals via image reconstruction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

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.

Tijl Grootswagers, Amanda K Robinson, and Thomas A Carlson. The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188:668–679, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Fabian Hutmacher. Why is there so much more research on vision than on any other sensory modality? *Frontiers in psychology*, 10:481030, 2019.

Helene Intraub. Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 7(3):604, 1981.

Hojin Jang and Frank Tong. Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nature Communications*, 15(1):1989, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zhen Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

Li Jiayu, Zhang Qiuzhu, Li Wenjuan, Zhang Junjun, Jin Zhenlan, and Li Ling. Neural structural underlying audiovisual working memory and visual dominance under cognitive load. *Scientific Reports*, 15(1):32778, 2025.

Christian Keysers, D-K Xiao, Peter Földiák, and David I Perrett. The speed of sight. *Journal of cognitive neuroscience*, 13(1):90–101, 2001.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Yueyang Li, Zijian Kang, Shengyu Gong, Wenhao Dong, Weiming Zeng, Hongjie Yan, Wai Ting Siok, and Nizhuan Wang. Neural-mcrl: Neural multimodal contrastive representation learning for eeg-based visual decoding. *arXiv preprint arXiv:2412.17337*, 2024b.

Minxu Liu, Donghai Guan, Chuhang Zheng, Chunwei Tian, Jie Wen, and Qi Zhu. Vieeg: Hierarchical visual neural representation for eeg brain decoding. *arXiv preprint arXiv:2505.12408*, 2025.

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. Mind's eye: Grounded language model reasoning through simulation. *arXiv preprint arXiv:2210.05359*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.

Rubén Moreno-Bote, David C Knill, and Alexandre Pouget. Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108(30):12491–12496, 2011.

Marc R Nuwer, Giancarlo Comi, Ronald Emerson, Anders Fuglsang-Frederiksen, Jean-Michel Guérit, Hermann Hinrichs, Akio Ikeda, Fransisco Jose C Luccas, and Peter Rappelsburger. Ifcn standards for digital recording of clinical eeg. *Electroencephalography and clinical Neurophysiology*, 106(3):259–261, 1998.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya 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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Fernanda L Ribeiro, Noah C Benson, and Alexander M Puckett. Human retinotopic mapping: From empirical to computational models of retinotopy. *Journal of Vision*, 25(8):14–14, 2025.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

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)*, pp. 1–5. IEEE, 2023.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023.

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

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*, pp. 2246–2257, 2025.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. 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*, pp. 1179–1189, 2023.

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, pp. 14486–14493, 2025a.

Wenjiang Zhang, Sifeng Wang, Yuwei Su, Xinyu Li, Chen Zhang, and Suyu Zhong. Neurobridge: Bio-inspired self-supervised eeg-to-image decoding via cognitive priors and bidirectional semantic alignment. *arXiv preprint arXiv:2511.06836*, 2025b.

Xiao Zheng and Wanzhong Chen. An attention-based bi-lstm method for visual object classification via eeg. *Biomedical Signal Processing and Control*, 63:102174, 2021.

A APPENDIX

A.1 DATASETS DETAILS

THINGS-EEG Gifford et al. (2022) is a large-scale EEG dataset involving 10 subjects, collected using the Rapid Serial Visual Presentation (RSVP) paradigm (Grootswagers et al., 2019; Intraub, 1981; Keysers et al., 2001). The EEG data were recorded with a 64-channel EASYCAP system using the standard 10-10 electrode placement Nuwer et al. (1998). The training set consists of 1654 concepts, each with 10 images, with each image repeated 4 times (i.e., 1654 concepts × 10 images/concept × 4 trials/image). The test set includes 200 concepts, each with 1 image, repeated 80 times (i.e., 200 concepts × 1 image/concept × 80 trials/image). Data preprocessing follows the method detailed in (Song et al., 2023; Wu et al., 2025), where the raw EEG data is bandpass filtered between 0.1 Hz and 100 Hz, with a sampling rate of 1000 Hz. Out of the 63 channels, 17 [1] electrodes located in the occipital and parietal areas related to visual processing are selected for analysis. To improve the signal-to-noise ratio (SNR), EEG repetitions are averaged, resulting in 16,540 training samples and 200 test samples per subject. The EEG data are stored in float16 format to enable faster reading speeds and reduce storage requirements.

