CFBA: CONTOUR-PRESERVING FOVEATED BLUR ALIGNMENT FOR NEURAL VISUAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding visual content from brain signals helps characterize human perception and supports brain–computer interfaces. However, existing approaches to visual supervision adaptation do not explicitly combine contour preservation with spatially varying detail suppression, potentially limiting neural–visual alignment. To address this limitation, we propose Contour-preserving Foveated Blur Alignment (CFBA), which combines explicit preservation of estimated contour pixels with spatially varying suppression of surrounding detail. EGNet predictions are refined through morphological closing and contour-length filtering to construct a contour mask. An exponential radial profile blends the original image with its Gaussian-blurred counterpart, retaining more original detail near the image center. Restoring the masked contour pixels produces the supervision target. A Residual Projection Encoder (RPE) aligns neural recordings with frozen CLIP ResNet-50 features of this target through symmetric contrastive learning. A separate ATM–ViT-H-14 branch learns alignment under original-image supervision. The RPE supports retrieval and classification. For reconstruction, a diffusion prior converts the ATM embedding into the primary condition, while a linear projection trained through the generation objective maps the RPE embedding into an auxiliary condition. IP-Adapters inject both conditions into a pretrained SDXL generator. Extensive experiments on THINGS-EEG and THINGS-MEG demonstrate that CFBA outperforms state-of-the-art methods across retrieval, classification, and reconstruction. The code is available at `https://anonymous.4open.science/r/CFBA-D9CD`.

1 INTRODUCTION

Neural visual decoding identifies or reconstructs viewed images from the associated brain activity (Kay et al., 2008; Miyawaki et al., 2008), providing insights into visual representations and supporting brain–computer interfaces. Functional magnetic resonance imaging (fMRI), combined with latent diffusion models, enables detailed natural-image reconstruction (Takagi & Nishimoto, 2023), but its slow hemodynamic response limits the temporal precision of visual decoding (Benchetrit et al., 2024). Electroencephalography (EEG) and magnetoencephalography (MEG) offer high temporal resolution, with EEG also providing comparatively accessible and portable acquisition (Song et al., 2024; Li et al., 2024). Advances in pretrained visual representations and generative models have consequently increased interest in EEG- and MEG-based classification, retrieval, and reconstruction (Song et al., 2024; Li et al., 2024; Benchetrit et al., 2024).

A common approach aligns neural features with representations of the original stimulus images. NICE (Song et al., 2024) uses contrastive alignment for zero-shot object recognition. ATM (Li et al., 2024) connects CLIP-aligned EEG embeddings to diffusion-based generation for classification, retrieval, and reconstruction. CognitionCapturer (Zhang et al., 2025) extends supervision to image, text, and depth, learning separate EEG representations and combining their conditions for reconstruction. Although these methods improve neural representation learning and generative conditioning, features of the original stimuli can include information that is not reliably recoverable from neural recordings, creating a mismatch between visual supervision and the information available to the neural encoder.

This mismatch has motivated efforts to adapt visual supervision. The uncertainty-aware blur prior (UBP) (Wu et al., 2025) regulates image detail using cross-modal uncertainty to improve retrieval. Human-aligned image encoders (Rajabi et al., 2025) incorporate perceptual judgments

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

into supervision targets, improving retrieval across neural architectures and recording modalities. ShrinkAdapter (Wu et al., 2026) adapts the visual teacher through a task-guided bottleneck. These results establish target design as an effective means of improving neural–visual alignment. Psychophysical evidence shows that aligned local elements form coherent contours and that spatial resolution declines with retinal eccentricity (Field et al., 1993; Strasburger et al., 2011). Together with Gestalt accounts of perceptual grouping (Wagemans et al., 2012), these findings motivate preserving shape-defining boundaries while selectively simplifying surrounding detail (Fig. 1). However, existing approaches to visual supervision adaptation do not explicitly protect estimated contour pixels during spatially varying detail suppression, leaving informative boundaries vulnerable to attenuation and potentially limiting decoding performance.

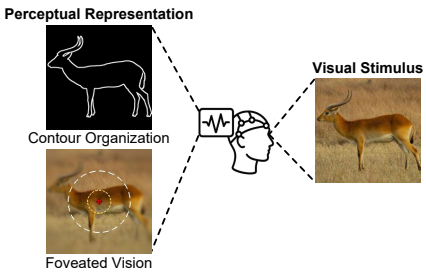

Figure 1: Contour organization and spatially nonuniform resolution motivate selective preservation of visual detail.

To address this limitation, we propose Contour-preserving Foveated Blur Alignment (CFBA). The framework refines EGNet predictions into a contour mask and preserves the selected pixels while increasing the contribution of Gaussian-blurred content toward the image periphery. A Residual Projection Encoder (RPE) aligns neural recordings with frozen CLIP ResNet-50 features of the resulting target. A separate original-image branch pairs an ATM with ViT-H-14. The RPE supports retrieval and classification, and its projected embedding complements the primary condition produced by the original-image diffusion prior. IP-Adapters inject both conditions into SDXL for reconstruction. Our contributions are threefold.

- To our knowledge, we are the first to jointly preserve estimated contour pixels and apply spatially varying context smoothing when constructing visual supervision for visual decoding.
- We propose CFBA, integrating selective visual supervision with independent alignment and complementary conditioning to support discriminative decoding and image reconstruction.
- CFBA surpasses all state-of-the-art methods across retrieval, classification, and reconstruction on both THINGS-EEG and THINGS-MEG.

2 Related Work

2.1 Neural visual decoding and visual supervision

Early fMRI studies demonstrated visual identification and reconstruction from brain activity (Kay et al., 2008; Miyawaki et al., 2008), with diffusion models later enabling detailed natural-image reconstruction (Takagi & Nishimoto, 2023). For EEG, NICE (Song et al., 2024), ATM (Li et al., 2024), and CognitionCapturer (Zhang et al., 2025) advance contrastive alignment, generative decoding, and multimodal conditioning, respectively. MEG decoding similarly combines representation learning with generation and examines the effects of visual features and response timing (Benchetrit et al., 2024). These methods develop neural representations around pretrained visual targets. Subsequent work adapts the targets themselves through uncertainty-aware foveated blur in UBP (Wu et al., 2025), perceptual alignment of image encoders (Rajabi et al., 2025), and task-adaptive visual teaching in ShrinkAdapter (Wu et al., 2026). Together, image transformation and feature adaptation establish complementary routes to improved supervision, motivating the joint preservation of informative contours and selective suppression of surrounding detail.

2.2 Perceptual organization and foveated spatial resolution

Perceptual organization groups local visual elements into coherent structures and separates them from their surroundings. Wagemans et al. (2012) review contour grouping and figure–ground organization within modern Gestalt theory, emphasizing relationships among visual elements. Field et al. (1993) show that detecting contours formed by oriented elements in distractors depends on their spatial and

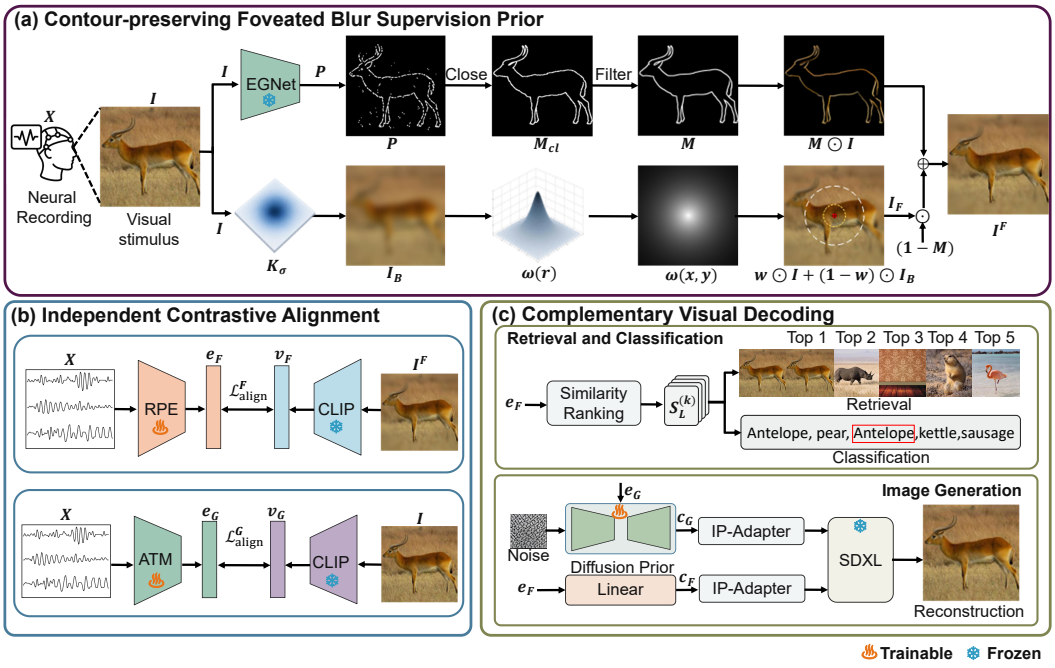

Figure 2: Overview of CFBA. **(a)** Contour-preserving foveated blur supervision prior. **(b)** Independent contrastive alignment. **(c)** Complementary visual decoding.

orientational relationships, supporting an association-field account of contour integration. These findings motivate retaining coherent boundaries when simplifying visual supervision.

