Neural-Visual Decoding via Cognitive-guided Adaptive Blurring and Information-Constrained Alignment

Fan Yin[1*] **Chuhang Zheng**[2*] **Peiliang Gong**[3] **Donghai Guan**[1†] **Qi Zhu**[3†]

[1]College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Nanjing, China

[2]The International Joint Institute of Tianjin University, Fuzhou,
Tianjin University, Tianjin, China

[3]College of Artificial Intelligence,
Nanjing University of Aeronautics and Astronautics, Nanjing, China

fanyin@nuaa.edu.cn chzheng@tju.edu.cn
plgong@nuaa.edu.cn dhguan@nuaa.edu.cn zhuqi@nuaa.edu.cn

Abstract

EEG-based visual decoding aims to establish a mapping between neural signals and visual semantics. However, it remains constrained by the dual challenges of severe information granularity mismatch and the low signal-to-noise ratio (SNR) of EEG signals. Existing approaches typically treat visual features as static, ignoring the dynamic selectivity of human vision and the frequency specificity of neural oscillations. To bridge this gap, we propose **CAIA**, a Cognitive-guided Adaptive blurring with Information-Constrained Alignment framework for Neural-Visual decoding. On the visual side, it simulates selective attention to adaptively reduce redundancy. Meanwhile, on the EEG side, it combines neural oscillation priors with an information-bottleneck-inspired spectral sparsity regularizer to enhance SNR. Specifically, we devise a cognitive-dynamics-based adaptive blurring mechanism that dynamically integrates center-biased and saliency-guided visual cues via cross-modal attention. Furthermore, we introduce a distribution-aware boundary calibration loss to robustly rectify alignment bias caused by outlier samples. Moreover, a cognitively guided information-screening method is proposed to select task-relevant EEG oscillations. Extensive experiments demonstrate that CAIA improves both subject-dependent and subject-independent average Top-1 and Top-5 accuracy in zero-shot brain-to-image retrieval, significantly outperforming prior methods. Our work validates that optimizing visual information density to match neural granularity offers a more interpretable and robust pathway for neural decoding. Code is available at https://github.com/mikumafu/CAIA.

1 Introduction

Understanding the brain's visual processing mechanism and decoding its neural signals is a cutting-edge research topic bridging neuroscience, cognitive science, and artificial intelligence [1, 2]. Establishing accurate stimulus-activity mappings advances brain-inspired intelligence and supports applications like Brain Computer Interfaces (BCIs) and visual prosthetics [3–5]. While deep learning has improved neural decoding, EEG's low SNR and cross-modal granularity mismatch hinder zero-shot generalization [6, 7].

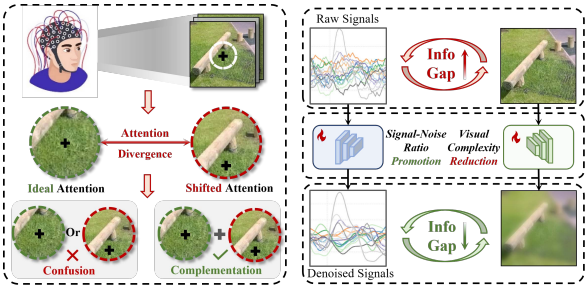
*Equal contribution.
†Corresponding author.

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

Functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), and electroencephalography (EEG) constitute the dominant neuroimaging modalities for visual neural decoding [8, 9]. While fMRI excels in spatial resolution, its temporal precision is fundamentally constrained by hemodynamic latency [10]. MEG affords millisecond-scale temporal fidelity but suffers from prohibitive costs and noise-sensitivity [11]. EEG is portable and temporally precise, ideal for real-time applications [12, 13], yet suffers from low SNR due to noise and individual variability [14]. Adaptively extracting visual-related neural features remains a core challenge.

To bridge the representational gap, some studies focus on designing more efficient EEG encoders. For instance, BrainVis [15] hierarchically fuses spatiotemporal features via temporal self-supervised encoding and frequency-domain LSTM. NICE [16] employs a spatial attention module to model inter-electrode correlations. ATM [17] combines spatiotemporal convolution with channel-wise attention to extract multi-domain EEG features. These methods essentially attempt to distill discriminative patterns from noisy data through complex encoders.

Alternatively, some approaches concentrate on visual feature optimization. Bratrix [18] extracts image background and depth maps to emphasize structural information. BraVL [6] leverages high-level semantic embeddings from vision-language models to filter pixel-level redundancy. Nevertheless, these methods predominantly optimize unimodal representations without explicitly bridging the representational gap between neural signals and visual stimuli. Recently, UBP [19] proposes blurring images based on the center bias of human visual perception to alleviate this modality gap. However, this approach remains constrained by its rigid center-bias assumption, thereby neglecting attentional shifts triggered by peripheral salient targets. Furthermore, UBP lacks a concomitant mechanism to enhance EEG signal quality, hindering genuine bidirectional optimization and compromising robustness against individual variability and paradigm heterogeneity.



(a) Attention drift. (b) Bidirectional modulation.

Figure 1: Cognitive-Guided Visual–EEG Alignment. (a) Integrating center bias and saliency guidance to counteract attention drift. (b) Bidirectional Visual–EEG information alignment.

Human visual processing is non-uniform across the visual field. By default, fixations tend toward the image center due to high-resolution foveal vision [20, 21]. However, salient peripheral targets can capture attention during dynamic viewing [22, 23]. As shown in fig. 1a, in typical EEG paradigms such as Rapid Serial Visual Presentation (RSVP), prolonged central fixation may induce attention fatigue and eye drift [24]. Consequently, a static center-blurring strategy could suppress neurally encoded peripheral information.

Moreover, EEG signals contain substantial task-irrelevant redundancy [25]. Given the distinct functional roles of neural oscillations across frequency bands [26], such redundancy may reflect a functional mismatch between certain bands and the decoding task. This motivates frequency-band screening—retaining task-relevant bands while suppressing irrelevant ones—to systematically improve SNR. The information bottleneck theory offers a framework that seeks an optimal trade-off between the compression of an input variable and the preservation of relevant information about the downstream task [27].

To address the above challenges, we propose **CAIA**, a Cognitive-guided Adaptive blurring with Information-Constrained Alignment framework for Neural-Visual decoding. As shown in fig. 1b, CAIA bridges modality gaps via bidirectional information modulation. Visually, a cognitive-dynamics-based blurring mechanism combines center-fixation and saliency-guided paths, dynamically fused via EEG-driven cross-modal attention to simulate selective attention and reduce information density. On the EEG side, a cognitively guided information-screening method combines neural oscillation priors with an information-bottleneck-inspired spectral sparsity regularizer to select task-relevant frequency bands and suppress interference, enhancing SNR. A distribution-aware boundary calibration loss further identifies and calibrates outliers based on similarity statistics, improving robustness to attention shifts and noise. Our main contributions can be summarized as follows:

- **We propose a cognitive-dynamics-based brain-inspired visual decoding approach** that utilizes center fixation and saliency blurring, dynamically fusing the two visual paths via EEG-driven cross-modal attention to simulate visual selective attention and bridge the modality gap.
- **We introduce a distribution-aware boundary calibration loss** to address outlier samples in EEG-image matching, which dynamically guides them toward the reference interval for robust cross-modal alignment.
- **We develop a cognitively guided information-screening method** that integrates neuroscientific priors to automatically select task-relevant EEG oscillatory features, suppress irrelevant bands, and enhance SNR.
- **Extensive experiments demonstrate that CAIA surpasses state-of-the-art methods**, achieving +19.4% Top-1 and +16.4% Top-5 accuracy gains over the strongest baseline UBP[19].

2 Related Works

2.1 Visual Neural Decoding

Neural decoding aims to reconstruct perceptual and cognitive content from neural signals, with significant recent progress in motor imagery [28], emotion recognition [29], disease diagnosis [30], and visual decoding [31–33, 8]. Visual decoding has garnered particular attention for its potential to elucidate the neural mechanisms underlying brain-environment interactions. While fMRI decoding benefits from high spatial resolution for detailed reconstruction [34, 35], EEG decoding capitalizes on millisecond-scale temporal precision, rendering it particularly amenable to real-time brain-computer interfaces and the analysis of dynamic cognitive processes [12, 13].

A core challenge in EEG visual decoding is establishing reliable cross-modal mappings. Early methods used handcrafted features that struggled with non-stationary trial dynamics [36]. Deep learning has since enabled specialized architectures [15–17, 37–40] that improve EEG feature encoding. However, EEG's low SNR inherently limits encoder expressivity and mapping accuracy, and these methods do not address the fundamental inter-modal information granularity mismatch. Recent work has turned to visual feature optimization, via image decomposition or blurring [18, 19, 41]—to mitigate this granularity gap. However, these approaches fail to incorporate a neural feedback mechanism that coordinates dynamic visual adjustment with EEG SNR enhancement, limiting their resilience to shifts in cognitive states and unseen trials. In contrast, CAIA introduces a bidirectional co-optimization framework that combines EEG-conditioned adaptive visual fusion with neural frequency screening, thereby achieving more robust cross-modal alignment.

2.2 Multi-Modal Contrastive Learning

Contrastive learning encourages agreement between matched representations while distinguishing unmatched samples, showing strong results in vision [42] and language [43]. Models like CLIP leverage large-scale image-text pairs for zero-shot generalization [44], inspiring its adaptation to neural decoding by aligning EEG representations with CLIP's visual space [45]. These methods typically use InfoNCE loss to pull matched pairs together and push unmatched pairs apart in a shared latent space.

Applying this paradigm directly to neural data faces two key issues. First, neural-visual pair quality is lower than natural image-text pairs. EEG contains substantial task-irrelevant noise (e.g., spontaneous oscillations, artifacts), leading to weak positive alignment and spurious correlations [46]. Second, common uniform optimization neglects variation in neural-visual pair similarities, whose distribution has been modeled using a Gaussian approximation in prior work [19]. Outliers (e.g., from attention drift or high noise) deviate from this main distribution, biasing gradients. Moreover, increasing model capacity or data augmentation may worsen noise overfitting [47], underscoring the need to integrate cognitive mechanisms—not just data-driven learning. Our CAIA addresses these challenges by coupling EEG-guided visual adaptation with distribution-aware boundary calibration, mitigating cross-modal mismatches that may arise from attentional fluctuations and neural noise to promote more robust contrastive alignment.