THINGS-MEG Hebart et al. (2023) involves 4 participants and is recorded with 271 channels. The experimental design includes a stimulus duration of 500 ms, followed by a blank screen lasting 1000 ± 200 ms. The training set contains 1854 concepts × 12 images × 1 repetition, and the test set includes 200 concepts × 1 image × 12 repetitions. During data processing, 200 test concepts are removed from the training set to construct a zero-shot task. The MEG data is segmented into trials ranging from 0 to 1000 ms after the stimulus onset. Preprocessing involves applying a bandpass filter between 0.1 Hz and 100 Hz, followed by baseline correction after down-sampling the data to 200 Hz. To enhance the SNR, all MEG repetitions of an image are averaged. The data are also stored in float16 format to optimize reading speed and minimize storage usage.

A.2 IMPLEMENTATION DETAILS

Environment. Our method is implemented in Python 3.10 and accelerated using PyTorch 2.0.1 with CUDA 11.8. The experimental environment is managed via Conda. Core dependencies include PyTorch, TorchVision, Transformers, Diffusers, NumPy, and SciPy. A complete list of required libraries is provided in the repository for reproducibility.All experiments are conducted on a workstation equipped with an Intel Core i7-14650HX CPU, a single NVIDIA RTX 3090 GPU with 24 GB memory, and 64 GB RAM.

Training Configuration. We train the model with a batch size of 256. Except for the EEG projector, all modules are trained for 60 epochs to achieve better convergence. A unified learning rate of 1e-4 is adopted across all experiments. Model parameters are optimized using the AdamW Loshchilov & Hutter (2017) optimizer with a weight decay of 1e-4. Early stopping is employed based on training loss and validation performance to mitigate overfitting. For the intra-subject setting, the loss coefficient of the high-level visual channel is set to 0.5. In the inter-subject setting, this coefficient is reduced to 0.1 to alleviate the impact of inter-individual variability in semantic understanding. All other hyperparameters are kept consistent across experiments.

A.3 COMPARISON METHOD

- **BraVL** Du et al. (2023) proposes a multimodal learning framework that jointly models brain signals, visual features, and linguistic representations using a mixture-of-experts mechanism. By dynamically weighting modality-specific experts, BraVL aims to enhance cross-modal alignment for EEG-based visual decoding.
- **NICE** Song et al. (2023) is a self-supervised EEG representation learning approach that leverages contrastive objectives to learn discriminative neural features. It incorporates attention mechanisms to capture spatial dependencies across EEG channels and improves generalization under limited labeled data.

[1] P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, O2

- **ATM** Li et al. (2024a) introduces an adaptive EEG encoder that integrates positional encoding with spatiotemporal modeling. The method emphasizes dynamic temporal feature extraction and spatial structure awareness to enhance EEG-based visual decoding performance.
- **Neural-MCRL** Li et al. (2024b) performs visual decoding by learning multimodal contrastive representations that align EEG signals with visual embeddings. It adopts a cross-modal contrastive learning strategy to reduce the semantic gap between neural and visual modalities.
- **CognitionCapturer** Zhang et al. (2025a) focuses on modeling cognitive processes by integrating multimodal information for EEG-based visual decoding. The method aligns neural signals with visual representations through a unified semantic space to improve decoding robustness.
- **VE-SDN** Chen et al. (2024) constructs a shared semantic space via a semantic decoupling strategy, separating visual semantics from modality-specific representations. This design facilitates more effective alignment between EEG and visual features during decoding.
- **UBP** Wu et al. (2025) introduces uncertainty-aware modeling by explicitly incorporating uncertainty priors into EEG representations. The approach enhances robustness and generalization by accounting for noise and variability inherent in neural signals.

A.4 Ablation Study of the Proposed Framework

Table 5 reports the detailed ablation results under different combinations of ABVP, BVFE, and MBCL. Removing all three components leads to severely degraded performance, indicating that effective EEG–image alignment cannot be achieved without structural or visual priors. Introducing BVFE alone does not improve performance and even results in a slight decline, suggesting that enforcing EEG-side structural constraints without addressing modality mismatch may amplify noise.

In contrast, incorporating MBCL yields a substantial performance gain, highlighting its critical role in mitigating heterogeneity between EEG signals and visual representations. When combined with MBCL, both ABVP and BVFE consistently provide additional improvements, demonstrating their complementary effects. The full model, which integrates all three components, achieves the best performance across all subjects, confirming that ABVP, BVFE, and MBCL jointly contribute to robust and effective cross-modal representation learning.