Spatial resolution varies across the visual field. Strasburger et al. (2011) review changes in acuity, contrast sensitivity, and pattern recognition with eccentricity. Curcio et al. (1990) document declining cone density away from the foveal center, providing anatomical evidence for nonuniform retinal sampling. These observations motivate spatially varying treatment of image detail, although blur alone does not capture peripheral limitations such as crowding. Contour organization and foveated resolution therefore provide complementary grounds for preserving selected boundaries while reducing surrounding detail.

3 Method

CFBA comprises visual-prior construction, independent contrastive alignment, and complementary decoding (Fig. 2). The transformed-image branch supports retrieval and classification and supplies an auxiliary reconstruction condition. The original-image branch provides the primary condition through a diffusion prior.

3.1 Problem Definition

Let $\mathcal{D} = \{(\mathbf{X}_i, I_i, y_i)\}_{i=1}^{N}$ contain N paired neural recordings, RGB stimuli, and category labels. Each recording $\mathbf{X}_i \in \mathbb{R}^{C_s \times T}$ contains C_s selected EEG channels or MEG sensors and T time samples. Each stimulus $I_i \in [0,1]^{H \times W \times 3}$ has height H and width W. EEG and MEG are evaluated separately. The transformation Φ constructs the target $I^F = \Phi(I)$ for branch F, while branch G uses the original image. Given $\mathbf{X}$, retrieval ranks candidate images, classification predicts a category, and reconstruction produces $\widehat{I}$.

3.2 Contour-preserving Foveated Blur Supervision Prior

As shown in Fig. 2(a), the prior combines a binary contour mask with an image-centered radial blending profile.

3.2.1 CONTOUR PRESERVATION

A frozen Edge Guidance Network (EGNet) (Zhao et al., 2019) produces a sigmoid-normalized prediction map $P \in [0,1]^{H\times W}$, resampled to the stimulus grid. Input resizing and BGR mean subtraction follow EGNet preprocessing. With $p_{\max} = \max_{x,y} P(x,y)$, the initial binary mask is

$$M_0(x,y) = \begin{cases} 0, & p_{\max} < \epsilon, \\ \mathbf{1}[P(x,y)/p_{\max} > \tau_c], & \text{otherwise}, \end{cases} \tag{1}$$

where (x,y) indexes a pixel, $\mathbf{1}[\cdot]$ is the indicator function, $\epsilon = 0.1$ rejects weak prediction maps, and $\tau_c = 0.4$ thresholds the peak-normalized response. Closing with a 3×3 structuring element S connects short gaps.

$$M_{\text{cl}} = (M_0 \oplus S) \ominus S, \qquad M = \text{Rasterize}_\kappa(\{c \in \mathcal{C}(M_{\text{cl}}) : \ell(c) \geq \delta_{\text{len}}\}). \tag{2}$$

Here, $\oplus$ and $\ominus$ denote dilation and erosion, $\mathcal{C}(M_{\text{cl}})$ is the extracted contour set, and $\ell(c)$ is the arc length of contour c in pixels. Contours meeting the minimum length δ_{len} are rasterized with thickness κ pixels, yielding $M \in \{0,1\}^{H\times W}$. Pixels with $M = 1$ form the protected contour region.

3.2.2 FOVEATED BLUR AND CONTOUR COMPOSITION

Motivated by the image-centered prior in UBP (Wu et al., 2025), we retain more original detail centrally and increase the blurred-image contribution toward the periphery. For the pixel grid $\Omega = \{0,\dots,W-1\} \times \{0,\dots,H-1\}$ and center $\mathbf{c} = ((W-1)/2, (H-1)/2)$, define

$$r(x,y) = \frac{\|(x,y) - \mathbf{c}\|_2}{d_{\max}}, \qquad d_{\max} = \max_{(x,y)\in\Omega} \|(x,y) - \mathbf{c}\|_2, \tag{3}$$

where $\|\cdot\|_2$ denotes the Euclidean norm and $r \in [0,1]$ is normalized radial distance. The original-image retention weight is

$$w(x,y) = R_e + (R_c - R_e)e^{-\gamma r(x,y)}, \qquad 0 \leq R_e < R_c \leq 1. \tag{4}$$

The parameters R_c and R_e specify the center value and asymptotic lower level, respectively, and $\gamma = 3$ controls radial decay. Thus w decreases with radius, reaching $R_e + (R_c - R_e)e^{-\gamma}$ at $r = 1$.

Let $I_B = K_\sigma * I$, where K_σ is a discretely normalized Gaussian kernel with standard deviation σ and support $k_B \times k_B$, and $*$ denotes channel-wise spatial convolution. The foveated image I_F and final supervision image I^F are

$$I_F = w \odot I + (1 - w) \odot I_B, \tag{5}$$

$$I^F = \Phi(I) = M \odot I + (1 - M) \odot I_F. \tag{6}$$

Here, $\odot$ is element-wise multiplication, with spatial masks broadcast across RGB channels. The composition retains the original values wherever $M = 1$ and applies the foveated blend elsewhere, combining contour preservation with stronger peripheral smoothing.

3.3 INDEPENDENT CONTRASTIVE ALIGNMENT

The two branches in Fig. 2(b) receive the same neural recording and produce

$$\begin{aligned} \mathbf{e}_F &= f_F^{\text{RPE}}(\mathbf{X}), \quad \mathbf{v}_F = F_{\text{RN50}}(I^F), \\ \mathbf{e}_G &= f_G^{\text{ATM}}(\mathbf{X}), \quad \mathbf{v}_G = F_{\text{H14}}(I). \end{aligned} \tag{7}$$

The neural encoders f_F^{RPE} and f_G^{ATM} are RPE and ATM, respectively. The image encoders F_{RN50} and F_{H14} use CLIP ResNet-50 and ViT-H-14. Branch F learns d-dimensional embeddings against frozen RN50 targets (Radford et al., 2021). Branch G independently trains ATM (Li et al., 2024) and ViT-H-14 to produce d_g-dimensional embeddings from original-image pairs. For each branch b, $\mathbf{e}_b$ and $\mathbf{v}_b$ denote neural and visual embeddings.

Residual projection encoder. The RPE flattens channel and time samples into $\mathbf{h}_0 = \mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{C_s T}$ and computes

$$\begin{aligned} \mathbf{h}_1 &= \mathcal{A}_1(\mathbf{W}_1 \mathbf{h}_0 + \mathbf{b}_1), \\ \mathbf{h}_{\mathrm{res}} &= \mathbf{h}_1 + \mathcal{A}_{\mathrm{res}}(\mathbf{W}_{\mathrm{res}} \mathbf{h}_1 + \mathbf{b}_{\mathrm{res}}), \\ \mathbf{e}_F &= \mathrm{LN}_{\mathrm{out}}(\mathbf{W}_{\mathrm{out}} \mathbf{h}_{\mathrm{res}} + \mathbf{b}_{\mathrm{out}}). \end{aligned} \tag{8}$$

For $j \in \{1, \mathrm{res}\}$, $\mathcal{A}_j(\mathbf{h}) = \mathrm{LN}_j(\mathrm{Dropout}(\mathrm{GELU}(\mathbf{h})))$ applies a Gaussian error linear unit, dropout, and stage-specific LayerNorm. Both hidden vectors have dimension 1024. The learned matrices $\mathbf{W}_1$, $\mathbf{W}_{\mathrm{res}}$, and $\mathbf{W}_{\mathrm{out}}$ have dimensions $1024 \times C_s T$, 1024×1024, and $d \times 1024$. The corresponding biases have dimensions 1024, 1024, and d, and $\mathrm{LN}_{\mathrm{out}}$ normalizes the final projection.

Symmetric contrastive objective. For branch $b \in \{F, G\}$ and a minibatch of B paired samples, let

$$s_{ij}^b = \exp(a_b) \overline{\mathbf{e}}_{b,i}^\top \overline{\mathbf{v}}_{b,j}, \qquad \overline{\mathbf{x}} = \mathbf{x} / \|\mathbf{x}\|_2, \tag{9}$$

where $i, j \in \{1, \ldots, B\}$ index neural and image samples, and a_b is a learnable logit scale initialized to $\log(1/0.07)$. Each branch independently minimizes

$$\mathcal{L}_{\mathrm{align}}^b = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(s_{ii}^b)}{\sum_{j=1}^{B} \exp(s_{ij}^b)} + \log \frac{\exp(s_{ii}^b)}{\sum_{j=1}^{B} \exp(s_{ji}^b)} \right]. \tag{10}$$

The two InfoNCE terms align neural-to-image and image-to-neural representations. Branch F updates RPE parameters, while branch G updates its ATM and ViT-H-14 encoders.