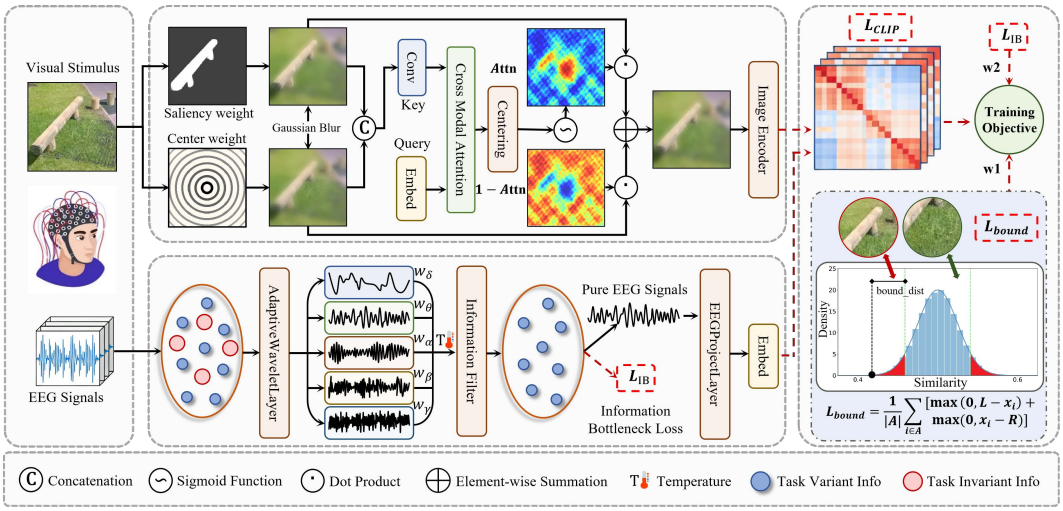


Figure 2: CAIA model framework: (a) dual-path adaptive blurring mechanism. (b) frequency-selective representation learning with IB-inspired spectral sparsity regularization. (c) distribution-aware boundary calibration loss.

3 Method

3.1 Overall Architecture

CAIA introduces a cognitively guided framework that aligns EEG and visual representations through adaptive blurring and information constraints, as illustrated in fig. 2. During training, EEG features are aligned with blurred image embeddings through contrastive learning, assisted by the boundary loss. During testing, EEG embeddings are matched to visual candidates based on similarity for retrieval evaluation.

3.2 Cognitive-Dynamics-Based Adaptive Blurring Mechanism

To mitigate information-density mismatch, we propose a dual-path adaptive blurring mechanism inspired by the complementary roles of center bias and stimulus-driven saliency in visual attention. Moving beyond static blurring strategies, this mechanism uses EEG features to adaptively balance the preservation of visual detail across image regions. It comprises two parallel blurring paths—saliency-guided differential blurring and center-guided radial blurring—which are fused via EEG-driven cross-modal attention to produce a neural-conditioned visual representation.

3.2.1 Saliency-Guided Differential Blurring

This branch leverages bottom-up visual saliency [23]. To simulate the "fine perception of salient regions, blurred perception of non-salient regions" characteristic, we apply spatially varying blur intensity inversely proportional to a saliency map. The saliency-guided blurred image X_A is generated as:

$$X_A(i,j) = w_S(i,j) \cdot X(i,j) + \big[1 - w_S(i,j)\big] \cdot \mathcal{G}_\sigma(X)(i,j) \tag{1}$$

$$w_S(i,j) = (1 - w_0) \cdot \text{saliency}(i,j) \tag{2}$$

where $w_S(i,j)$ is the saliency-based preservation weight at pixel (i,j), $w_0 \in [0,1]$ governs the overall strength of the blurring effect, $\text{saliency}(i,j)$ is obtained from a salient object detection model, and $\mathcal{G}_\sigma$ is a Gaussian blur operator.

3.2.2 Center-Guided Radial Blurring

This path incorporates the prior of visual center bias—the tendency of the human visual system to direct initial and early fixations toward the geometric center of stimuli [21]. We simulate this

characteristic using a radial blur profile that preserves clarity at the center and gradually increases blur toward the periphery [19]. The center-guided blurred image X_B is computed as:

$$X_B(i,j) = w_R(i,j) \cdot X(i,j) + \big[1 - w_R(i,j)\big] \cdot \mathcal{G}_\sigma(X)(i,j) \tag{3}$$

$$w_R(i,j) = (1 - w_0) \cdot \exp\big(-g \cdot d(i,j)\big) \tag{4}$$

where $w_R(i,j)$ is the center-bias preservation weight at pixel (i,j), $w_0 \in [0,1]$ governs the overall strength of the blurring effect, $d(i,j)$ is the normalized distance from the image center, and g controls the blur decay rate.

3.2.3 Cross-Modal Attention Fusion Strategy

The two blurring paths have complementary strengths and context-dependent applicability. To dynamically balance the "default center" and "stimulus capture" mechanisms, we introduce a cross-modal attention fusion strategy. Here, EEG features serve as the query Q, providing neural context for adaptive fusion, while features from the concatenated dual-path blurred images form the key K. This allows the model to estimate pixel-wise fusion weights A that are informed by the neural signal:

$$S = Q^\top K \in \mathbb{R}^{1 \times HW} \tag{5}$$

$$A = \mathrm{sigmoid}\big(S - \bar{S}\big) \in \mathbb{R}^{1 \times HW} \tag{6}$$

$$X_{\text{fused}} = A \odot X_A + (1 - A) \odot X_B \tag{7}$$

We adopt the sigmoid activation in lieu of softmax to compute pixel-wise fusion weights. This design avoids the sum-to-one constraint imposed by softmax, which can cause overconfidence in a single region and limit attention allocation across multiple salient areas [48]. The sigmoid function enables independent modulation per pixel, permitting simultaneous attention to several regions. To preserve relative spatial contrast amid this independence, we apply a centering operation by subtracting the mean score $\bar{S}$, thereby amplifying differential attention strength across the image and enhancing the spatial coherence of the fused output.

3.3 Cognitively Guided Information-Screening Method

This module selects task-relevant spectral information from EEG signals by leveraging neuroscientific priors on the frequency-specific functional roles of neural oscillations. Its design is inspired by the Information Bottleneck (IB) principle, which emphasizes preserving task-relevant information while limiting irrelevant input information [27].

In visual decoding, the EEG representation should preserve stimulus-related features while reducing sensitivity to irrelevant spectral activity. Cross-modal contrastive learning guides the representation toward the visual stimulus, and a regularized spectral selection stage controls how different frequency components contribute to EEG encoding.

Specifically, we first decompose EEG signals into sub-bands using an adaptive wavelet transform. The decomposition scale $k_i = k_i^0 \cdot \gamma_i$ is adjusted through a learnable parameter γ_i, enabling task-specific spectral tuning. A spectral gating layer then assigns soft selection weights $\{m_i\}$ to the sub-bands through a temperature-controlled softmax:

$$m_i = \frac{\exp(w_i/\tau)}{\sum_{j=1}^{n} \exp(w_j/\tau)} \tag{8}$$

where w_i are learnable logits, n is the number of sub-bands, and τ controls the concentration of the selection weights. The gated sub-band features are fused and passed to the EEG encoder to obtain Z.

To encourage selective spectral usage, we regularize the entropy of the band-selection distribution $M = \{m_i\}_{i=1}^n$:

$$H(M) = -\sum_{i=1}^{n} m_i \log m_i \tag{9}$$

Minimizing $H(M)$ encourages the weights to concentrate on fewer sub-bands. The scales and band weights are learned jointly with the contrastive objective, so the emphasis on each spectral component evolves with its contribution to neural-visual alignment.

3.4 Distribution-Aware Boundary Calibration Loss

Matched EEG-visual pairs exhibit varying degrees of alignment. Attentional fluctuations and neural noise may contribute to atypical similarity scores, introducing cross-modal mismatches that can interfere with contrastive learning. We therefore use the similarity distribution to construct an adaptive reference interval for boundary calibration. A Gaussian reference model provides a location-and-scale approximation for defining this interval, following prior work [19]. The observed similarity distribution is illustrated in fig. 3.

Let s_i denote the similarity score of the i-th matched pair, and let $\bar{s}$ and σ be running estimates of the mean and standard deviation, updated from training-batch similarity statistics. With $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$, where Φ is the standard normal cumulative distribution function, the reference interval is $[\bar{s} - z_{1-\alpha/2}\sigma,\ \bar{s} + z_{1-\alpha/2}\sigma]$. Matched pairs outside this interval form the set $\mathcal{M}$ and are penalized according to their distance from the corresponding boundary:

$$\mathcal{L}_{\text{bound}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left(\ell_i^{\text{left}} + \ell_i^{\text{right}} \right) \tag{10}$$

$$\ell_i^{\text{left}} = \max\left(\bar{s} - z_{1-\alpha/2}\sigma - s_i, 0\right) \tag{11}$$

$$\ell_i^{\text{right}} = \max\left(s_i - \bar{s} - z_{1-\alpha/2}\sigma, 0\right) \tag{12}$$

When $\mathcal{M}$ is empty, we define $\mathcal{L}_{\text{bound}} = 0$.

The penalty draws outlying scores toward the reference interval, moderating extreme deviations during contrastive alignment. Its gradients jointly adjust the EEG and fused visual representations, while the running statistics allow the reference boundaries to follow the evolving similarity distribution. Per-subject distributional diagnostics and comparisons with empirical percentile boundaries are provided in Appendix A.7.

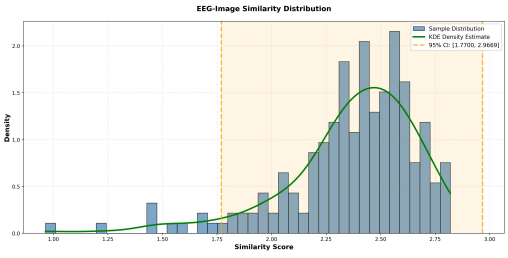


Figure 3: EEG-image similarity distribution.

Combining the components, our full training objective is:

$$\mathcal{L}_{\text{Overall}} = \mathcal{L}_{\text{CLIP}} + w_1 \mathcal{L}_{\text{bound}} + w_2 H(M) \tag{13}$$

where $\mathcal{L}_{\text{CLIP}}$ is the CLIP-based contrastive loss [44], $H(M)$ is the entropy of the band-selection distribution, and w_1 and w_2 balance boundary calibration and spectral regularization.

4 Experiment

4.1 Datasets and Implementation Details

We evaluate the proposed framework and comparison methods on two publicly available neuroimaging datasets derived from the THINGS stimulus library.

THINGS-EEG [14] records electroencephalography (EEG) from 10 subjects under an RSVP paradigm. The training set comprises 1,654 categories (10 images per category, 4 repetitions each), while the test set contains 200 novel categories (1 image per category, 80 repetitions). Following UBP [19], we enhance the signal-to-noise ratio by averaging repeated trials of the same stimulus, using a post-stimulus window of 0–250 ms and selecting 17 visually related channels spanning occipital, parietal, and lateral temporal regions.

THINGS-MEG [49] contains magnetoencephalography (MEG) recordings from 4 subjects (271 channels). The training set covers all 1,654 categories (12 images per category, single presentation), and the test set comprises 200 novel categories (1 image per category, 12 repetitions).

We adopt the ResNet-50 vision encoder from the OpenCLIP library as our visual feature extractor. All experiments were implemented in PyTorch on an NVIDIA RTX 3090 GPU. We used the Adam optimizer with a batch size of 128 and a learning rate of 1e-4 for subject-dependent settings, and a batch size of 256 with a learning rate of 1e-5 for subject-independent settings. The loss weights w_1 and w_2 in eq. (13) are set to 0.01. The blur decay rate g is set to 3.