Table 5: Top-1 and Top-5 Accuracy (%) of Different ABVP/BVFE/MBCL Combinations on EEG Dataset

ABVP	BVFE	MBCL	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
✗	✗	✗	31.00 62.00	24.00 58.00	26.00 60.50	30.50 60.00	17.00 47.50	28.00 61.00	27.50 61.00	32.50 63.50	24.50 55.50	30.50 64.50	27.15 59.35
✗	✓	✗	26.00 57.00	28.50 54.50	19.50 59.00	29.50 55.00	16.00 42.00	28.00 60.50	23.50 53.00	31.00 59.00	21.50 47.50	29.50 57.50	25.30 54.50
✗	✗	✓	69.00 96.50	70.00 93.00	65.50 93.00	57.00 87.00	52.50 85.50	74.50 95.50	64.00 92.00	68.00 96.00	61.50 87.00	72.00 95.50	65.40 92.10
✓	✗	✓	78.50 98.00	79.50 98.50	72.50 96.50	65.00 94.00	65.50 93.50	83.00 98.50	74.50 95.50	81.50 99.50	73.00 92.50	85.50 99.50	75.85 96.60
✗	✓	✓	72.50 95.00	71.00 96.00	73.00 95.50	59.50 89.50	60.50 90.00	76.50 95.00	71.00 93.00	72.50 97.00	68.50 92.00	71.50 97.00	69.65 94.00
✓	✓	✓	**82.50** **99.00**	**85.00** **99.00**	**80.00** **98.00**	**67.00** **95.00**	**74.50** **99.50**	**86.50** **98.50**	**78.50** **96.00**	**86.50** **99.50**	**77.50** **95.50**	**86.50** **99.50**	**80.45** **97.95**

A.5 Results of Visual and EEG Encoder Variants

Tables 6 and 7 summarize the architectural configurations of the brain and vision encoders used in our experiments. The compared models span a wide range of parameter scales and embedding dimensions, enabling a fair and comprehensive analysis of architectural effects.

For the brain encoders, we include both lightweight CNN-based architectures (e.g., EEGNet, Shallownet, and TSConv) and a higher-capacity projection-based encoder, covering commonly adopted

designs for EEG representation learning. For the vision encoders, multiple ResNet variants with increasing depth and model capacity are evaluated, allowing us to examine the influence of visual encoder scale and embedding dimensionality on cross-modal alignment. This design ensures that the observed performance differences can be attributed to architectural characteristics rather than model size bias.

Table 6: Details of Brain Encoders

Brain Encoder	Params
EEGProjectLayer	5.40 M
EEGNet	2.34 M
Shallownet	2.56 M
TSconv	2.56 M

Table 7: Details of Vision Encoders

Vision Encoder	Params	Emb dim
ResNet18	11.69 M	512
ResNet34	21.80 M	512
ResNet50	38.32 M	2048
ResNet101	56.26 M	2048
ResNet152	60.19 M	2048

To further analyze the generalization behavior of the proposed framework, we report performance comparisons across different combinations of EEG encoders and ResNet-based image encoders with varying depths, as summarized in Tables 8–12. These experiments involve four representative EEG encoders and five visual backbones, resulting in over one thousand trained model variants.

Across all visual backbones, EEGProjectLayer consistently achieves the best Top-1 and Top-5 performance, indicating its strong ability to learn stable and discriminative EEG representations. In contrast, EEGNet, ShallowNet, and TSConv show larger performance fluctuations across subjects, especially when paired with deeper image encoders, suggesting reduced robustness to inter-subject variability.

Regarding visual backbone depth, performance improves from ResNet18 to ResNet50 but does not consistently benefit from deeper architectures such as ResNet101 and ResNet152. This indicates that increasing image encoder capacity alone is insufficient for improving EEG-based visual decoding and may even hinder optimization.Overall, these results demonstrate that the framework exhibits stable performance trends across different encoder configurations.

Table 8: Top-1 and Top-5 Accuracy (%) of Different ResNet18-based Model Architectures on EEG Dataset

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
EEGProjectLayer	Top-1	76.00	83.50	76.50	63.50	72.00	81.50	76.00	83.50	73.00	79.00	76.45
	Top-5	96.00	99.00	97.50	91.00	93.50	97.50	94.50	99.00	92.50	98.00	95.85
EEGNet	Top-1	71.50	68.50	59.00	52.00	66.50	73.00	60.00	76.00	61.50	72.00	66.00
	Top-5	91.50	96.50	90.00	84.00	93.50	96.50	88.00	95.00	89.00	95.00	91.90
Shallownet	Top-1	58.50	55.50	56.00	51.50	42.50	63.50	51.50	57.50	47.50	60.00	54.40
	Top-5	90.00	89.50	89.00	81.50	82.50	90.50	84.00	89.00	81.00	88.00	86.50
TSconv	Top-1	70.50	69.00	67.00	55.50	63.00	77.50	66.50	77.00	61.00	71.00	67.80
	Top-5	96.00	93.00	95.50	84.00	93.00	96.00	88.50	95.50	87.00	95.00	92.35