3.4 Complementary Visual Decoding

Fig 2(c) shows retrieval and classification from branch F and reconstruction conditioned on both branches.

3.4.1 Retrieval and Classification

For candidate images $I_j \in \mathcal{I}_{\mathrm{cand}}$ and categories $y \in \mathcal{Y}$, define

$$s^{\mathrm{ret}}(\mathbf{X}, I_j) = \overline{\mathbf{e}}_F^\top \overline{F_{\mathrm{RN50}}(\Phi(I_j))}, \tag{11}$$

$$s^{\mathrm{cls}}(\mathbf{X}, y) = \overline{\mathbf{e}}_F^\top \overline{F_T(p(y))}. \tag{12}$$

Here, $\mathcal{I}_{\mathrm{cand}}$ and $\mathcal{Y}$ are the image and category candidate sets, F_T is the frozen text encoder paired with RN50, and $p(y)$ is the category prompt. Retrieval ranks transformed candidate-image features by cosine similarity. Classification ranks category-text features and predicts the highest-scoring category.

3.4.2 Image Reconstruction

Primary diffusion prior. A conditional diffusion prior predicts the original-image embedding $\mathbf{z}_0 = \mathbf{v}_G$ from $\mathbf{e}_G$. At step $t \in \{1, \ldots, T_p\}$,

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t} \mathbf{v}_G + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_p, \tag{13}$$

$$\mathcal{L}_{\mathrm{prior}} = \mathbb{E}\left[\|\boldsymbol{\epsilon}_p - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{e}_G)\|_2^2 \right]. \tag{14}$$

Here, T_p is the number of prior diffusion steps, $\bar{\alpha}_t = \prod_{j=1}^{t} (1 - \beta_j)$ for noise variances $\beta_j \in (0, 1)$, and $\boldsymbol{\epsilon}_p \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_g})$ is Gaussian noise with identity covariance. The denoiser $\boldsymbol{\epsilon}_\theta$, parameterized by θ, predicts this noise conditioned on $\mathbf{e}_G$. The expectation averages over training pairs, diffusion steps, and noise. At inference, reverse diffusion from Gaussian noise produces the primary condition $\mathbf{c}_G = \widehat{\mathbf{z}}_0 \in \mathbb{R}^{d_g}$.

Auxiliary projection and joint conditioning. An affine projection maps the CFBA embedding into the shared conditioning space.

$$\mathbf{c}_F = \mathbf{A}_F \mathbf{e}_F + \mathbf{q}_F \in \mathbb{R}^{d_g}, \tag{15}$$

where $\mathbf{A}_F \in \mathbb{R}^{d_g \times d}$ and $\mathbf{q}_F \in \mathbb{R}^{d_g}$ are learned parameters. Separate IP-Adapters (Ye et al., 2023) inject $\mathbf{c}_G$ and $\mathbf{c}_F$ into pretrained SDXL (Podell et al., 2024) with the same scale ω.

$$\mathbf{H}_{\text{out}} = \mathbf{H}_{\text{base}} + \omega \sum_{b \in \{G,F\}} \text{softmax}(\mathbf{Q}\mathbf{K}_b^\top / \sqrt{d_h})\mathbf{V}_b. \tag{16}$$

The matrices $\mathbf{H}_{\text{base}}$ and $\mathbf{H}_{\text{out}}$ are the base and conditioned attention outputs. Queries $\mathbf{Q}$ come from generator features, while keys $\mathbf{K}_b$ and values $\mathbf{V}_b$ come from adapter tokens for $\mathbf{c}_b$. The scalar d_h is the attention-head key/query dimension, and softmax normalizes scores over conditioning tokens.

Generation objective. The auxiliary projection learns through the image-generation loss. Let $\mathbf{u}_0 = \mathcal{E}_{\text{VAE}}(I) \in \mathbb{R}^{d_u}$ be the SDXL variational autoencoder latent, with d_u entries in vectorized form. For step $s \in \{1, \ldots, T_g\}$,

$$\mathbf{u}_s = \sqrt{\bar{\gamma}_s}\mathbf{u}_0 + \sqrt{1 - \bar{\gamma}_s}\boldsymbol{\epsilon}_g, \tag{17}$$

$$\mathcal{L}_{\text{gen}} = \mathbb{E}\left[\|\boldsymbol{\epsilon}_g - \boldsymbol{\epsilon}_\psi(\mathbf{u}_s, s, \mathbf{c}_G, \mathbf{c}_F)\|_2^2\right]. \tag{18}$$

Here, T_g is the number of image-diffusion steps, $\bar{\gamma}_s = \prod_{j=1}^{s}(1 - \eta_j)$ for variances $\eta_j \in (0, 1)$, and $\boldsymbol{\epsilon}_g \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_u})$ is independent image-latent noise. The function $\boldsymbol{\epsilon}_\psi$ denotes the conditioned image denoiser and its adapter pathways. The expectation covers training examples, diffusion steps, noise, and sampled primary conditions. Gradients through the auxiliary adapter update $\mathbf{A}_F$ and $\mathbf{q}_F$.

At inference, $\widehat{I} = \mathcal{G}_{\text{SDXL}}(\boldsymbol{\xi}, \mathbf{c}_G, \mathbf{c}_F)$, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_u})$ is initial latent noise and $\mathcal{G}_{\text{SDXL}}$ comprises conditional denoising and VAE decoding. Both conditions are derived from $\mathbf{X}$.

4 Experiments and Results

4.1 Comparison with SOTA Methods

Due to space limitations, Appendix A.1 provides dataset descriptions, preprocessing, evaluation metrics, baselines, and implementation details. We evaluate CFBA against state-of-the-art methods on THINGS-EEG and THINGS-MEG across retrieval, classification, and reconstruction. CFBA achieves state-of-the-art overall performance across the three tasks on both datasets, as summarized in Fig. 3. Appendix A.2 reports subject-wise retrieval in Tables 5 and 6, classification in Table 7, and reconstruction in Table 8. Retrieval examples and reconstruction comparisons appear in Figs. 6 and 7 (Appendix A.2). They illustrate recognizable target matches and improved reconstruction of object structure, while similar non-target candidates reveal remaining ambiguity in the aligned representation. The detailed tables also show leading mean retrieval and classification performance in both modalities. For reconstruction, CFBA leads on five of seven EEG metrics and all seven MEG metrics, indicating benefits across pixel-level and learned-feature evaluations.

4.2 Ablation Study

CFBA components. To assess the contribution of each component in the visual prior, we compare CFBA with variants that remove context smoothing, the visual prior, or contour preservation. As shown in Table 1, removing any component degrades retrieval, classification, and reconstruction on both THINGS-EEG and THINGS-MEG. The complete CFBA configuration achieves the best results across all reported metrics, supporting the joint use of contour preservation and foveated context smoothing. Qualitative comparisons appear in Appendix A.2 (Fig. 7). The displayed reconstructions better preserve object structure and contextual appearance, supporting the use of the transformed-image representation alongside original-image conditioning. Compared with foveated smoothing alone, restoring contours improves every reported metric on both datasets (Table 1). This comparison isolates the benefit of protecting boundaries within the smoothed target.

Visual-target alignment. To assess how the prior changes neural–visual alignment, we compare matched-pair similarity and discriminability across four targets. Matched-pair similarity is the mean diagonal entry of the EEG–image similarity matrix, and discriminability is the mean diagonal minus the mean off-diagonal similarity. Table 2 shows that CFBA improves both measures over original-image and foveated-blur targets. These results support retaining contours while simplifying the surrounding image.

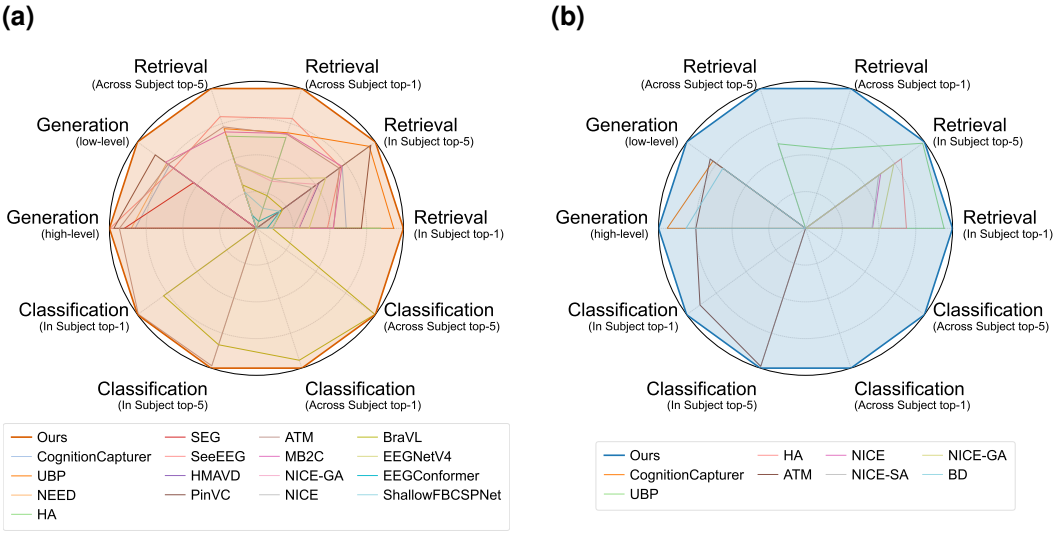

Figure 3: Comparison across retrieval, classification, and reconstruction on **(a)** THINGS-EEG and **(b)** THINGS-MEG.