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

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 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
Subject-dependent (intra-subject): train and test on one subject																						
BraVL [6]	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 [16]	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
ATM [17]	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	24.4	51.5	29.1	63.5	26.1	58.1
CogCap [41]	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
VE-SDN [45]	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	70.0
MB2C [50]	23.6	56.3	22.6	50.5	26.3	60.1	34.8	67.0	21.3	53.0	31.0	62.3	25.0	54.8	39.0	69.3	27.5	59.3	33.1	70.8	28.4	60.3
Neural-MCRL [51]	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
UBP [19]	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
CAIA (Ours)	**74.0**	**98.0**	**80.0**	**97.5**	**69.0**	**94.0**	51.0	**92.5**	**67.5**	**94.0**	**67.5**	**96.5**	**63.0**	**95.5**	**83.0**	**99.0**	**69.0**	**96.5**	**79.0**	**97.5**	**70.3**	**96.1**
Subject-independent (inter-subject): leave one subject out for test																						
BraVL [6]	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 [16]	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
ATM [17]	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	11.9	33.8
CogCap [41]	**16.3**	**42.3**	16.2	37.9	8.8	26.8	**15.4**	37.6	10.1	31.7	14.0	35.4	10.7	26.9	13.9	34.2	9.0	32.4	15.3	38.6	13.0	34.4
MB2C [50]	10.5	28.1	11.3	32.8	8.8	27.6	13.6	33.5	10.6	27.5	12.1	33.1	11.5	31.8	12.0	32.1	12.1	31.3	16.1	42.1	11.9	32.0
Neural-MCRL [51]	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 [19]	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
CAIA (Ours)	14.0	39.5	**21.5**	**57.5**	13.5	**37.0**	14.0	**53.5**	**12.0**	**40.5**	**16.0**	**43.0**	**16.5**	42.0	15.0	**40.5**	10.0	35.5	**31.0**	**67.5**	**16.4**	**45.7**

4.2 Overall Performance

We compare the proposed method with state-of-the-art approaches, including BraVL [6], NICE [16], ATM [17], CogCap [41], VE-SDN [45], MB2C [50], Neural-MCRL [51], and UBP [19].

As summarized in table 1 and table 2, CAIA establishes a new state of the art on both EEG and MEG test sets under subject-dependent and subject-independent protocols.

Table 2: Overall Top-1/5 accuracy (%) on THINGS-MEG for 200-way zero-shot recognition.

Method	Sub 1		Sub 2		Sub 3		Sub 4		Avg	
	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5
Subject-dependent										
MB2C [50]	9.3	33.6	20.6	49.2	18.2	44.3	10.2	33.6	14.6	40.2
NICE [16]	11.5	35.6	25.7	54.4	21.0	47.8	11.2	35.2	17.4	43.3
ATM [17]	8.0	29.3	30.2	61.5	20.3	50.5	11.8	33.3	17.6	43.7
UBP [19]	15.0	**38.0**	46.0	80.5	27.3	59.0	18.5	43.5	26.7	55.2
CAIA	**18.5**	**38.0**	**47.0**	**89.0**	**37.0**	**73.5**	**24.5**	**50.0**	**31.8**	**62.6**
Subject-independent										
UBP [19]	2.0	5.7	1.5	**17.2**	2.7	10.5	2.5	8.0	2.2	10.4
CAIA	**9.5**	**27.0**	**4.0**	15.0	**9.5**	**21.5**	**7.0**	**23.5**	**7.5**	**21.8**

On THINGS-EEG, CAIA attains subject-dependent Top-1/Top-5 accuracy of **70.3**% **/ 96.1**%, surpassing the UBP baseline by +19.4% / +16.4%. This marked improvement validates that the proposed adaptive blurring mechanism and EEG-side information screening collectively mitigate the cross-modal granularity mismatch. In the more challenging subject-independent setting, CAIA achieves **16.4**% **/ 45.7**% (UBP: 12.4% / 33.4%), indicating improved robustness to inter-subject variability.

On THINGS-MEG, CAIA achieves subject-dependent accuracies of **31.8**% **/ 62.6**% and subject-independent accuracies of **7.5**% **/ 21.8**%, outperforming UBP by +5.1% / +7.4% and +5.3% / +11.4%, respectively. These results indicate that the framework can leverage neural oscillation priors to identify stable cross-modal relationships even with weaker inter-subject signals, evidencing promising cross-modal generalization.

4.3 Ablation Experiments

We conduct ablation studies to evaluate three core components of CAIA individually and in combination: the adaptive blurring mechanism ($ATTN$), the information-screening representation learner ($IBWave$), and the boundary calibration loss (L_{bound}). Results are reported in table 3.

Disabling all modules (leaving only the EEG projection head, center-biased blurring, and the CLIP contrastive loss) yields a baseline performance of 54.5% Top-1 and 87.3% Top-5 accuracy on THINGS-EEG. Enabling $ATTN$ alone improves performance to 62.1% / 92.6%, demonstrating the efficacy of center-saliency fusion in reducing visual redundancy. Using $IBWave$ in isolation achieves 64.0% / 91.6%, indicating that frequency-band screening suppresses EEG noise and enhances the SNR. Employing L_{bound} alone leads to a slight degradation (52.4% / 85.7%). This may reflect the difficulty of imposing a similarity-range constraint on representations that still contain substantial cross-modal mismatch. By identifying atypical pair similarities and guiding them toward the reference interval, the boundary loss provides a calibration signal for cross-modal alignment. Combined with adaptive visual fusion and spectral screening, this signal guides representation refinement through adjustments to visual detail and neural spectral emphasis.

Pairwise combinations reveal synergistic effects. $ATTN + IBWave$ attains 64.9% / 93.6%, highlighting complementary benefits from visual and EEG-side optimization. $ATTN + L_{bound}$ yields 64.0% / 93.0%, supporting the complementary roles of adaptive fusion and boundary calibration. $IBWave + L_{bound}$ achieves 64.9% / 92.4%, as noise-suppressed EEG features facilitate more reliable distribution modeling.

Table 3: Ablation study on THINGS-EEG (average Top-1 / Top-5, %).

Modules			Avg	
IBWave	ATTN	L_{bound}	Top-1	Top-5
×	×	×	54.5	87.3
×	×	✓	52.4	85.7
×	✓	×	62.1	92.6
×	✓	✓	64.0	93.0
✓	×	×	64.0	91.6
✓	×	✓	64.9	92.4
✓	✓	×	64.9	93.6
✓	✓	✓	**70.3**	**96.1**

The full CAIA model integrates all three components, achieving the best overall performance of 70.3% Top-1 and 96.1% Top-5 accuracy. This optimal result demonstrates the effectiveness of the complete design loop: $IBWave$ supplies denoised EEG features, $ATTN$ generates cognitively aligned blurred images, and L_{bound} calibrates residual mismatches, collectively establishing robust cross-modal alignment.

4.4 Model Analysis

4.4.1 Cross-Modal Attention Analysis

As shown in fig. 4, the adaptive blurring mechanism produces spatially varying fusion weights, balancing center-biased and saliency-guided processing across image regions. This allows visual detail to be selectively preserved through complementary blurring pathways.

To evaluate the effectiveness of cross-modal attention, we compare the proposed mechanism with alternative image-processing strategies, including the original image, center-only blur, saliency-only blur, fixed linear blending, and a global learnable mask. Quantitative results in table 4 show that EEG-guided fusion outperforms these alternatives. By conditioning the balance between center-bias and saliency cues on neural features, the mechanism adapts visual detail to the neural context, supporting more effective cross-modal alignment.

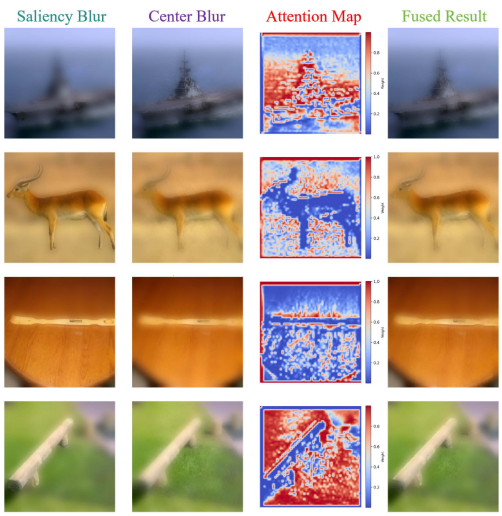


Figure 4: Visualization of the cognitive-dynamics-based adaptive blurring mechanism. Red regions indicate greater weighting of the center-biased pathway, whereas blue regions indicate greater weighting of the saliency-guided pathway.

4.4.2 Information-Guided Frequency Band Screening Analysis

After information-guided screening, the Beta band receives the highest selection weight (see Appendix A.6 for details), consistent with its reported involvement in visual attention, feature integration, and target recognition. Beta oscillations have been associated with information flow between visual and frontoparietal attention networks, providing a neurophysiological context for the learned spectral preference [52, 53].

Time-frequency visualizations illustrate the spectral reshaping induced by the IB-inspired screening module (fig. 5). The signal amplitude decreases from approximately $[-1, 1]$ to $[-0.3, 0.3]$, accompanied by selective attenuation across frequency ranges. Components above 40 Hz, which may include EMG-related interference, are markedly attenuated, while activity below 5 Hz is reduced in certain channels. These patterns are consistent with the module's denoising motivation, reducing potentially interfering spectral activity while emphasizing components relevant to visual decoding.

The EEG topography (fig. 5) shows a spatial concentration of signal intensity over the occipital-parietal region, consistent with prior reports [25]. This posterior pattern may reflect contributions

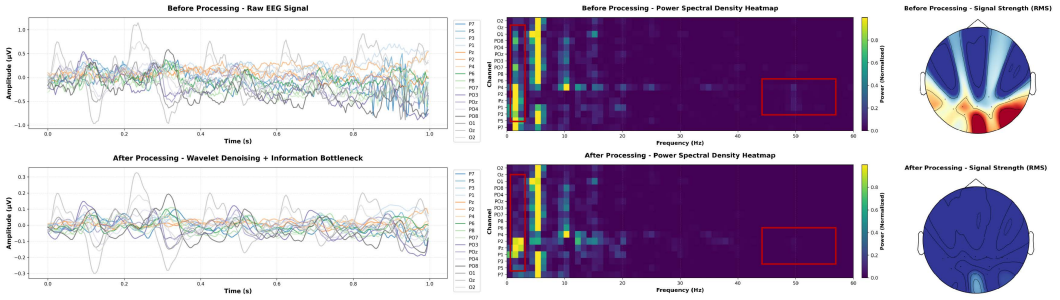
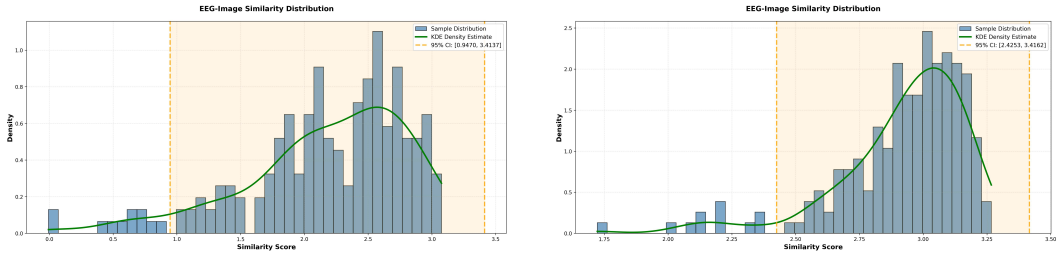


Figure 5: Visualization of EEG signals before and after information-guided frequency band screening in the time domain (left), frequency domain (middle), and EEG topographic map (right).