Table 9: Top-1 and Top-5 Accuracy (%) of Different ResNet34-based Model Architectures on EEG Dataset

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
EEGProjectLayer	Top-1	78.00	79.00	81.00	67.50	72.50	86.00	80.00	85.50	74.50	80.50	78.45
	Top-5	97.50	99.00	99.50	94.00	95.00	98.00	96.00	99.50	94.50	99.00	97.20
EEGNet	Top-1	75.00	66.00	61.00	51.00	58.50	76.50	58.00	76.50	62.00	73.50	65.80
	Top-5	94.50	96.00	91.50	84.50	88.00	96.50	87.00	94.50	91.50	96.00	92.00
Shallownet	Top-1	64.00	60.00	63.50	52.00	55.00	72.00	63.00	61.50	45.00	56.50	59.25
	Top-5	91.50	91.00	94.00	84.50	87.00	95.50	88.00	91.50	76.50	88.50	88.80
TSconv	Top-1	74.00	68.50	71.50	52.00	63.50	82.00	65.50	78.50	62.00	74.50	69.20
	Top-5	95.50	94.00	95.00	88.00	90.50	96.50	89.00	95.50	89.00	97.00	93.00

Table 10: Top-1 and Top-5 Accuracy (%) of Different ResNet50-based Model Architectures on EEG Dataset

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
EEGProjectLayer	Top-1	82.50	85.00	80.00	67.00	74.50	86.50	78.50	86.50	77.50	86.50	80.45
	Top-5	99.00	99.00	98.00	95.00	99.50	98.50	96.00	99.50	95.50	99.50	97.95
EEGNet	Top-1	73.50	69.00	60.00	55.00	63.00	81.50	60.50	76.00	62.00	77.50	67.80
	Top-5	94.00	94.50	92.00	87.00	95.50	98.00	90.50	95.50	91.50	96.00	93.45
Shallownet	Top-1	71.00	68.50	51.00	49.50	42.50	72.00	56.50	60.50	49.50	55.00	57.60
	Top-5	93.00	92.00	87.00	82.50	79.00	96.50	85.50	91.00	81.50	86.50	87.45
TSconv	Top-1	74.00	66.50	66.50	54.00	64.00	85.00	64.50	76.50	66.00	77.00	69.40
	Top-5	95.00	93.00	94.50	83.00	91.00	98.50	89.00	96.50	93.00	97.00	93.05

Table 11: Top-1 and Top-5 Accuracy (%) of Different ResNet101-based Model Architectures on EEG Dataset

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
EEGProjectLayer	Top-1	81.50	81.50	78.00	63.50	70.50	81.00	76.00	84.50	72.50	84.00	77.30
	Top-5	98.50	97.50	97.00	91.00	94.00	97.50	94.00	99.50	93.00	98.50	96.05
EEGNet	Top-1	71.00	69.00	57.50	50.50	57.50	78.00	61.50	75.50	60.00	71.00	65.15
	Top-5	92.00	94.00	89.00	81.50	87.00	98.50	86.00	95.50	89.00	95.50	90.80
Shallownet	Top-1	61.00	65.00	61.50	48.00	53.00	70.00	58.00	64.00	47.00	49.00	57.65
	Top-5	89.50	92.00	93.00	85.50	86.00	95.00	87.50	91.50	84.00	85.50	88.95
TSconv	Top-1	73.00	73.00	67.50	52.00	64.50	79.50	64.50	82.50	60.00	70.50	68.70
	Top-5	96.50	94.00	93.00	85.50	91.00	97.00	89.00	98.50	89.00	96.50	93.00

Table 12: Top-1 and Top-5 Accuracy (%) of Different ResNet152-based Model Architectures on EEG Dataset

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
EEGProjectLayer	Top-1	79.00	81.00	78.00	65.00	71.00	82.00	75.00	82.00	77.00	81.00	77.10
	Top-5	96.50	97.00	96.50	92.50	93.00	97.00	95.50	99.50	93.50	98.00	95.90
EEGNet	Top-1	76.50	72.50	62.00	54.00	65.00	79.00	64.50	77.50	66.00	74.50	69.15
	Top-5	94.50	93.50	89.50	84.00	92.50	97.00	90.00	96.00	92.00	97.00	92.60
Shallownet	Top-1	59.00	51.20	50.00	57.00	59.00	74.00	69.00	62.00	40.50	45.50	56.72
	Top-5	88.00	85.00	86.50	87.00	89.00	94.50	87.50	93.00	72.00	81.50	86.40
TSconv	Top-1	75.50	66.50	66.50	56.00	62.00	79.50	62.00	77.00	60.50	72.00	67.75
	Top-5	97.00	93.00	93.00	86.50	92.50	97.50	88.00	95.00	88.50	95.00	92.60