Table 1: CFBA ablation on THINGS-EEG and THINGS-MEG. Accuracies are proportions. w/o context uses contours only, and w/o prior uses original images.

Method	Retrieval				Classification				Reconstruction						
	Intra-subject		Inter-subject		Intra-subject		Inter-subject		Low-level		High-level				
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
THINGS-EEG															
w/o context	0.270	0.605	0.067	0.228	0.033	0.119	0.014	0.052	0.102	0.333	0.621	0.684	0.561	0.571	0.714
w/o prior	0.422	0.766	0.094	0.306	0.043	0.151	0.015	0.057	0.158	0.420	0.755	0.832	0.750	0.738	0.573
w/o contour	0.497	0.797	0.131	0.335	0.067	0.179	0.019	0.068	0.182	0.466	0.815	0.891	0.770	0.804	0.544
CFBA	**0.545**	**0.840**	**0.176**	**0.465**	**0.075**	**0.210**	**0.020**	**0.071**	**0.194**	**0.486**	**0.840**	**0.904**	**0.787**	**0.832**	**0.522**
THINGS-MEG															
w/o context	0.048	0.175	0.016	0.059	0.014	0.053	0.010	0.038	0.079	0.273	0.635	0.681	0.589	0.559	0.742
w/o prior	0.229	0.504	0.019	0.095	0.031	0.115	0.011	0.042	0.113	0.366	0.740	0.827	0.785	0.723	0.596
w/o contour	0.265	0.550	0.031	0.145	0.058	0.170	0.013	0.044	0.133	0.397	0.804	0.883	0.801	0.789	0.571
CFBA	**0.281**	**0.570**	**0.038**	**0.161**	**0.066**	**0.195**	**0.015**	**0.048**	**0.142**	**0.412**	**0.834**	**0.899**	**0.820**	**0.815**	**0.543**

Table 2: EEG–image alignment under different visual targets.

Visual target	Matched-pair similarity ↑	Discriminability ↑
Contour only	0.0496	0.0277
Original image	0.1052	0.1133
Foveated blur	0.1269	0.1164
CFBA	**0.1619**	**0.1375**

Blur strategies and contour detectors. To evaluate the construction of the prior, we vary the smoothing strategy and contour detector while retaining the remaining components. Table 3 compares uniform, contour-distance, and foveated blur, together with morphological gradients, Canny edges, and EGNet contours. Distance blur uses distance from contours, and Morph. gradient denotes morphological gradients. Foveated blur and EGNet achieve the best results within their respective comparisons across all reported metrics, supporting the selected combination.

Neural and image encoders. To determine which encoder pairing supports retrieval, we compare neural encoders across image backbones in Fig. 4. RPE achieves the highest accuracy with each

Table 3: Blur strategies and contour detectors on THINGS-EEG. Accuracies are proportions.

Variant	Retrieval				Classification				Reconstruction						
	Intra-subject		Inter-subject		Intra-subject		Inter-subject		Low-level		High-level				
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
Blur strategy															
Distance blur	0.484	0.804	0.110	0.355	0.061	0.197	0.014	0.063	0.186	0.482	0.838	0.902	0.783	0.831	0.538
Uniform blur	0.508	0.831	0.146	0.408	0.065	0.208	0.017	0.063	0.184	0.485	0.826	0.896	0.777	0.821	0.534
Foveated blur	**0.545**	**0.840**	**0.176**	**0.465**	**0.075**	**0.210**	**0.020**	**0.071**	**0.194**	**0.486**	**0.840**	**0.904**	**0.787**	**0.832**	**0.522**
Contour detector															
Morph. gradient	0.500	0.811	0.132	0.331	0.067	0.202	0.014	0.066	0.187	0.481	0.811	0.890	0.782	0.827	0.530
Canny	0.525	0.827	0.150	0.415	0.067	0.194	0.017	0.062	0.176	0.481	0.819	0.892	0.782	0.826	0.531
EGNet	**0.545**	**0.840**	**0.176**	**0.465**	**0.075**	**0.210**	**0.020**	**0.071**	**0.194**	**0.486**	**0.840**	**0.904**	**0.787**	**0.832**	**0.522**

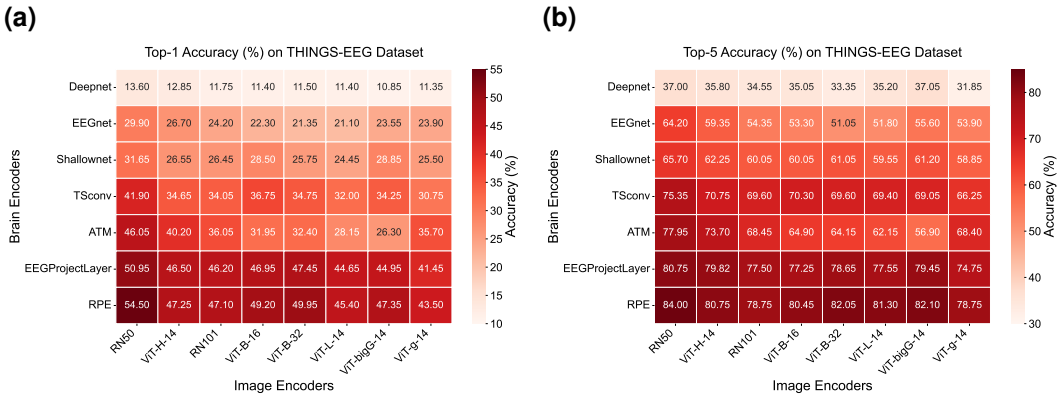

Figure 4: Neural–image encoder comparison on THINGS-EEG. **(a)** Top-1 and **(b)** Top-5 intra-subject retrieval accuracy.

Table 4: Primary and auxiliary encoder assignments for reconstruction on THINGS-EEG. Each pair lists the neural encoder followed by the image encoder.

Primary	Auxiliary	Low-level		High-level				
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
RPE–RN50	RPE–RN50	0.1563	0.3339	0.5649	0.5155	0.5042	0.5025	0.7370
ATM–ViT-H-14	ATM–ViT-H-14	0.1787	0.3836	0.8031	0.8503	0.7667	0.6829	0.6222
RPE–RN50	ATM–ViT-H-14	0.1492	0.3516	0.5209	0.5163	0.5292	0.5398	0.6804
ATM–ViT-H-14	RPE–RN50	**0.1944**	**0.4862**	**0.8398**	**0.9041**	**0.7871**	**0.8319**	**0.5223**

tested image backbone, and RN50 provides better retrieval targets than ViT-H-14. These results support RPE–RN50 for the transformed-image branch. The paired-encoder comparison also favors RPE–RN50 across retrieval and classification (Appendix A.5, Table 19), showing that the larger encoder pair does not necessarily provide better discriminative alignment. We separately vary the primary and auxiliary encoder assignments for reconstruction in Table 4. ATM–ViT-H-14 as the primary branch and RPE–RN50 as the auxiliary branch outperform the homogeneous and reversed assignments on all seven metrics. This result is consistent with original-image conditioning and prior-specific conditioning serving complementary roles. The additional encoder results show that the retrieval advantage of RPE–RN50 extends to classification and cross-subject evaluation (Table 19). This supports its use beyond within-subject image matching.

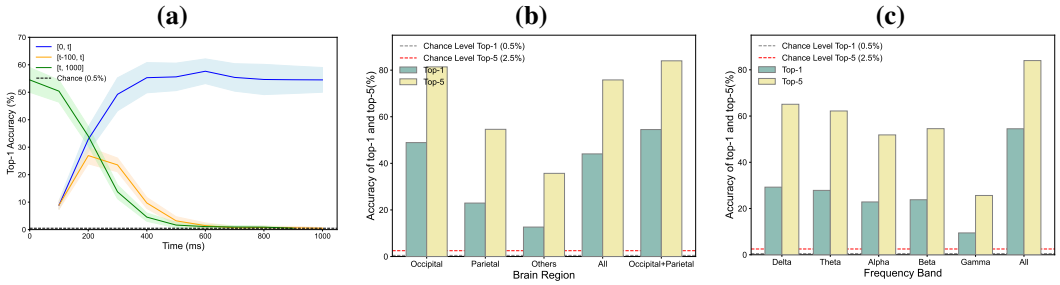

Figure 5: EEG retrieval across **(a)** temporal windows, **(b)** channel subsets, and **(c)** frequency bands.