(a) before (b) after

Figure 6: EEG-image similarity distribution before and after calibration using the distribution-aware loss.

from visual processing in early visual areas (V1–V4) and attentional modulation in parietal regions, including those around the intraparietal sulcus, offering a neurophysiological interpretation of the spatial emphasis learned by the screening module.

4.4.3 Calibration Effect of Distribution-Aware Boundary Loss

As depicted in fig. 6, incorporating the distribution-aware boundary loss into CAIA yields a more compact EEG-image similarity distribution, with a higher mean similarity and fewer outliers beyond the reference interval. This pattern is consistent with its calibration role in guiding atypical pair similarities toward the reference interval while contrastive learning strengthens cross-modal alignment. By moderating extreme deviations that may arise from attentional fluctuations or neural noise, boundary calibration complements adaptive visual fusion and spectral screening in refining neural-visual alignment.

Table 4: Image-fusion comparison on THINGS-EEG (average Top-1 / Top-5, %).

Method	Top-1	Top-5
Original	64.6	92.1
Center_blur	<u>64.9</u>	<u>92.4</u>
Saliency_blur	59.3	88.0
Linear	63.2	91.5
Learnable_mask	64.1	92.1
Attention (Ours)	**70.3**	**96.1**

5 Conclusion

CAIA presents a cognitively guided framework that mitigates the information granularity mismatch in EEG-visual decoding via bidirectional modulation. Visually, an adaptive blurring mechanism fuses center-biased and saliency-guided cues through EEG-driven cross-modal attention. On the neural side, an IB-inspired frequency-screening module selectively enhances task-relevant frequency bands to improve SNR, while a distribution-aware boundary loss calibrates alignment outliers.

Extensive experiments on THINGS-EEG and THINGS-MEG demonstrate that CAIA consistently surpasses state-of-the-art methods in both subject-dependent and subject-independent zero-shot retrieval. Ablation studies support the complementary roles of the proposed components, while visual

analyses illustrate adaptive visual fusion and selective attenuation of potentially interfering spectral activity.

The adopted cognitive priors—center bias and saliency capture—presuppose an experimental paradigm with sustained central fixation, and their effectiveness may degrade under free-viewing conditions or paradigms without a well-defined fixation point. The current evaluation also relies on repeated-trial averaging and does not incorporate eye-tracking-based validation of fixation patterns or attentional shifts. This work offers a principled, cognition-informed approach to EEG-visual alignment, advancing both performance and interpretability. Future work may incorporate eye-tracking-based validation and extend the framework to multi-subject modeling, single-trial decoding, and real-time BCI scenarios.

Acknowledgments and Disclosure of Funding

This work was supported in part by Natural Science Foundation of China under Grant 62472220, Grant 62371234 and Grant 62076129.

References

[1] Yu Wang, Heyang Liu, Yuhao Wang, Chuan Xuan, Yixuan Hou, Sheng Feng, Hongcheng Liu, Yusheng Liao, and Yanfeng Wang. Progress, challenges and future of linguistic neural decoding with deep learning. *Communications Biology*, 8(1):1350, 2025. doi: 10.1038/s42003-025-08511-z.

[2] Eric Y. Wang, Paul G. Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A. Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, Dat Tran, Jiakun Fu, Casey M. Schneider-Mizell, Nuno Maçarico da Costa, R. Clay Reid, Forrest Collman, Katrin Franke, Alexander S. Ecker, Jacob Reimer, Xaq Pitkow, Fabian H. Sinz, Andreas S. Tolias, and MICrONS Consortium. Foundation model of neural activity predicts response to new stimulus types. *Nature*, 640(8058):470–477, April 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08829-y.

[3] Dingkun Liu, Xin Xu, Dongyang Li, Jie Li, Xinguang Yu, Zhipei Ling, and Bo Hong. Intracranial brain-computer interface spelling using localized visual motion response. *NeuroImage*, 258:119363, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119363.

[4] M. Criaud, M. Wulff, A.A. Alegria, G.J. Barker, V. Giampietro, and K. Rubia. Increased left inferior fronto-striatal activation during error monitoring after fmri neurofeedback of right inferior frontal cortex in adolescents with attention deficit hyperactivity disorder. *NeuroImage: Clinical*, 27:102311, 2020. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2020.102311.

[5] Vivien Gaillet, Annarita Cutrone, Fiorenzo Artoni, Paola Vagni, Ariastity Mega Pratiwi, Sandra Alejandra Romero, Dario Lipucci Di Paola, Silvestro Micera, and Diego Ghezzi. Spatially selective activation of the visual cortex via intraneural stimulation of the optic nerve. *Nature Biomedical Engineering*, 4(2):181–194, February 2020. ISSN 2157-846X. doi: 10.1038/s41551-019-0446-8.

[6] 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. doi: 10.1109/TPAMI.2023.3263181.

[7] Erda Zhou, Xiner Wang, Jizhi Liang, Yang Liu, Qinrong Yang, Xingchen Ran, Lei Xia, Xiang Zou, Changjiang Liu, Liuyang Sun, Lei Peng, Liang Chen, Ying Mao, Zehan Wu, Tiger H. Tao, and Zhitao Zhou. Chronically stable, high-resolution micro-electrocorticographic brain-computer interfaces for real-time motor decoding. *Advanced Science*, 12(45):e06663, December 2025. doi: 10.1002/advs.202506663.

[8] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14453–14463, 2023. doi: 10.1109/CVPR52729.2023.01389.

[9] Yohann Benchetrit, Hubert Banville, and Jean-Remi King. Brain decoding: toward real-time reconstruction of visual perception. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 7846–7858, 2024. doi: 10.48550/arXiv.2310.19812.

[10] Jingyuan Sun, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens. Contrast, attend and diffuse to decode high-resolution images from brain activities. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 12332–12348. Curran Associates, Inc., 2023. doi: 10.48550/arXiv.2305.17214.

[11] Dheeraj Rathee, Haider Raza, Sujit Roy, and Girijesh Prasad. A magnetoencephalography dataset for motor and cognitive imagery-based brain-computer interface. *Scientific Data*, 8(1):120, April 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00899-7.

[12] Kai Keng Ang and Cuntai Guan. Eeg-based strategies to detect motor imagery for control and rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(4):392–401, 2017. doi: 10.1109/TNSRE.2016.2646763.

[13] Anna Bellini, Davide Gusmeo Curti, Marco Cursi, Giordano Cecchetti, Federica Agosta, Giovanna F. Fanelli, and Massimo Filippi. Predictors of seizure detection and EEG clinical impact in an italian tertiary emergency department. *Journal of Neurology*, 271(8):5137–5145, August 2024. doi: 10.1007/s00415-024-12464-4.

[14] 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. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119754. URL `https://www.sciencedirect.com/science/article/pii/S1053811922008758`.

[15] 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)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889805.

[16] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition, 2024.

[17] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 102822–102864. Curran Associates, Inc., 2024. doi: 10.52202/079017-3266.

[18] Zehui Feng, Chenqi Zhang, Mingru Wang, Minuo Wei, Shiwei Cheng, Cuntai Guan, and Ting Han. Unveiling deep semantic uncertainty perception for language-anchored multi-modal vision-brain alignment, 2025.

[19] Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2246–2257, 2025. doi: 10.1109/CVPR52734.2025.00215.

[20] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson. Human photoreceptor topography. *The Journal of Comparative Neurology*, 292(4):497–523, February 1990. doi: 10.1002/cne.902920402.

[21] Rotem Mairon and Ohad Ben-Shahar. Stimulus center bias persists irrespective of its position on the display. *Journal of Eye Movement Research*, 18(6):77, December 2025. ISSN 1995-8692. doi: 10.3390/jemr18060077.

[22] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, March 2001. ISSN 1471-0048. doi: 10.1038/35058500.

[23] Sandeep Chand Kumain, Maheep Singh, and Lalit Kumar Awasthi. Revisited visual saliency detection with deep learning: A review of recent advancements. *ACM Comput. Surv.*, 58(6), December 2025. ISSN 0360-0300. doi: 10.1145/3773697.

[24] Aasef Shaikh. Human gaze holding and its disorders. *Annals of Otology and Neurotology*, 02:33–40, 03 2019. doi: 10.1055/s-0039-1693834.

[25] Daowen Xiong, Liangliang Hu, Jiahao Jin, Yikang Ding, Congming Tan, Jing Zhang, and Yin Tian. Interpretable cross-modal alignment network for eeg visual decoding with algorithm unrolling. *IEEE Transactions on Neural Networks and Learning Systems*, 36(11):19894–19908, 2025. doi: 10.1109/TNNLS.2025.3592646.

[26] Eric Lowet, Peter De Weerd, Mark J. Roberts, and Avgis Hadjipapas. Tuning neural synchronization: The role of variable oscillation frequencies in neural circuits. *Frontiers in Systems Neuroscience*, Volume 16 - 2022, 2022. ISSN 1662-5137. doi: 10.3389/fnsys.2022.908665.

[27] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377, 1999. URL `https://arxiv.org/abs/physics/0004057`.

[28] Xuefei Zhao, Dong Liu, Li Ma, Quan Liu, Kun Chen, Shane Xie, and Qingsong Ai. Deep cnn model based on serial-parallel structure optimization for four-class motor imagery eeg classification. *Biomedical Signal Processing and Control*, 72:103338, 2022. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2021.103338.

[29] Dongdong Li, Li Xie, Zhe Wang, and Hai Yang. Brain emotion perception inspired eeg emotion recognition with deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9): 12979–12992, 2024. doi: 10.1109/TNNLS.2023.3265730.

[30] Mário L. Vicchietti, Fernando M. Ramos, Luiz E. Betting, and Andriana S. L. O. Campanharo. Computational methods of EEG signals analysis for Alzheimer's disease classification. *Scientific Reports*, 13(1): 8184, 2023. doi: 10.1038/s41598-023-32664-8.

[31] Tao Fang, Qian Zheng, and Gang Pan. Alleviating the semantic gap for generalized fmri-to-image reconstruction. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[32] Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5899–5908, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3613832.

[33] Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119121.

[34] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. doi: 10.48550/arXiv.1907.02431.

[35] Jiangxue Li and Peize Zhao. Deep learning applications in fmri – a review work. In *Proceedings of the 2023 13th International Conference on Bioscience, Biochemistry and Bioinformatics*, ICBBB '23, page 75–80, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398190. doi: 10.1145/3586139.3586150.