A.6 Results under Different Image Processing Methods

Tables 13 and 14 report the Top-1 and Top-5 accuracy of different image processing methods with and without ABVP, while Figure 9 qualitatively illustrates how ABVP modulates visual degradation across different processing strategies. Overall, introducing ABVP consistently improves performance across most image processing methods, demonstrating the robustness and general applicability of the proposed visual prior. Notable gains are observed for degradation-based methods such as Gaussian noise, low resolution, mosaic, cyanscale, and Gaussian blur, with Gaussian noise achieving the highest Top-1 accuracy when combined with ABVP. In contrast, color jitter shows only marginal improvement, as it primarily alters chromatic distributions rather than structural visual content, limiting the effectiveness of adaptive blurring.

Without ABVP, performance varies across image processing methods and remains close to that of the original images, suggesting that uniform global degradation may suppress task-relevant central information while reducing redundant peripheral cues. By explicitly modeling the foveated structure of human vision, ABVP adaptively reallocates visual information, as visually reflected in Figure 9 and quantitatively validated by the consistent improvements in both Top-1 and Top-5 accuracy.

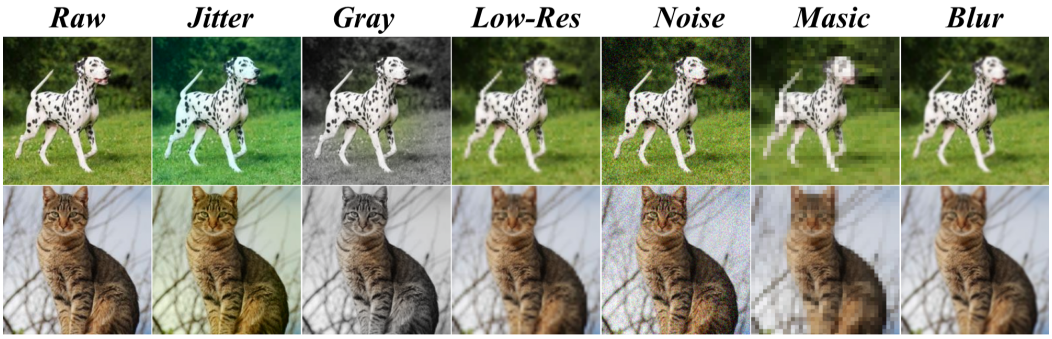

Figure 9: Effect of Incorporating ABVP across Different Image Processing Methods

Table 13: Top-1 Accuracy (%) of Different Augmentation Methods on EEG Dataset

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
Color Jitter	w/o ABVP	73.50	73.00	71.50	64.00	61.00	74.50	72.00	72.00	67.00	76.00	70.45
	w/ ABVP	75.50	73.50	76.00	62.00	63.50	73.00	71.50	76.50	69.00	76.00	**71.65**
Gaussian Noise	w/o ABVP	71.50	71.00	71.50	57.00	59.50	73.50	71.00	73.50	66.50	76.50	69.15
	w/ ABVP	88.50	87.00	84.50	77.00	82.00	90.50	88.00	92.50	80.50	91.00	**86.15**
Low Resolution	w/o ABVP	76.00	73.50	71.50	59.00	61.50	76.00	69.00	76.00	70.00	75.50	70.80
	w/ ABVP	81.00	84.50	82.50	67.00	70.00	85.50	78.50	86.00	77.50	86.00	**79.85**
Mosaic	w/o ABVP	71.00	71.00	65.00	54.50	55.00	72.50	68.00	72.00	64.50	71.50	66.50
	w/ ABVP	83.00	83.50	81.50	67.00	72.00	84.00	78.50	87.00	77.50	82.50	**79.65**
cyan	w/o ABVP	64.00	66.50	63.50	48.50	54.50	64.00	61.00	63.50	58.00	68.50	61.20
	w/ ABVP	79.00	83.00	77.00	65.50	68.50	79.50	74.00	83.00	72.00	81.50	**76.30**
Gaussian blur	w/o ABVP	77.00	74.50	76.00	59.50	63.50	79.00	73.00	75.50	72.00	77.00	72.70
	w/ ABVP	82.50	85.00	80.00	67.00	74.50	86.50	78.50	86.50	77.50	86.50	**80.45**