4.3 Temporal, Spatial, and Spectral Analysis

To identify which recorded information supports decoding, we vary the post-stimulus window, EEG channel subset, and frequency band. Figure 5 summarizes the results. Among non-overlapping 100-ms windows, the 100–200 ms interval yields the strongest retrieval performance. Occipital–Parietal channels outperform the other tested subsets in retrieval and classification. The full passband outperforms individual frequency bands, with Delta achieving the highest retrieval and classification performance among the individual bands. Appendix A.3 provides temporal, spatial, and spectral visualizations in Figs. 8–10, followed by quantitative results in Tables 9–14. The cumulative-window results show that additional late responses do not consistently improve accuracy, while delayed-onset windows confirm the importance of early information. Reconstruction metrics favor partly different temporal and spatial configurations, indicating that input selection depends on the decoding task. For example, cumulative 0–600 ms EEG yields the strongest discriminative results, whereas full-window input maximizes SSIM (Table 10). The temporal optimum therefore depends on the information emphasized by the evaluation metric.

4.4 Hyperparameter Analysis

To assess sensitivity to the visual prior, we vary contour-mask thickness (Section A.4.1, Table 15), Gaussian kernel size (Section A.4.2, Table 16), contour threshold (Section A.4.3, Table 17), and minimum contour length (Section A.4.4, Table 18). The comparisons favor a two-pixel mask, a 45×45 kernel, a threshold of 0.4, and a minimum length of 100 pixels. These settings balance contour retention with suppression of surrounding detail, improving both discriminative decoding and reconstruction. The shared optima across tasks support a common prior configuration, while declines with thicker masks or stronger smoothing indicate sensitivity to the balance between preserved structure and simplified context.

5 Conclusion and Future Work

CFBA combines contour preservation with foveated context smoothing to construct visual supervision for neural decoding. Independent alignment and complementary conditioning improve retrieval, classification, and reconstruction on THINGS-EEG and THINGS-MEG. The prior uses an image-centered weighting profile and estimated contours, which limit its adaptation to individual viewing behavior and contour-detection errors. Performance also varies across subjects, and the current evaluation is restricted to static images. Future work will investigate gaze-adaptive priors, subject adaptation, and decoding of dynamic scenes.

6 AI Use Statement

We used a large language model to improve the clarity and grammar of the manuscript. The authors reviewed the revisions and take responsibility for the scientific content.

7 Ethics Statement

This study analyzes publicly available THINGS-EEG and THINGS-MEG recordings and involves no new participant recruitment. Neural data require privacy safeguards, and potential applications of visual decoding should respect informed consent and restrictions on secondary use.

8 Reproducibility Statement

We use public datasets and provide the code link in the abstract. Appendix A.1 describes preprocessing, evaluation protocols, and implementation details, and Appendices A.2–A.5 provide detailed results and ablations.

References

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: Toward real-time reconstruction of visual perception, 2024. URL https://arxiv.org/abs/2310.19812. Version 3.

Christine A. Curcio, Kenneth R. Sloan, Robert E. Kalina, and Anita E. Hendrickson. Human photoreceptor topography. *The Journal of Comparative Neurology*, 292(4):497–523, 1990. doi: 10.1002/cne.902920402. URL https://christineacurcio.com/PRtopo/Curcio_JCompNeurol1990_PRtopo_searchable.pdf.

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. URL https://doi.org/10.1109/TPAMI.2023.3263181.

David J. Field, Anthony Hayes, and Robert F. Hess. Contour integration by the human visual system: Evidence for a local "association field". *Vision Research*, 33(2):173–193, 1993. doi: 10.1016/0042-6989(93)90156-Q. URL https://pubmed.ncbi.nlm.nih.gov/8447091/.

Shuai Huang, Huan Luo, Haodong Jing, Qixian Zhang, Litao Chang, Yating Feng, Xiao Lin, Chendong Qin, Han Chen, Shuwen Jia, Siyi Sun, and Yongxiong Wang. NEED: Cross-subject and cross-task generalization for video and image reconstruction from EEG signals. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5761. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fca83589e85cb061631b7ebc5db5d6bd-Abstract-Conference.html.

Haodong Jing, Yongqiang Ma, Panqi Yang, Haibo Hua, and Nanning Zheng. Pinpointing visual content: Disentangled features in multimodal model for EEG representation learning and decoding. *Knowledge-Based Systems*, 329:114212, 2025. doi: 10.1016/j.knosys.2025.114212. URL https://doi.org/10.1016/j.knosys.2025.114212.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452:352–355, 2008. doi: 10.1038/nature06713.

Jun-Mo Kim, Woohyeok Choi, Sang-Jun Park, Keun-Soo Heo, Young-Han Son, Ji-Hye Oh, Dong-Hee Shin, and Tae-Eui Kam. SeeEEG: Semantic-aware EEG-based multi-modal retrieval-augmented generation for high-fidelity visual brain decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4883–4892, 2025. URL https://openaccess.thecvf.com/content/ICCV2025W/HCV/html/Kim_SeeEEG_Semantic-aware_EEG-based_Multi-Modal_Retrieval-Augmented_Generation_for_High-Fidelity_Visual_Brain_ICCVW_2025_paper.html.

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. URL https://doi.org/10.1088/1741-2552/aace8c.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ba5f1233efa77787ff9ec015877dbd1f-Abstract.html.

Lu Meng and Chuanhao Yang. Semantics-guided hierarchical feature encoding generative adversarial network for visual image reconstruction from brain activity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:1267–1283, 2024.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C. Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008. doi: 10.1016/j.neuron.2008.11.004.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nona Rajabi, Antonio H. Ribeiro, Miguel Vasco, Farzaneh Taleb, Mårten Björkman, and Danica Kragic. Human-aligned image models improve visual decoding from the brain. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51009–51038, 2025. URL https://proceedings.mlr.press/v267/rajabi25a.html.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730. URL https://doi.org/10.1002/hbm.23730.

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. URL https://doi.org/10.1109/TNSRE.2022.3230250.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dhLIno8FmH.

Hans Strasburger, Ingo Rentschler, and Martin Jüttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):13, 2011. doi: 10.1167/11.5.13. URL https://doi.org/10.1167/11.5.13.

Kaili Sun, Xingyu Miao, Bing Zhai, Haoran Duan, and Yang Long. Decoding visual neural representations by multimodal with dynamic balancing. *Expert Systems with Applications*, 297(Part B):129382, 2026. doi: 10.1016/j.eswa.2025.129382. URL https://doi.org/10.1016/j.eswa.2025.129382.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Takagi_High-Resolution_Image_Reconstruction_With_Latent_Diffusion_Models_From_Human_Brain_CVPR_2023_paper.html.

Johan Wagemans, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Rüdiger von der Heydt. A century of Gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological Bulletin*, 138(6):1172–1217, 2012. doi: 10.1037/a0029333. URL https://doi.org/10.1037/a0029333.

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*, pp. 8992–9000, 2024. doi: 10.1145/3664647.3681292. URL https://doi.org/10.1145/3664647.3681292.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2246–2257, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_Bridging_the_Vision-Brain_Gap_with_an_Uncertainty-Aware_Blur_Prior_CVPR_2025_paper.html.

Lukun Wu, Jie Li, Ziqi Ren, Kaifan Zhang, and Xinbo Gao. Shrinking the teacher: An adaptive teaching paradigm for asymmetric EEG-vision alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17859–17867, 2026. doi: 10.1609/aaai.v40i21.38844. URL https://ojs.aaai.org/index.php/AAAI/article/view/38844.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. CognitionCapturer: Decoding visual stimuli from human EEG signal with multimodal information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):14486–14493, 2025. doi: 10.1609/aaai.v39i13.33587. URL https://ojs.aaai.org/index.php/AAAI/article/view/33587.

Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8779–8788, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Zhao_EGNet_Edge_Guidance_Network_for_Salient_Object_Detection_ICCV_2019_paper.html.

Yanan Zhu, Ziwei Xiang, Jiamin Wu, Jinyang Guo, Hongyuan Zhang, Chunfeng Song, Hongjian Fang, Qihao Zheng, Yufei Guo, Xuelong Li, and Xianglong Liu. BRENA: Brain-inspired hierarchical neural alignment framework for visual decoding from EEG signals, 2025. URL https://openreview.net/forum?id=ybKEYQvM0L.

A Appendix

A.1 Experimental Settings

Datasets and preprocessing. THINGS-EEG contains recordings from 10 participants. Its training split includes 1,654 concepts with 10 images per concept and four presentations per image, giving 66,160 trials per participant. The test split contains 200 unseen concepts with one image per concept and 80 repetitions, giving 16,000 trials. Recordings are acquired at 1,000 Hz, filtered to 0.1–100 Hz, baseline-corrected over the 200 ms pre-stimulus interval, and downsampled to 250 Hz. We use 17 occipital–parietal channels over 0–1,000 ms and average the 80 repetitions of each test image. THINGS-MEG contains recordings from four participants with 271 channels. The main set contains 1,854 concepts with 12 images per concept and one presentation per image. The repeated test set contains 200 concepts with one image per concept and 12 repetitions. We exclude the test concepts from training, filter recordings to 0.1–40 Hz, apply baseline correction, and downsample to 200 Hz. We use the 0–1,000 ms interval and average the 12 test repetitions per image.