[36] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.10.067.

[37] 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. doi: 10.1088/1741-2552/aace8c.

[38] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

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

[40] Dongyang Li, Haoyang Qin, Mingyang Wu, Chen Wei, and Quanying Liu. Brainflora: Uncovering brain concept representation via multimodal neural embeddings, 2025.

[41] Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information, 2024.

[42] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL `https://arxiv.org/abs/2002.05709`.

[43] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552.

[44] 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, 2021. URL `https://arxiv.org/abs/2103.00020`.

[45] Hongzhou Chen, Lianghua He, Yihang Liu, and Longzhen Yang. Visual neural decoding via improved visual-eeg semantic consistency. *ArXiv*, 2024. doi: 10.48550/arXiv.2408.06788.

[46] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Self-supervised learning for label- efficient sleep stage classification: A comprehensive evaluation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1333–1342, 2023. doi: 10.1109/TNSRE.2023.3245285.

[47] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16670–16681, June 2022.

[48] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. doi: 10.48550/arXiv.1806.01768.

[49] 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, feb 2023. ISSN 2050-084X. doi: 10.7554/eLife.82580.

[50] Yayun Wei, Lei Cao, Hao Li, and Yilin Dong. Mb2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 8992–9000, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681292.

[51] 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. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2025. doi: 10.1109/ICME59968.2025.11210130.

[52] Andreas K Engel and Pascal Fries. Beta-band oscillations—signalling the status quo? *Current Opinion in Neurobiology*, 20(2):156–165, 2010. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2010.02.015. Cognitive neuroscience.

[53] André Moraes Bastos, Julien Vezoli, Conrado Arturo Bosman, Jan-Mathijs Schoffelen, Robert Oostenveld, Jarrod Robert Dowdall, Peter De Weerd, Henry Kennedy, and Pascal Fries. Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, 85(2):390–401, 2015. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2014.12.018.

[54] Xuan-Hao Liu, Yan-Kai Liu, Tianyi Zhou, Bao-Liang Lu, and Wei-Long Zheng. Mindcross: Fast new subject adaptation with limited data for cross-subject video reconstruction from brain signals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17589–17597, Mar. 2026. doi: 10.1609/aaai.v40i21.38814.

[55] Yeongwoo Jang, Daye Jung, Seunghyun Song, Hunjun Lee, and Jangwoo Kim. Infinimind: A learning-optimized large-scale brain-computer interface. In *Proceedings of the 52nd Annual International Symposium on Computer Architecture*, ISCA '25, page 1969–1985, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712616. doi: 10.1145/3695053.3731067.

[56] Yuta Sasatake and Kojiro Matsushita. P300 erp system utilizing wireless visual stimulus presentation devices. *Sensors (Basel, Switzerland)*, 25(12):3592, 2025. ISSN 1424-8220. doi: 10.3390/s25123592.

A Supplementary Material

A.1 Details of the Datasets

THINGS-EEG [14] is a large-scale neuroimaging resource designed to facilitate zero-shot visual object recognition from neural signals. The dataset comprises electroencephalography recordings from ten healthy adult participants with normal or corrected-to-normal vision. EEG signals were acquired using a 64-channel EASYCAP arranged according to the standard 10–10 system and a BrainVision actiCHamp amplifier, sampled at 1,000 Hz with online filtering between 0.1 and 100 Hz and referenced to the Fz electrode. Visual stimulation followed a Rapid Serial Visual Presentation (RSVP) protocol. An orthogonal detection task was employed to ensure sustained attention. After excluding target trials and retaining the specified repetitions, the training and test sets contain 82,160 trials per subject in total. The stimulus set was drawn from the THINGS image database, encompassing 1,854 distinct object concepts organized into 27 higher-level categories (e.g., animals, vehicles, food items, tools, sports equipment). These concepts were partitioned into 1,654 training categories (10 images per category, 4 repetitions each) and 200 hold-out test categories (1 image per category, 80 repetitions for signal averaging).

THINGS-MEG [49] contains magnetoencephalography recordings from four healthy participants, with 271 MEG channels retained for analysis. The experimental protocol involved 500 ms stimulus presentation followed by a $1{,}000 \pm 200$ ms blank interval, with participants performing an oddball detection task. The dataset comprises 1,654 training concepts (12 images per category, single presentation) and 200 test concepts (1 image per category, 12 repetitions). MEG data were band-pass filtered (0.1–100 Hz), epoched from 0–250 ms post-stimulus, and downsampled to 200 Hz, with trial averaging applied to enhance sensor-level signal stability.

A.2 Comparison Methods

BraVL [6] develops a multimodal variational autoencoding framework based on mixture-of-product-of-experts to jointly embed brain activity, visual features, and textual descriptions into a shared latent representation. It maximizes both intra- and inter-modality mutual information to ensure consistent joint generation and improve data efficiency in zero-shot neural decoding scenarios.

NICE [16] presents a self-supervised framework that aligns image and EEG representations via contrastive learning for zero-shot object recognition. It maximizes the cosine similarity between paired image stimuli and EEG responses while pushing non-paired samples apart, enabling the EEG encoder to extract semantics consistent with visual features without categorical labels. Additionally, plug-and-play self-attention and graph attention modules are introduced to capture spatial correlations among EEG channels, supporting the analysis of spatial contributions to visual decoding.

ATM [17] proposes an end-to-end zero-shot visual reconstruction framework that projects EEG signals into a shared embedding space aligned with CLIP. It employs channel-wise Transformer encoders with temporal-spatial convolution to capture semantic and spatial correlations, facilitating tasks like classification, retrieval, and image generation.

CogCap [41] presents a unified multimodal decoding framework that incorporates text and depth alongside EEG and image modalities. It uses dedicated encoders to align EEG with each modality, capturing "beyond-image" information, and a lightweight diffusion prior to map EEG embeddings into the CLIP space for high-fidelity reconstruction.

VE-SDN [45] disentangles semantic and domain-specific features through mutual-information-based semantic decoupling and cross-modal alignment. It constructs a shared semantic space for visual and EEG representations and introduces an intra-class geometric consistency constraint to improve semantic consistency.

MB2C [50] employs a dual-GAN architecture with cycle consistency loss and contrastive learning to generate and inversely translate modality-related features. This approach narrows the modality gap between EEG and visual features, ensuring semantically aligned embeddings reside in a shared representation space.

Neural-MCRL [51] achieves multimodal alignment via semantic bridging and cross-attention mechanisms, ensuring intra-modal semantic completion and inter-modal consistency. It introduces NESTA

to adaptively capture spectral-temporal EEG patterns with subject-specific transformations for robust visual decoding.

UBP [19] proposes an Uncertainty-aware Blur Prior framework that quantifies uncertainty within paired brain-visual data to estimate the inherent mismatch. It dynamically blurs high-frequency details of visual stimuli accordingly, thereby reducing information discrepancy and enhancing alignment robustness for zero-shot neural decoding.

A.3 Statistical Robustness and Capacity Control

A.3.1 Multi-Seed Evaluation

We evaluate CAIA under five random seeds in the subject-dependent, 200-way zero-shot retrieval setting on THINGS-EEG (table 5).

Table 5: CAIA Top-1/Top-5 accuracy (%) across five random seeds on THINGS-EEG.

Seed	Metric	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Avg
0	Top-1	74.0	80.0	69.0	51.0	67.5	67.5	63.0	83.0	69.0	79.0	70.30
	Top-5	98.0	97.5	94.0	92.5	94.0	96.5	95.5	99.0	96.5	97.5	96.10
1	Top-1	72.5	78.0	66.0	48.5	64.5	65.0	61.0	80.0	67.5	76.5	67.95
	Top-5	97.5	97.0	93.5	91.5	93.0	95.5	94.5	98.0	96.0	96.5	95.30
2	Top-1	75.5	83.0	70.5	54.5	68.0	66.5	66.0	85.5	71.0	82.0	72.25
	Top-5	98.0	98.0	94.5	93.0	94.5	96.0	96.0	99.0	97.0	98.0	96.40
3	Top-1	70.0	77.0	65.0	47.0	62.5	64.5	60.0	80.0	65.5	75.0	66.65
	Top-5	97.0	96.5	93.0	91.0	92.5	95.0	94.0	98.0	95.5	96.0	94.85
4	Top-1	73.0	81.5	68.0	52.0	70.0	65.5	64.5	84.0	72.0	80.5	71.10
	Top-5	97.5	98.0	94.0	92.5	95.0	96.0	95.5	98.5	97.0	97.5	96.15

Across five seeds, CAIA achieves Top-1 and Top-5 accuracies of 69.65% $\pm$ 1.03% and 95.76% $\pm$ 0.29% (mean $\pm$ SE), with 95% confidence intervals of [66.79%, 72.51%] and [94.95%, 96.57%], respectively. These results support the stability of the reported retrieval performance across random initializations.

A.3.2 Paired Comparison with UBP

We additionally compare CAIA with UBP using seed-matched runs at seed 4. The ten subjects serve as paired observations, and we apply two-sided paired t-tests to their accuracy differences. As reported in table 6, CAIA improves mean Top-1 and Top-5 accuracy by 21.40 and 17.60 percentage points, respectively. The differences are statistically significant for both Top-1 ($t(9) = 7.13$, $p < 0.0001$) and Top-5 ($t(9) = 11.35$, $p < 0.0001$), supporting the improvement over UBP across subjects under the matched seed.

Table 6: Seed-matched Top-1/Top-5 accuracy (%) of UBP and CAIA on THINGS-EEG at seed 4.

Method	Metric	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Avg
UBP	Top-1	42.0	49.5	49.5	48.5	42.5	56.0	48.0	56.5	44.5	60.0	49.70
	Top-5	70.5	79.5	81.0	75.5	72.5	82.0	78.5	84.0	75.5	86.5	78.55
CAIA	Top-1	73.0	81.5	68.0	52.0	70.0	65.5	64.5	84.0	72.0	80.5	71.10
	Top-5	97.5	98.0	94.0	92.5	95.0	96.0	95.5	98.5	97.0	97.5	96.15

A.3.3 Capacity-Matched Comparison

To examine whether increased model capacity accounts for the performance gain, we expand the EEG projection head of UBP to approximately 68 million trainable parameters without adding the proposed CAIA components. The resulting variant achieves 51.35% Top-1 and 80.20% Top-5 accuracy (table 7), remaining substantially below CAIA's 70.30% and 96.10%. This comparison suggests that increasing encoder capacity alone does not account for the observed improvement.

Table 7: Capacity-matched UBP Top-1/Top-5 accuracy (%) on THINGS-EEG.

Metric	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Avg
Top-1	44.0	51.5	51.5	49.5	44.5	58.0	49.0	58.5	46.0	61.0	51.35
Top-5	72.5	81.5	83.0	76.5	74.5	83.5	80.0	85.5	77.5	87.5	80.20

A.4 Detailed Ablation Results

table 8 reports the per-subject results for all eight combinations of $IBWave$, $ATTN$, and L_{bound}, supplementing the ablation analysis in the main text.