Table 14: Top-5 Accuracy (%) of Different Augmentation Methods on EEG Dataset

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
Color Jitter	w/o ABVP	95.50	96.00	94.00	90.00	90.00	96.00	93.50	97.00	92.50	98.00	94.25
	w/ ABVP	96.00	95.50	96.00	91.00	92.00	95.50	93.50	97.50	92.00	98.00	**94.70**
Gaussian Noise	w/o ABVP	95.50	95.50	93.00	87.50	90.00	95.00	93.00	95.50	91.00	97.00	93.30
	w/ ABVP	99.50	99.00	99.00	97.00	97.50	99.50	98.00	100.00	95.50	99.50	**98.45**
Low Resolution	w/o ABVP	98.00	97.00	95.00	90.00	90.00	95.00	63.50	97.00	93.50	97.00	91.60
	w/ ABVP	99.00	98.50	97.50	94.00	95.00	97.50	97.50	99.00	94.50	99.50	**97.20**
Mosaic	w/o ABVP	95.00	91.50	93.00	85.50	86.00	91.50	92.50	95.00	90.50	95.00	91.55
	w/ ABVP	99.00	98.00	98.90	93.00	95.00	99.00	98.50	99.50	95.50	99.50	**97.59**
cyan	w/o ABVP	90.50	92.50	90.00	84.50	84.50	91.50	90.50	91.50	90.00	92.50	89.80
	w/ ABVP	97.50	98.00	97.00	97.00	64.50	98.00	97.00	99.00	95.00	98.00	**94.10**
Gaussian blur	w/o ABVP	96.00	98.00	95.50	89.00	90.50	95.00	94.00	97.50	94.00	97.50	94.70
	w/ ABVP	99.00	99.00	98.00	95.00	99.50	98.50	96.00	99.50	95.50	99.50	**97.95**

A.7 RESULTS UNDER DIFFERENT VISUAL FUNCTION PRIORS

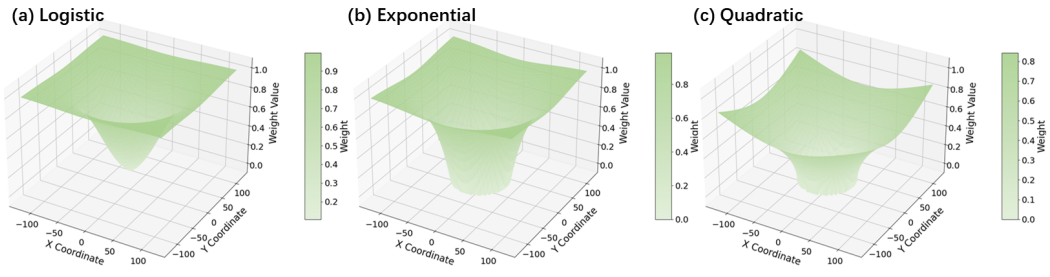

Figure 10: Visualization of Retinal Topology Fitting Function

To investigate the impact of visual function priors on EEG-image alignment, we evaluated three biologically inspired formulations—Logistic, Exponential (Exp), and Quadratic (Quad)—alongside a baseline without priors. Figure 10 qualitatively illustrates their spatial modulation across the foveated visual field. The Logistic prior (Figure 10a) emphasizes the central region with smoothly decaying peripheral contributions, the Exponential prior (Figure 10b) increases weights gradually beyond the fovea, and the Quadratic prior (Figure 10c) highlights peripheral regions with a shallower central dip.

Mathematically, the Exponential and Quadratic priors are defined as

$$w_{exp}(r) = \text{clip}\big(1 - \exp(-\lambda \cdot \max(r - r_0, 0)), 0, 1\big)$$

$$w_{quad}(r) = \text{clip}\big((\max(r - r_0, 0))^{\gamma}, 0, 1\big)$$

where r is the radial distance from the fovea, r_0 the foveal boundary, and λ, γ control the attenuation shape.

Table 15 summarizes the Top-1 and Top-5 accuracy for each subject and the overall average performance. It is evident that all prior-based methods outperform the baseline without a prior, highlighting the effectiveness of incorporating structural constraints into adaptive blurring. Among the variants, the logistic prior consistently achieves the highest accuracy across most subjects, with an average Top-1 of 80.45% and Top-5 of 97.95%. The exponential and quadratic priors also improve performance relative to the baseline but show slightly lower Top-1 and Top-5 values, likely because their attention allocation does not perfectly match the foveated structure of human retinal topography. The fully independent variant (w/o prior) exhibits competitive results but lacks a systematic central-peripheral bias, leading to a small but noticeable performance drop.