Evaluation metrics. Retrieval and classification use 200 image and category-text candidates, respectively, and are evaluated with Top-1 and Top-5 accuracy. Intra-subject evaluation trains and tests on the same participant, whereas inter-subject evaluation holds out one participant for testing. Reconstruction is assessed using PixCorr and SSIM for low-level similarity and AlexNet(2), AlexNet(5), Inception, CLIP, and SwAV for feature-level similarity. Higher values are better except for SwAV. Bold numbers indicate the best results within each dataset or comparison block, including ties. Captions distinguish percentages from proportions, and results are averaged across participants unless indicated otherwise.

Baselines. The compared methods include NICE and its variants (Song et al., 2024), ATM (Li et al., 2024), CognitionCapturer (Zhang et al., 2025), UBP (Wu et al., 2025), BRENA (Zhu et al., 2025), NEED (Huang et al., 2025), human-aligned encoders (Rajabi et al., 2025), SEG (Meng & Yang, 2024), SeeEEG (Kim et al., 2025), HMAVD (Sun et al., 2026), PinVC (Jing et al., 2025), MB2C (Wei et al., 2024), BraVL (Du et al., 2023), EEGNetV4 (Lawhern et al., 2018), EEGConformer (Song et al., 2023), ShallowFBCSPNet (Schirrmeister et al., 2017), and MEG decoding (Benchetrit et al., 2024). The result tables identify the methods compared for each task.

Implementation details. Experiments use PyTorch on a single NVIDIA RTX 4090 with 24 GB memory. The transformed-image branch uses frozen OpenCLIP RN50 with OpenAI pretrained weights and 1,024-dimensional features. RPE receives 17 EEG channels with 250 samples or 271 MEG channels with 200 samples. Its hidden width is 1,024 and its dropout probability is 0.5. We train the alignment network for up to 50 epochs using AdamW with learning rate 10^{-4}, weight decay 10^{-4}, and batch size 1,024. The learnable logit-scale parameter is initialized to $\log(1/0.07)$. Early stopping uses a patience of five epochs.

A.2 Overall Performance

Retrieval. To examine subject-level performance, Tables 5 and 6 report intra-subject and inter-subject retrieval. CFBA achieves the highest reported mean Top-1 and Top-5 accuracy in both settings on both datasets. On THINGS-EEG, mean intra-subject Top-1 accuracy increases from 50.0% with UBP to 54.5%, and mean inter-subject Top-1 accuracy increases from 12.0% to 17.5%. On THINGS-MEG, the corresponding CFBA accuracies are 28.125% and 3.75%. Fig 6 illustrates Top-1 matches and Top-5 rankings. Some non-target candidates share object categories or shapes with the stimulus, suggesting that the aligned representation captures useful visual relationships even when it misses the exact image. The lower inter-subject results indicate that transferring this representation across participants remains more difficult than decoding within a participant.

Classification. To evaluate transfer from image-instance matching to category prediction, Table 7 compares zero-shot classification. CFBA achieves the highest Top-1 and Top-5 accuracy on both datasets, reaching 7.45% and 20.95% on THINGS-EEG and 6.63% and 19.50% on THINGS-MEG. The gains across both datasets show that the image-aligned representation transfers to category-text matching. Classification therefore benefits from the prior beyond identifying individual stimulus images, although the absolute accuracies indicate substantial remaining difficulty.

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

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Mean	
Method	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Intra-subject																						
BraVL	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE	13.2	39.5	13.5	40.3	14.5	42.7	20.6	52.7	10.1	31.5	16.5	44.0	17.0	42.1	22.9	56.1	15.4	41.6	17.4	45.8	16.1	43.6
NICE-SA	13.3	40.2	12.1	36.1	15.3	39.6	15.9	49.0	9.8	34.4	14.2	42.4	17.9	43.6	18.2	50.2	14.4	38.7	16.0	42.8	14.7	41.7
NICE-GA	15.2	40.1	13.9	40.1	14.7	42.7	17.6	48.9	9.0	29.7	16.4	44.4	14.9	43.1	20.3	52.1	14.1	39.7	19.6	46.7	15.6	42.8
Neural-MCRL	24.2	52.5	17.3	45.9	22.0	50.8	27.7	56.1	16.6	40.8	23.6	48.9	25.3	49.3	33.1	62.8	21.5	46.7	24.1	56.8	23.5	51.1
ATM-S	25.6	60.4	22.0	54.5	25.0	62.4	31.4	60.9	12.9	43.0	21.3	51.1	30.5	61.5	38.8	72.0	34.4	51.5	29.1	63.5	28.5	60.4
MB2C	23.7	56.3	22.7	50.5	26.3	60.2	34.8	67.0	21.3	53.0	31.0	62.3	25.0	54.8	39.0	69.3	27.5	59.3	33.2	70.8	28.5	60.4
CognitionCapturer	31.4	79.7	31.4	77.8	38.2	**85.7**	40.4	**85.8**	24.4	66.3	34.8	78.8	34.7	**81.0**	48.1	**88.6**	37.4	79.4	35.6	79.3	35.6	80.2
VE-SDN	32.6	63.7	34.4	69.9	38.7	73.5	39.8	72.0	29.4	58.6	34.5	68.8	34.5	68.3	49.3	79.8	39.0	69.6	39.8	75.3	37.2	69.9
PinVC	32.6	77.9	32.5	78.9	35.0	85.4	39.8	85.1	26.5	73.4	34.7	78.2	36.5	80.4	48.3	83.7	37.8	84.0	36.6	80.7	38.9	80.8
UBP	41.0	70.5	48.5	84.0	50.0	81.0	50.0	79.5	43.0	72.5	**58.0**	83.0	47.0	79.0	56.0	86.5	44.5	76.5	61.5	88.5	50.0	80.1
CFBA	**52.0**	**84.0**	**50.0**	**85.0**	**52.0**	85.5	**53.5**	82.5	**48.5**	**74.5**	**58.0**	**87.5**	**51.5**	78.5	**60.0**	87.0	**56.5**	**85.0**	**63.0**	**90.5**	**54.5**	**84.0**
Inter-subject																						
BraVL	2.3	8.0	1.5	6.3	1.4	5.9	1.7	6.7	1.5	5.6	1.8	7.2	2.1	8.1	2.2	7.6	1.6	6.4	2.3	8.5	1.8	7.0
NICE	7.6	22.8	5.9	20.5	6.0	22.3	6.3	20.7	4.4	18.3	5.6	22.2	5.6	19.7	6.3	22.0	5.7	17.6	8.4	28.3	6.2	21.4
NICE-SA	7.0	22.6	6.6	23.2	7.5	23.7	5.4	21.4	6.4	22.2	7.5	22.5	3.8	19.1	8.5	24.4	7.4	22.3	9.8	29.6	7.0	23.1
NICE-GA	5.9	21.4	6.4	22.7	5.5	20.1	6.1	21.0	4.7	19.5	6.2	22.5	5.9	19.1	7.3	25.3	4.8	18.3	6.2	26.3	5.9	21.6
Neural-MCRL	10.1	26.1	9.0	22.1	5.8	18.2	8.6	21.0	5.1	15.5	7.9	22.0	7.3	19.9	10.6	26.6	10.6	24.7	12.1	28.2	8.7	22.4
ATM-S	10.5	26.8	7.1	24.8	11.9	33.8	14.7	39.4	7.0	23.9	11.1	35.8	**16.1**	**43.5**	15.0	40.3	4.9	22.7	**20.5**	46.5	11.8	33.7
MB2C	10.5	28.2	11.3	32.8	8.8	27.7	13.7	33.5	10.7	27.5	12.2	33.2	11.5	31.8	12.0	32.2	12.2	31.3	16.2	42.2	11.9	32.0
UBP	11.5	29.0	13.5	40.0	6.5	25.5	13.0	33.0	9.0	33.5	13.5	30.0	10.5	25.0	10.0	32.5	15.5	39.5	17.0	42.5	12.0	33.1
CFBA	**13.5**	**44.5**	**24.5**	**52.0**	**13.5**	**39.5**	**16.0**	**46.0**	**20.0**	**45.0**	**18.5**	**43.0**	11.0	38.5	**17.5**	**47.5**	**21.5**	**51.5**	19.5	**57.0**	**17.5**	**46.4**

Table 6: Subject-wise 200-way zero-shot retrieval on THINGS-MEG. Values are percentages.

	Subject 1		Subject 2		Subject 3		Subject 4		Mean	
Method	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Intra-subject										
NICE	9.6	27.8	18.5	47.8	14.2	41.6	9.0	26.6	12.8	36.0
NICE-SA	9.8	28.1	18.6	46.4	10.5	38.4	11.7	27.2	12.7	35.0
NICE-GA	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.3	42.3
UBP	**13.0**	34.0	**49.5**	**83.0**	26.5	**65.0**	17.0	42.5	26.5	56.125
CFBA	**13.0**	**37.0**	47.0	82.5	**31.0**	62.5	**21.5**	**46.0**	**28.125**	**57.0**
Inter-subject										
UBP	**2.5**	5.0	3.0	15.0	1.5	10.0	1.5	9.0	2.125	9.75
CFBA	1.5	**7.5**	**5.0**	**22.0**	**4.5**	**18.0**	**4.0**	**17.0**	**3.75**	**16.125**

Table 7: 200-way zero-shot classification on THINGS-EEG and THINGS-MEG. Values are percentages.