Table 8: Ablation study results (%) of 200-way zero-shot retrieval on THINGS-EEG: Top-1 and Top-5.

Modules			Sub 1		Sub 2		Sub 3		Sub 4		Sub 5		Sub 6		Sub 7		Sub 8		Sub 9		Sub 10		Avg	
IBWave	ATTN	L_{bound}	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5
×	×	×	51.5	82.0	55.0	89.0	58.0	89.5	45.0	82.0	45.5	79.0	57.5	92.0	59.0	91.5	64.5	92.5	47.5	83.0	61.5	92.5	54.5	87.3
×	×	✓	51.5	85.0	43.5	80.5	49.5	84.5	48.0	83.0	42.0	81.0	56.5	90.5	59.5	87.5	62.5	93.0	42.0	80.0	68.5	91.5	52.4	85.7
×	✓	×	60.5	91.0	66.0	94.5	61.5	93.5	54.0	90.5	52.0	84.0	68.5	95.5	60.5	93.0	67.5	95.5	58.5	91.5	71.5	96.5	62.1	92.6
×	✓	✓	66.5	92.5	70.0	94.5	66.5	**95.0**	**54.5**	89.5	52.0	87.0	69.5	95.5	61.5	91.0	69.5	95.5	59.0	92.0	70.5	97.0	64.0	93.0
✓	×	×	66.5	93.5	63.0	93.5	61.0	90.5	49.0	84.5	59.5	86.5	68.0	95.0	**65.5**	92.5	75.5	98.0	61.0	88.5	70.5	93.5	64.0	91.6
✓	×	✓	68.0	95.5	69.0	93.5	62.0	90.5	51.0	85.0	62.0	89.0	69.5	95.5	61.5	91.5	74.5	97.5	60.0	90.5	71.0	95.0	64.9	92.4
✓	✓	×	66.5	94.5	68.5	95.0	66.5	94.0	53.5	89.0	57.0	90.0	**70.0**	95.5	61.0	92.0	73.5	98.5	60.0	92.5	72.5	95.0	64.9	93.6
✓	✓	✓	**74.0**	**98.0**	**80.0**	**97.5**	**69.0**	94.0	51.0	**92.5**	**67.5**	**94.0**	67.5	**96.5**	63.0	**95.5**	**83.0**	**99.0**	**69.0**	**96.5**	**79.0**	**97.5**	**70.3**	**96.1**

A.5 Adaptive Visual Fusion Analysis

A.5.1 Image-Fusion Strategy Comparison

table 9 reports the per-subject comparison of image-fusion strategies. EEG-guided attention achieves the highest mean Top-1 and Top-5 accuracy, outperforming individual blurring pathways, fixed linear blending, and a global learnable mask. These results support adapting the balance between center-biased and saliency-guided processing to the neural context.

Table 9: Image fusion method comparison results (%) of 200-way zero-shot retrieval on THINGS-EEG: Top-1 and Top-5 (intra-subject).

	Sub 1		Sub 2		Sub 3		Sub 4		Sub 5		Sub 6		Sub 7		Sub 8		Sub 9		Sub 10		Avg	
Method	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5
Original	67.0	91.5	64.0	94.0	**70.0**	94.0	45.0	84.0	66.5	88.0	72.0	95.0	59.5	90.5	73.0	95.0	63.0	93.0	66.0	96.0	64.6	92.1
Center_blur	68.0	95.5	69.0	93.5	62.0	90.5	**51.0**	85.0	62.0	89.0	69.5	95.5	61.5	91.5	74.5	97.5	60.0	90.5	71.0	95.0	64.9	92.4
Saliency_blur	60.0	88.5	54.5	88.0	62.5	92.5	47.5	80.5	51.0	83.5	70.0	90.0	60.5	90.0	67.5	95.0	50.5	80.0	68.5	92.0	59.3	88.0
Linear	66.0	91.0	62.0	93.5	66.0	93.5	47.5	81.5	60.5	84.5	72.5	95.5	**65.5**	**95.5**	68.5	96.5	60.0	91.0	63.0	92.5	63.2	91.5
Learnable_mask	66.0	93.0	62.0	94.0	**70.0**	**95.0**	46.0	82.0	61.5	87.0	**74.5**	95.0	65.0	95.0	68.0	94.0	61.5	92.0	66.5	93.5	64.1	92.1
Attention (Ours)	**74.0**	**98.0**	**80.0**	**97.5**	69.0	94.0	**51.0**	**92.5**	**67.5**	**94.0**	67.5	**96.5**	63.0	**95.5**	**83.0**	**99.0**	**69.0**	**96.5**	**79.0**	**97.5**	**70.3**	**96.1**

A.5.2 Saliency Detection Model Comparison

We evaluate three saliency detection backends in the saliency-guided blurring branch (table 10). U^2Net is a deep salient-object detector pretrained on generic saliency benchmarks. Spectral Residual and Fine-Grained are provided by OpenCV and derive saliency from spectral irregularities and local contrast, respectively.

Fine-Grained achieves the highest mean accuracy (70.3% Top-1, 96.1% Top-5). Its contrast-based saliency may provide a useful balance between local visual detail and broader image structure, guiding adaptive blurring to preserve distinctive features while reducing background detail. The lower performance of U^2Net may reflect a mismatch between its generic salient-object predictions and the visual detail useful for EEG-image alignment. Spectral Residual may place greater emphasis on isolated spectral irregularities, yielding preservation priorities that are less suitable for this decoding task.

A.5.3 EEG-Shuffling Analysis

To examine the role of EEG conditioning in adaptive fusion, we replace the EEG input to the cross-modal attention module with that of a randomly selected different sample within the batch, while all other components retain their original inputs. This perturbation reduces mean Top-1 and Top-5

Table 10: Comparison of different saliency detection models within CAIA on THINGS-EEG (200-way zero-shot retrieval, subject-dependent, Top-1/Top-5, %).

	Sub 1		Sub 2		Sub 3		Sub 4		Sub 5		Sub 6		Sub 7		Sub 8		Sub 9		Sub 10		Avg	
Method	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5
U^2Net	67.5	95.5	75.0	96.5	**73.0**	**96.5**	**55.5**	91.0	63.0	92.0	**75.0**	**96.5**	62.0	94.5	80.5	97.5	63.0	93.0	77.5	97.0	69.2	95.0
Spectral Residual	62.0	91.0	71.5	96.0	71.0	96.0	47.5	88.0	66.5	**95.0**	71.0	96.5	62.0	94.0	78.0	97.5	62.0	94.0	74.5	96.5	66.6	94.5
Fine-Grained	**74.0**	**98.0**	**80.0**	**97.5**	69.0	94.0	51.0	**92.5**	**67.5**	94.0	67.5	**96.5**	**63.0**	**95.5**	**83.0**	**99.0**	**69.0**	**96.5**	**79.0**	**97.5**	**70.3**	**96.1**

accuracy from 70.3% / 96.1% to 64.5% / 92.0% (table 11). The spatial correlation between attention maps obtained with paired and shuffled EEG is 0.0057.

As shown in fig. 7, paired EEG produces spatially structured fusion patterns, whereas shuffled EEG and random noise yield more fragmented maps. These observations support the contribution of sample-specific neural context to adaptive visual fusion.

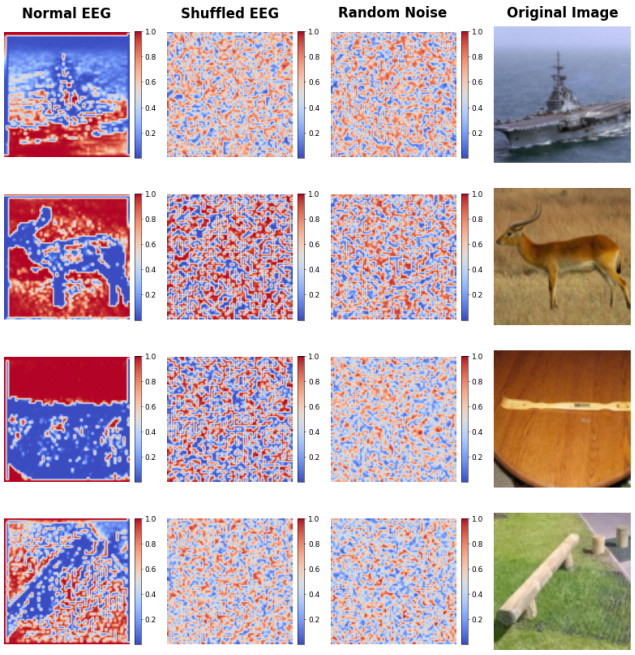


Figure 7: Attention maps with paired EEG, shuffled EEG, and random noise, alongside the corresponding stimulus images.

Table 11: Top-1/Top-5 accuracy (%) with shuffled EEG inputs to the attention module on THINGS-EEG.

Metric	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Avg
Top-1	66.5	68.0	63.5	50.0	63.0	68.0	60.0	75.0	60.5	70.0	64.5
Top-5	94.5	94.0	91.0	84.0	88.0	94.5	92.0	97.0	91.0	94.0	92.0

A.5.4 Blur Strength Sensitivity

Moderate blurring ($w_0 = 0.5$) achieves the highest mean accuracy, while lighter ($w_0 = 0.3$) and heavier ($w_0 = 0.7$) blurring yield lower performance (table 12). This pattern supports balancing visual information reduction with semantic preservation.

Table 12: Sensitivity analysis of the blur strength parameter w_0 on THINGS-EEG (200-way zero-shot retrieval, subject-dependent, Top-1/Top-5, %).

	Sub 1		Sub 2		Sub 3		Sub 4		Sub 5		Sub 6		Sub 7		Sub 8		Sub 9		Sub 10		Avg	
w_0	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5
Light ($w_0 = 0.3$)	68.0	93.0	77.0	94.0	**82.0**	**99.0**	50.0	**97.0**	60.5	92.5	**80.5**	**97.5**	58.0	93.0	**85.0**	98.5	57.5	94.5	76.0	**97.5**	69.5	95.7
Heavy ($w_0 = 0.7$)	68.0	95.0	77.0	97.0	69.0	92.5	**58.0**	95.0	59.0	92.0	70.5	95.5	57.0	92.0	74.5	95.5	58.5	92.0	77.0	94.5	66.9	94.1
Standard ($w_0 = 0.5$)	**74.0**	**98.0**	**80.0**	**97.5**	69.0	94.0	51.0	92.5	**67.5**	**94.0**	67.5	96.5	**63.0**	**95.5**	83.0	**99.0**	**69.0**	**96.5**	**79.0**	**97.5**	**70.3**	**96.1**

A.6 Frequency Screening Analysis

A.6.1 Frequency Filter Comparison

We compare IBWave against three static band-pass filters (table 13): a high-frequency filter (> 40 Hz), a low-frequency filter (< 8 Hz), and a mid-frequency filter (8–40 Hz).