Table 15: Top-1 and Top-5 Accuracy (%) of Different Blur and Adaptive Methods on EEG Dataset

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
Logistic	Top-1	82.50	85.00	80.00	67.00	75.00	86.50	78.50	86.50	77.50	86.50	80.45
	Top-5	99.00	99.00	98.00	95.00	95.50	98.50	96.00	99.50	95.50	99.50	97.95
Exp	Top-1	80.50	80.50	76.50	67.50	70.00	86.00	77.00	82.00	74.00	83.00	77.70
	Top-5	97.50	97.50	97.50	94.00	95.00	98.00	95.00	99.00	94.50	98.50	96.65
Quad	Top-1	78.50	78.50	74.50	61.50	70.00	82.00	75.50	79.00	76.50	80.50	75.60
	Top-5	96.50	97.50	96.50	92.00	94.00	97.50	94.00	98.00	93.50	98.00	95.70
w/o Prior	Top-1	75.50	74.50	73.50	60.50	66.00	78.00	72.00	78.00	73.50	77.00	72.80
	Top-5	97.50	96.00	96.00	89.00	92.00	95.50	93.50	98.00	92.50	98.00	94.75

A.8 VISUALIZATION AND ANALYSIS OF EEG–VISUAL SIMILARITY

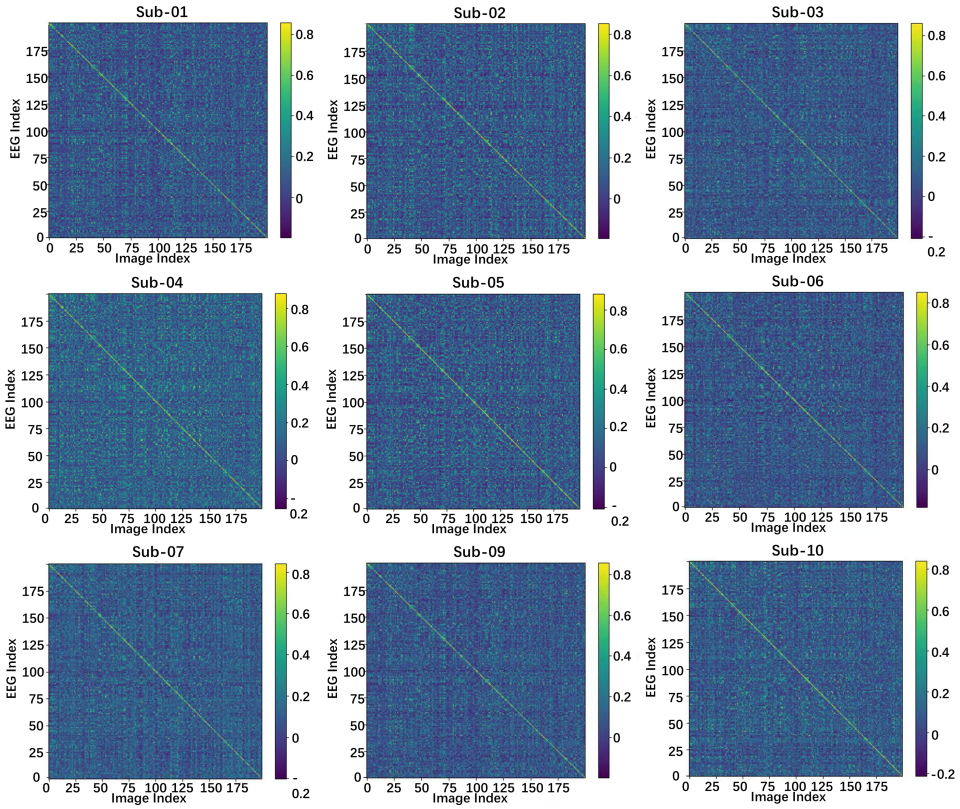

Figure 11: Similarity matrices for all subjects except Subject 8

Figure 11 illustrates the EEG–image similarity matrices for all subjects (excluding Subject 8) under the intra-subject setting. Across all subjects, a prominent diagonal-dominant structure is clearly observed, indicating a strong one-to-one correspondence between the learned EEG representations and their corresponding visual stimuli. In contrast, the off-diagonal values consistently remain at a low level and do not exhibit block-wise or column-wise patterns, suggesting that the learned representations are highly discriminative and free from systematic bias or shortcut correlations. Notably, despite the substantial inter-subject variability present in the raw EEG signals, the overall similarity structure remains consistent across subjects, demonstrating the robustness and stability of the proposed alignment framework.