Dataset	Method	Accuracy	
		Top-1	Top-5
THINGS-EEG	BraVL	5.82	17.45
	ATM	7.40	20.60
	CFBA	**7.45**	**20.95**
THINGS-MEG	ATM	5.88	19.25
	CFBA	**6.63**	**19.50**

Reconstruction. To evaluate complementary conditioning, Table 8 compares low-level and feature-level reconstruction metrics. CFBA leads on five of seven metrics on THINGS-EEG and all seven on THINGS-MEG. Its EEG PixCorr and SSIM reach 0.1944 and 0.4862, respectively. Fig 7 compares the stimulus, transformed target, original-image baseline, and CFBA reconstruction. The examples illustrate improved object structure and contextual fidelity with complementary conditioning. The transformed target retains selected boundaries while simplifying context, and its neural representation supplies information that complements the original-image condition. The remaining EEG metric

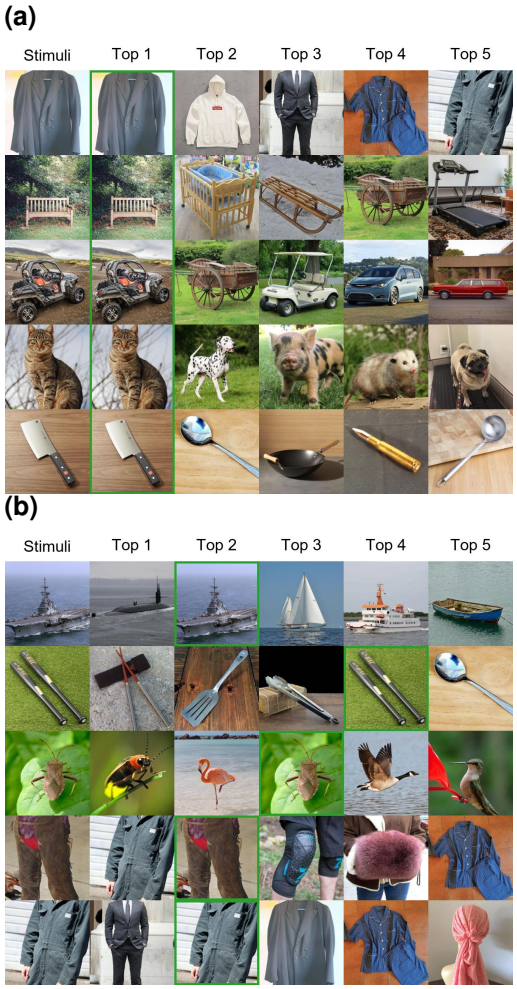

Figure 6: EEG retrieval examples. **(a)** Top-1 predictions. **(b)** Top-5 candidate rankings.

Table 8: Reconstruction performance on THINGS-EEG and THINGS-MEG.

		Low-level		High-level				
Dataset	Method	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
THINGS-EEG	ATM	0.1600	0.3450	0.7760	0.8660	0.7340	0.7860	0.5820
	PinVC	0.1890	0.3900	0.7940	0.8810	**0.8060**	0.8090	0.5630
	CognitionCapturer	0.1500	0.3470	0.7540	0.6230	0.6690	0.7150	0.5900
	SeeEEG	0.1300	0.3440	0.8090	0.9040	0.7840	**0.8690**	0.5390
	CFBA	**0.1944**	**0.4862**	**0.8398**	**0.9041**	0.7871	0.8319	**0.5223**
THINGS-MEG	ATM	0.1040	0.3400	0.6130	0.6720	0.6190	0.6030	0.6510
	CFBA	**0.1420**	**0.4118**	**0.8339**	**0.8993**	**0.8200**	**0.8146**	**0.5434**

differences show that improvements in structural fidelity do not imply uniform gains in every feature space.

A.3 Temporal, Spatial, and Spectral Analysis

To examine the information available under different recording configurations, we vary temporal windows, channel subsets, and frequency bands. Figs 8–10 present the visual comparisons before the quantitative tables.



Figure 7: Qualitative reconstruction comparison. Rows show stimuli, CFBA visual targets, original-image baseline reconstructions, and CFBA reconstructions.

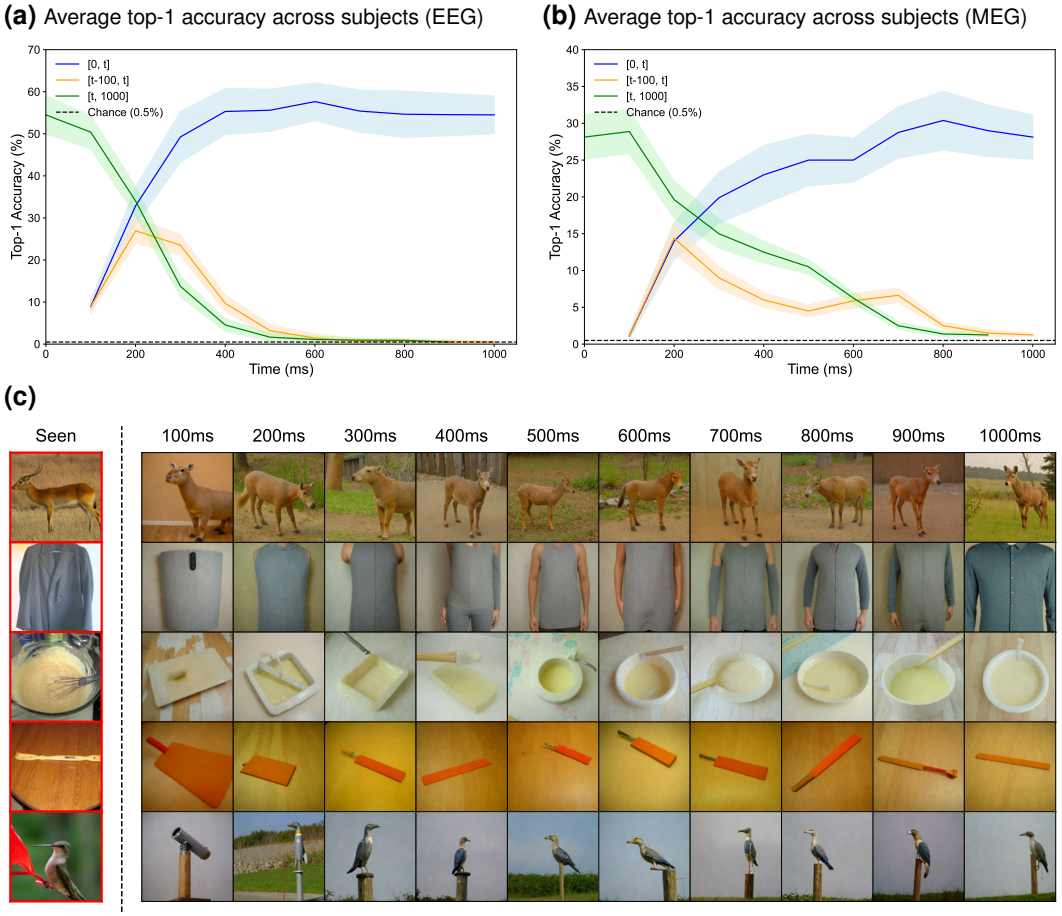

Figure 8: Temporal analysis. **(a, b)** Retrieval with different EEG and MEG windows. **(c)** Reconstructions with cumulative EEG windows.

Temporal dynamics. Fig 8 compares retrieval across cumulative EEG and MEG windows and illustrates reconstruction as the EEG interval expands. Longer early windows improve retrieval and make reconstructed objects more recognizable, with smaller changes at later endpoints.

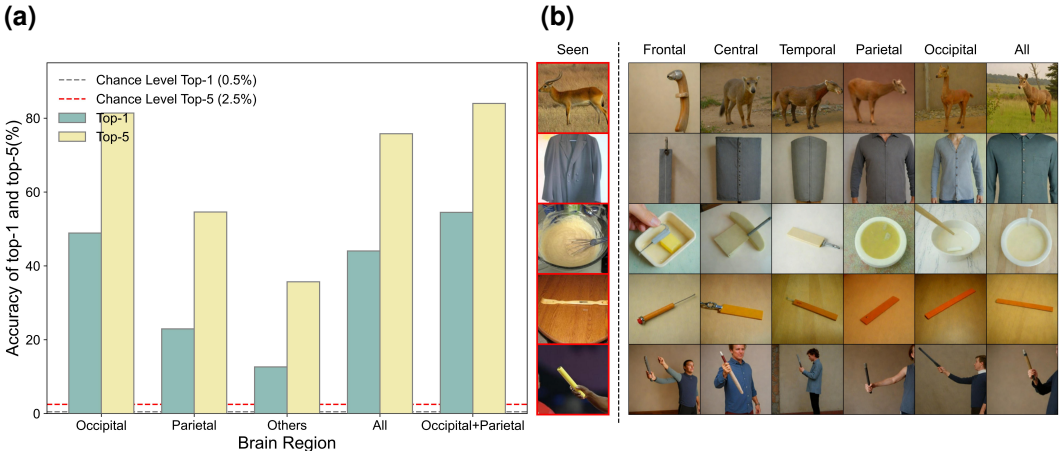

Figure 9: EEG channel analysis. **(a)** Retrieval accuracy. **(b)** Reconstructions with different channel subsets.