Among the static filters, the mid-frequency filter achieves the highest mean accuracy (58.1% Top-1, 90.1% Top-5), consistent with the Beta-dominant weighting learned by our model and its reported involvement in visual attention. However, its accuracy remains substantially below that of IBWave (70.3% Top-1, 96.1% Top-5). Static filters rely on fixed frequency boundaries, whereas IBWave jointly optimizes decomposition scales and band-selection weights through the decoding objective. This allows the spectral emphasis to adapt to the information useful for neural-visual alignment. These results support the benefit of learnable, task-adaptive frequency screening over fixed band-pass filtering.

Table 13: Comparison of IBWave with static band-pass filters on THINGS-EEG (200-way zero-shot retrieval, subject-dependent, Top-1/Top-5, %).

	Sub 1		Sub 2		Sub 3		Sub 4		Sub 5		Sub 6		Sub 7		Sub 8		Sub 9		Sub 10		Avg	
Method	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5
High-Frequency	22.0	46.5	24.5	63.0	11.5	43.0	18.5	48.5	9.0	29.5	20.5	53.0	23.0	57.0	24.5	58.5	23.0	56.5	27.5	58.0	20.4	51.4
Low-Frequency	43.5	86.5	38.5	78.5	48.0	88.5	39.0	82.5	44.5	84.0	44.5	87.5	52.0	87.0	61.5	91.5	40.0	84.0	53.5	95.5	46.5	86.6
Mid-Frequency	62.0	93.5	63.5	93.0	65.5	**94.0**	**52.5**	86.5	59.0	86.5	66.0	94.5	44.5	84.0	51.5	89.5	45.5	84.0	70.5	95.0	58.1	90.1
IBWave (Ours)	**74.0**	**98.0**	**80.0**	**97.5**	**69.0**	**94.0**	51.0	**92.5**	**67.5**	**94.0**	**67.5**	**96.5**	**63.0**	**95.5**	**83.0**	**99.0**	**69.0**	**96.5**	**79.0**	**97.5**	**70.3**	**96.1**

A.6.2 Entropy Regularization Ablation

To examine the contribution of entropy regularization, we retain the learnable decomposition scales and band-selection weights while removing the entropy term. This variant achieves 67.8% Top-1 and 94.2% Top-5 accuracy, compared with 70.3% and 96.1% for the full model (table 14). The decrease supports the role of entropy regularization in encouraging selective spectral weighting for visual decoding.

Table 14: Top-1/Top-5 accuracy (%) without entropy regularization on THINGS-EEG.

Metric	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Avg
Top-1	71.0	75.0	68.5	56.5	60.5	68.0	63.0	74.5	65.5	75.0	67.8
Top-5	95.0	96.0	94.0	91.5	90.5	95.0	92.5	96.5	94.0	97.0	94.2

A.6.3 Spectral Preference and Weight Stability

The learned band-selection weights exhibit a preference for the Beta band (figs. 8 and 9), consistent with its reported involvement in visual processing and attention [52, 53]. Across seeds, within-subject weight variation remains no greater than 0.005 for all bands, while between-subject differences are larger, reaching 0.038 for the Beta band. This pattern suggests a reproducible spectral preference alongside subject-specific adaptation. The observed weight ranges are reported in table 15.

A.6.4 Subject-Independent Screening Analysis

Under the subject-independent setting, spectral selection places greater emphasis on the Beta band (fig. 9). The time-domain amplitude range changes from approximately $[-1, 1]$ to $[-0.6, 0.8]$, while components above 50 Hz are attenuated and activity in the 20–40 Hz range becomes more prominent

Table 15: Ranges of learned band-selection weights across subjects and seeds.

Band	Weight Range
Delta (δ)	0.126–0.152
Theta (θ)	0.133–0.161
Alpha (α)	0.218–0.248
Beta (β)	0.302–0.340
Gamma (γ)	0.144–0.163

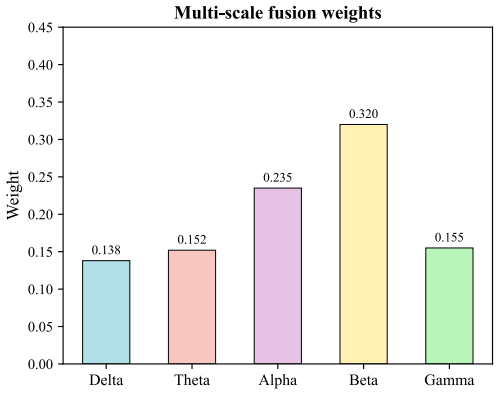


Figure 8: Band-selection weights in the subject-dependent setting.

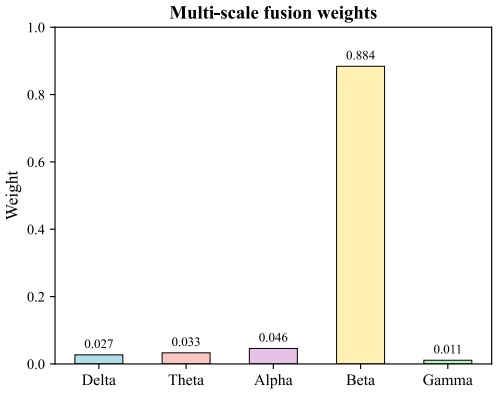


Figure 9: Band-selection weights in the subject-independent setting.

(fig. 10). The posterior topographic pattern extends toward the right-parietal region, which may reflect contributions from attentional modulation and visual information integration. Compared with the subject-dependent visualization, the attenuation is less pronounced, suggesting that screening across subjects must accommodate more heterogeneous neural responses.

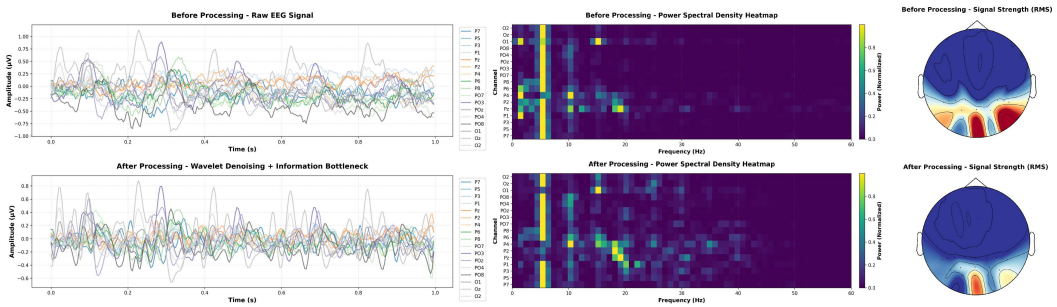


Figure 10: EEG signals before and after frequency screening in the subject-independent setting, shown in the time domain (left), frequency domain (middle), and EEG topographic map (right).

A.7 Boundary Calibration Analysis

A.7.1 Distributional Diagnostics and Reference Boundaries

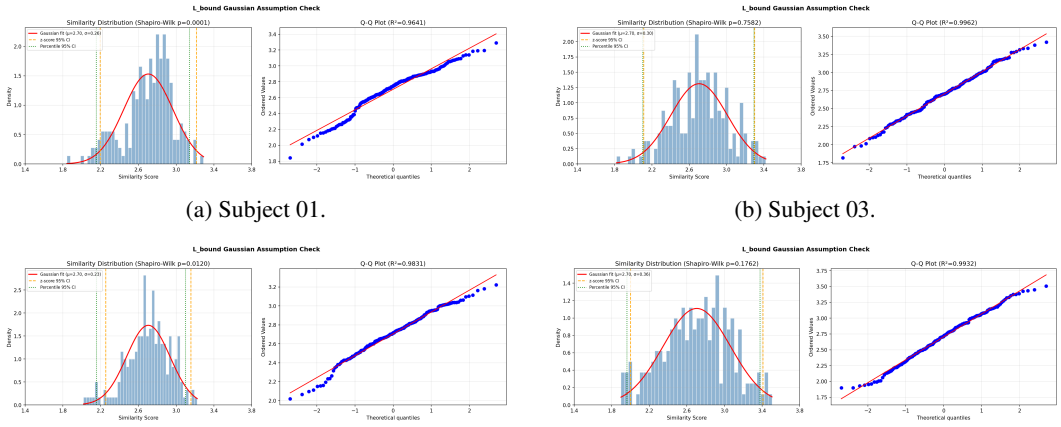
We examine matched-pair similarity distributions separately for all ten subjects using Shapiro–Wilk tests and normal Q–Q plots (table 16). Normality is not rejected for subjects 03, 06, and 09 at the 5% significance level, whereas the remaining subjects exhibit departures from a strict Gaussian distribution. Nevertheless, the squared Q–Q correlations exceed 0.96 for all subjects, supporting the Gaussian model as a reference approximation for boundary construction.

fig. 11 shows representative distributions for subjects 01, 03, 07, and 09. Subjects 01 and 07 exhibit more pronounced deviations from the normal reference line, while subjects 03 and 09 follow it more

Table 16: Per-subject distributional diagnostics on THINGS-EEG.

Metric	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10
Shapiro–Wilk p	0.0001	0.0007	0.7582	0.0054	0.0229	0.2535	0.0120	0.0010	0.1762	0.0003
Q–Q R^2	0.964	0.977	0.996	0.983	0.985	0.992	0.983	0.978	0.993	0.972

closely. The Gaussian-reference and empirical percentile boundaries remain visually close in these examples, suggesting that moderate distributional deviations need not produce markedly different reference intervals.



(a) Subject 01. (b) Subject 03.

(c) Subject 07. (d) Subject 09.

Figure 11: Per-subject similarity distributions and normal Q–Q plots. Orange dashed and green dotted lines indicate Gaussian-reference and empirical percentile boundaries, respectively.

A.7.2 Comparison with Alternative Losses

We compare boundary calibration with SupCon and hard-example weighting. These alternatives achieve 66.7% / 94.6% and 64.1% / 93.0% Top-1/Top-5 accuracy, respectively, below CAIA's 70.3% / 96.1% (table 17). Together with the component ablations, these results support boundary calibration as a complementary signal for adaptive visual fusion and spectral screening.

Table 17: Top-1/Top-5 accuracy (%) for alternative losses on THINGS-EEG.

Method	Top-1	Top-5
SupCon	<u>66.7</u>	<u>94.6</u>
Hard-example weighting	64.1	93.0
CAIA	**70.3**	**96.1**

A.7.3 Reference Interval Sensitivity

We evaluate reference interval coverage levels of 90%, 95%, and 99% (table 18). Mean Top-1 accuracy remains between 69.9% and 70.3%, while Top-5 accuracy ranges from 94.7% to 96.1%. The 95% setting achieves the highest mean accuracy for both metrics, indicating limited sensitivity within the tested range.

A.8 Computational Cost Analysis

We compare the computational cost of CAIA and UBP [19] in table 19. CAIA introduces additional parameters and FLOPs through adaptive visual fusion and frequency screening, with an inference latency of 14.29 ms compared with 5.88 ms for UBP.

Table 18: Sensitivity to reference interval coverage on THINGS-EEG (200-way zero-shot retrieval, subject-dependent, Top-1/Top-5, %).