It is worth noting that Subject 4 exhibits relatively elevated similarity responses in the off-diagonal regions compared to other subjects, indicating increased ambiguity in cross-concept discrimination. This phenomenon may be attributed to individual differences in neural signal quality, cognitive state, or variability in visual perception and attentional engagement during data acquisition. Such reduced separability in the learned representation space is consistent with the comparatively lower decoding accuracy observed for this subject, further supporting the correspondence between the similarity structure and quantitative performance metrics.

A.9 Visualization of Top-5 Retrieval Results for Representative Samples



Figure 12: Good Cases:Top-5 Retrieval Results for Various Stimuli

We present qualitative Top-5 retrieval results on the THINGS-EEG dataset, covering both successful and failure cases, as illustrated in Figure 12 and Figure 13, respectively.

In the successful cases, the retrieved images closely match the target stimuli at the semantic level, indicating effective EEG–visual alignment. Notably, the retrieved candidates preserve key visual characteristics of the stimuli, such as overall shape and structural configuration, rather than exact instance-level matching.

In the failure cases, although the Top-1 retrieval is incorrect, semantically related images frequently appear among the Top-5 results, suggesting that the model captures coarse semantic information from EEG signals while finer-grained distinctions remain challenging. These qualitative observations provide further evidence of the robustness of the learned cross-modal representations.

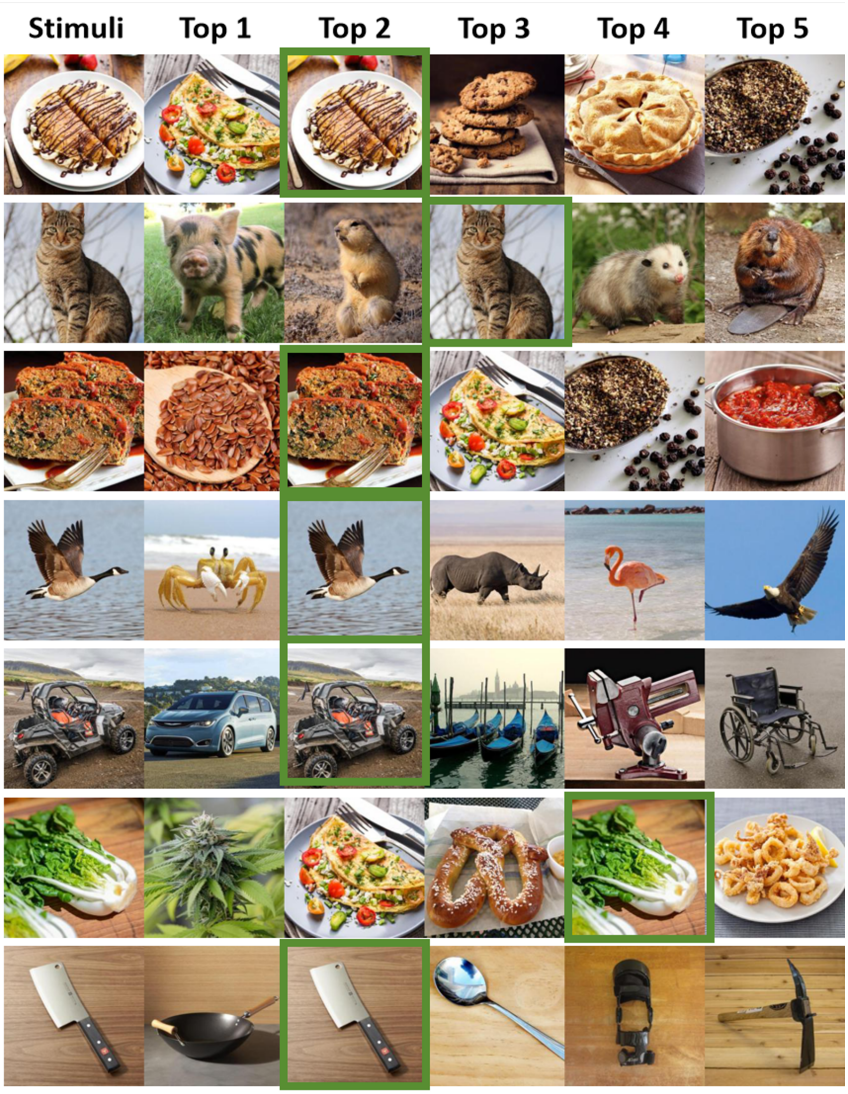

Figure 13: Bad Cases:Top-5 Retrieval Results for Various Stimuli