	Sub 1		Sub 2		Sub 3		Sub 4		Sub 5		Sub 6		Sub 7		Sub 8		Sub 9		Sub 10		Avg	
$1-\alpha$	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5
99%	63.0	97.0	**80.0**	**97.5**	**79.0**	**99.0**	50.5	88.5	63.0	93.0	77.5	95.0	**64.5**	93.5	80.0	97.0	64.5	94.0	80.0	**97.5**	70.2	95.2
90%	69.5	97.5	78.5	96.5	69.0	97.5	**56.0**	**92.5**	58.5	90.5	**79.0**	95.5	64.0	93.0	80.0	97.5	63.5	90.0	**80.5**	96.0	69.9	94.7
95%	**74.0**	**98.0**	**80.0**	**97.5**	69.0	94.0	51.0	**92.5**	**67.5**	**94.0**	67.5	**96.5**	63.0	**95.5**	**83.0**	**99.0**	**69.0**	**96.5**	79.0	**97.5**	**70.3**	**96.1**

Prior work discusses latency targets below 200 ms for EEG decoding [54] and tighter requirements in some implantable applications [55], while wireless EEG acquisition can itself introduce delays of tens of milliseconds [56]. These considerations place the additional model execution time in the context of a broader acquisition and decoding pipeline. CAIA's measured inference latency remains below these reported latency targets, suggesting potential for latency-sensitive decoding applications, although end-to-end feasibility depends on acquisition, preprocessing, and trial aggregation.

The additional training and inference costs accompany substantial retrieval gains, providing a practical trade-off between computational efficiency and decoding performance.

Table 19: Computational cost comparison on an NVIDIA RTX 3090.

Method	Params (M)	FLOPs (G)	Inference (ms)	Train Step (ms)
UBP	43.72	5.423	5.88 ± 0.58	8.35 ± 0.49
CAIA	107.69	6.559	14.29 ± 0.44	34.79 ± 0.42

A.9 Retrieval Examples and Embedding Visualizations

A.9.1 Retrieval Examples

fig. 12 shows representative examples with correct Top-1 predictions, illustrating successful matching between EEG responses and their corresponding visual stimuli.

fig. 13 presents examples with incorrect Top-1 predictions. Visually similar or semantically related alternatives appear among the Top-5 candidates, suggesting that the learned representations retain useful category-level information even when the target is not ranked first. These cases highlight the remaining difficulty of distinguishing visually or semantically similar stimuli.

A.9.2 UMAP Visualization and Stability

fig. 14 visualizes the EEG and image representations using UMAP. The modality-specific projections display broadly similar spatial organization, while the joint projection illustrates the distribution of correctly and incorrectly retrieved samples. These visualizations provide a qualitative view of the learned cross-modal representations.

To examine the stability of this geometric correspondence, we repeat the Procrustes analysis under five UMAP random seeds with $n_{\text{neighbors}} = 15$ and $\text{min_dist} = 0.1$. The matched-pair disparity ranges from 0.108 to 0.121, with a mean of 0.115, compared with approximately 0.993 for the random baseline (table 20). These results suggest that the observed geometric correspondence is stable across the tested UMAP initializations.

Table 20: Procrustes disparity across five UMAP random seeds.

UMAP Seed	0	1	2	3	4
Matched pairs	0.114	0.119	0.108	0.121	0.111
Random baseline	0.993	0.993	0.994	0.992	0.993

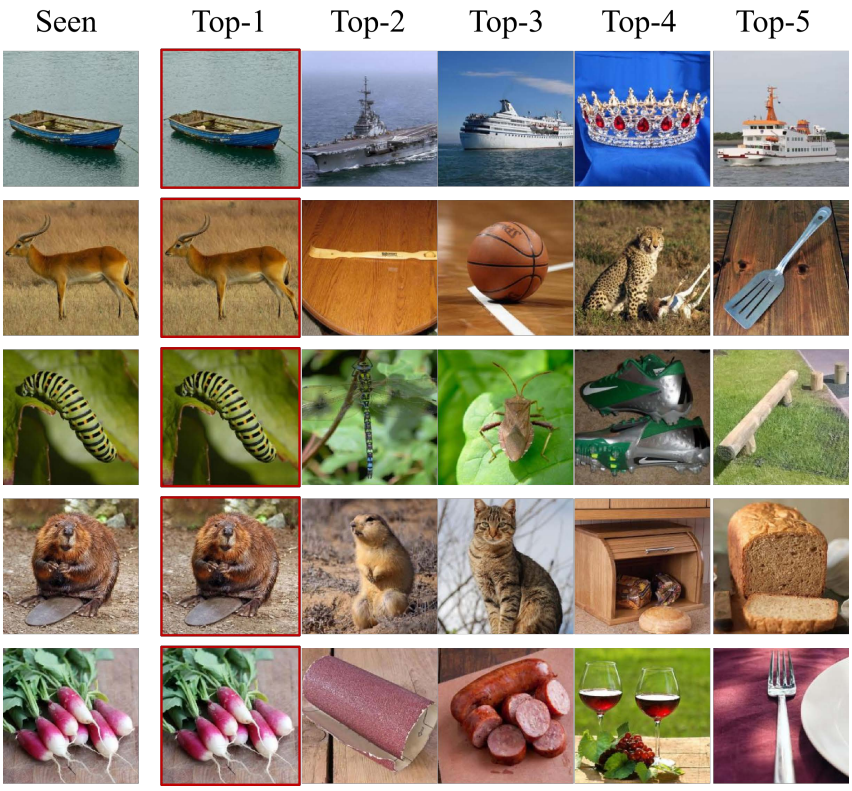


Figure 12: Successful retrieval examples.

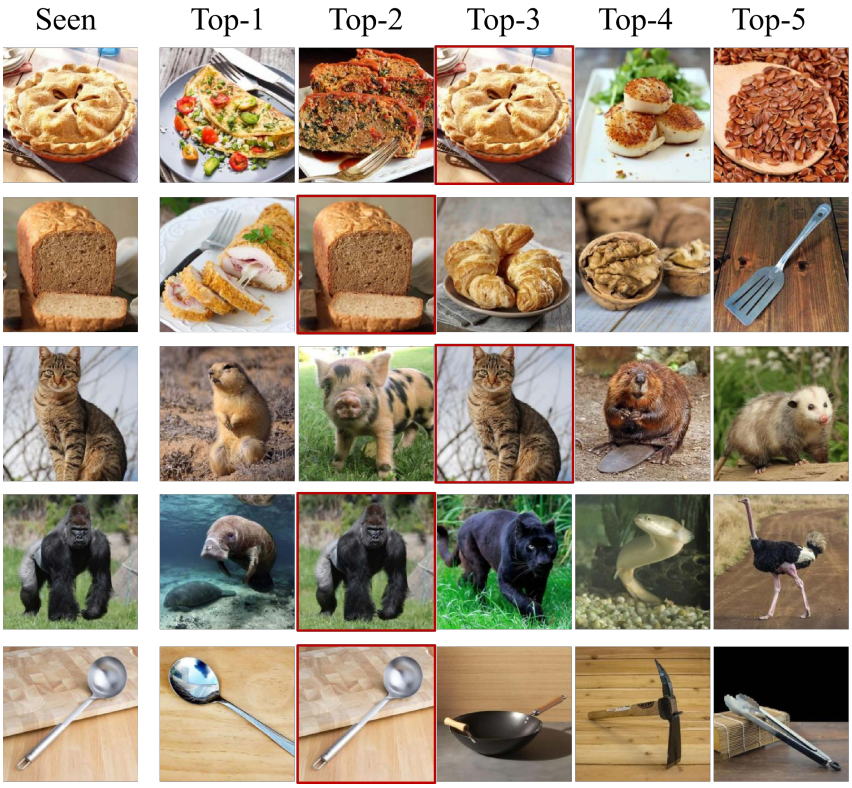
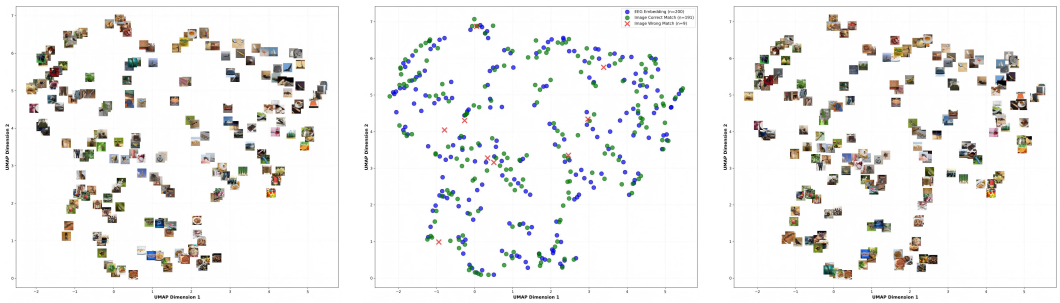


Figure 13: Failed retrieval examples.



(a) EEG representations. (b) Joint representations. (c) Image representations.

Figure 14: UMAP visualizations of EEG and image representations. In (b), green and red indicate correct and incorrect retrievals, respectively.

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 state the proposed CAIA framework and its components, supported by the experiments in section 4 and the supplementary analyses in Appendices A.3–A.9.

   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: section 5 discusses limitations related to sustained central fixation, repeated-trial averaging, and the absence of eye-tracking-based validation.

   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 does not include theoretical results; it focuses on empirical evaluation of the proposed method.

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: sections 3 and 4.1 describe the model, datasets, preprocessing, and training settings. Additional dataset details are provided in Appendix A.1, and code is available through the repository linked in the abstract.

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: Code is available at `https://github.com/mikumafu/CAIA`. THINGS-EEG and THINGS-MEG are publicly available and cited in section 4.1 and Appendix A.1.

   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: section 4.1 details the data splits, preprocessing, optimizer, batch sizes, and learning rates for both subject-dependent and subject-independent settings.

   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: Appendix A.3 reports standard errors and 95% confidence intervals across five seed-level mean accuracies, together with two-sided paired t-tests across ten subjects for seed-matched CAIA and UBP runs. Confidence intervals use Student's t distribution with four degrees of freedom.

   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: section 4.1 identifies the NVIDIA RTX 3090 GPU. Appendix A.8 and table 19 report model size, FLOPs, inference latency, and training step time for UBP and CAIA.

   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 uses publicly available datasets and adheres to standard ethical practices. No human subjects were directly involved. The paper does not violate any ethical guidelines.

   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: [No]

Justification: This work is foundational research aimed at improving EEG-visual alignment. It focuses on algorithmic design and empirical validation on public datasets, and does not directly address deployment-level societal impacts.

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 paper does not release any high-risk models or datasets; it only introduces an alignment framework that does not inherently pose misuse risks.

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 used datasets and prior methods are cited. The datasets are publicly available research datasets; specific licenses are stated in their respective publications.

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: The code repository linked in the abstract provides implementation documentation.

    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 uses existing public datasets; no new data collection from human subjects was conducted.

    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 paper does not involve primary human subjects research; it solely uses previously published datasets.

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: No LLMs were used as part of the core method; the research does not rely on LLMs in any significant way.

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.