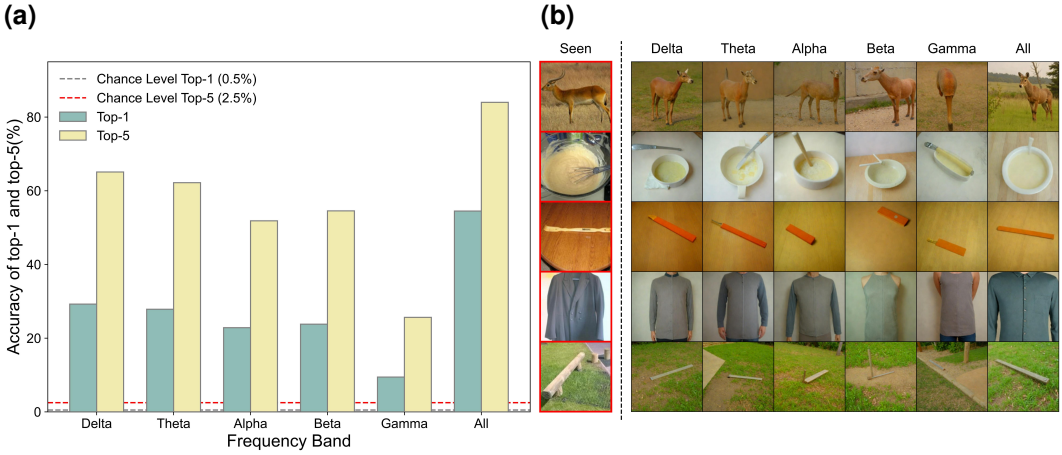

Figure 10: EEG frequency analysis. **(a)** Retrieval accuracy. **(b)** Reconstructions from individual bands and the full passband.

Channel subsets. Fig 9 compares EEG channel configurations. Occipital–parietal electrodes provide the strongest retrieval, and the reconstruction examples show useful visual information in posterior recordings.

Frequency bands. Fig 10 compares individual EEG bands with the full passband. Delta and theta support stronger retrieval than the other individual bands, while the full passband performs best. Reconstruction examples show how the selected band affects object appearance.

Quantitative temporal results. Table 9 evaluates non-overlapping 100-ms windows. The 100–200 ms interval gives the best retrieval and classification, while reconstruction metrics peak at different intervals. Table 10 evaluates cumulative windows. The 0–600 ms interval gives the highest retrieval and classification accuracy, although the best reconstruction window depends on the metric. Table 11 delays the window onset while fixing the endpoint at 1,000 ms. Removing early responses substantially reduces discriminative accuracy. Together, these comparisons support retaining early post-stimulus information. The 0–600 ms window reaches 57.65% intra-subject Top-1 retrieval, exceeding the full 0–1,000 ms window. However, SSIM peaks with the full window, showing that the most useful integration interval depends on whether the objective is discrimination or reconstruction.

Quantitative spatial results. Table 12 shows that occipital–parietal channels outperform each other tested subset for retrieval and classification. Table 13 reports reconstruction. This subset leads on

Table 9: Decoding from non-overlapping 100-ms EEG windows. Accuracies are percentages.

Window (ms)	Retrieval				Classification				Reconstruction						
	Intra-subject		Inter-subject		Intra-subject		Inter-subject		Low-level		High-level				
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
0–100	8.90	24.75	3.75	13.55	1.20	6.05	0.30	2.05	0.1651	0.4671	0.8030	0.8554	0.7121	0.7609	0.5582
100–200	**26.95**	**58.90**	**11.00**	**36.10**	**5.05**	**14.65**	**1.50**	**4.75**	**0.1835**	0.4831	**0.8088**	**0.8835**	**0.7820**	0.7892	0.5463
200–300	23.55	56.35	10.35	21.95	4.10	11.80	1.30	4.55	0.1832	0.4902	0.8044	0.8832	0.7400	**0.8124**	**0.5390**
300–400	9.70	29.50	7.20	18.95	4.15	10.95	1.05	3.95	0.1645	**0.5087**	0.7694	0.8278	0.7111	0.7673	0.5468
400–500	3.20	13.85	5.70	17.50	2.25	6.65	0.95	1.80	0.1502	0.5004	0.7598	0.8143	0.6836	0.7292	0.5598
500–600	1.45	5.50	4.35	14.00	2.10	5.70	0.80	1.50	0.1598	0.4998	0.7559	0.7995	0.6814	0.7104	0.5633
600–700	0.65	4.25	2.15	10.50	1.95	3.25	0.80	1.30	0.1471	0.4946	0.7732	0.8342	0.7121	0.7367	0.5552
700–800	0.75	3.20	0.90	7.10	1.35	2.50	0.55	0.95	0.1570	0.5030	0.7552	0.8292	0.6966	0.7376	0.5572
800–900	0.65	3.15	0.12	3.30	1.20	2.10	0.50	0.70	0.1454	0.5037	0.7247	0.7747	0.6510	0.6944	0.5724
900–1000	0.50	3.15	0.09	1.25	0.85	1.30	0.15	0.65	0.1546	0.4920	0.7665	0.8268	0.6882	0.7404	0.5612

Table 10: Decoding from cumulative EEG windows. Accuracies are percentages.

Window (ms)	Retrieval				Classification				Reconstruction						
	Intra-subject		Inter-subject		Intra-subject		Inter-subject		Low-level		High-level				
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
0–100	8.90	24.75	3.75	13.55	1.20	6.05	0.30	2.05	0.1651	0.4671	0.8030	0.8554	0.7121	0.7609	0.5582
0–200	32.85	65.55	12.75	36.75	4.25	16.00	1.15	5.40	0.1897	0.4720	0.8202	0.8801	0.7839	0.8172	0.5286
0–300	49.25	80.50	16.25	45.00	6.35	19.60	1.65	6.60	0.1885	0.4848	0.8324	0.9030	0.7590	0.8295	0.5243
0–400	55.30	84.95	17.75	46.75	7.10	20.75	1.85	7.00	0.1930	0.4821	0.8247	0.9072	0.7560	0.8207	0.5241
0–500	55.60	85.45	17.80	47.05	7.50	21.00	2.05	7.20	0.1984	0.4825	0.8397	**0.9137**	**0.8016**	0.8354	0.5231
0–600	**57.65**	**85.85**	**18.25**	**47.20**	**7.95**	**21.25**	**2.10**	**7.25**	**0.1999**	0.4848	**0.8398**	0.9090	0.7937	**0.8577**	**0.5181**
0–700	55.40	84.85	17.75	46.80	7.75	21.05	1.95	7.15	0.1927	0.4844	0.8177	0.9001	0.7633	0.8247	0.5255
0–800	54.65	84.70	17.60	46.75	7.55	20.95	1.70	7.00	0.1942	0.4829	0.8131	0.8942	0.7687	0.8509	0.5240
0–900	54.55	84.25	17.56	46.55	7.40	21.00	1.85	7.15	0.1938	0.4838	0.8256	0.8941	0.7986	0.8351	0.5225
0–1000	54.50	84.00	17.55	46.45	7.45	20.95	1.95	7.05	0.1944	**0.4862**	**0.8398**	0.9041	0.7871	0.8319	0.5223

Table 11: Decoding from EEG windows with delayed onset. Accuracies are percentages.

Window (ms)	Retrieval				Classification				Reconstruction						
	Intra-subject		Inter-subject		Intra-subject		Inter-subject		Low-level		High-level				
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
0–1000	**54.50**	**84.00**	**17.55**	**46.45**	**7.45**	**20.95**	**1.95**	**7.05**	**0.1944**	0.4862	**0.8398**	**0.9041**	**0.7871**	0.8319	**0.5223**
100–1000	50.40	83.40	16.75	42.00	7.00	18.45	1.90	6.85	0.1919	0.4856	0.8220	0.9001	0.7856	0.8327	0.5285
200–1000	34.05	67.20	13.05	38.20	6.45	16.20	1.65	5.90	0.1849	0.4840	0.7987	0.8954	0.7805	**0.8395**	0.5230
300–1000	13.75	37.85	8.40	21.90	4.10	12.60	1.30	4.45	0.1738	**0.4982**	0.7862	0.8585	0.7400	0.7777	0.5389
400–1000	4.55	16.10	6.30	12.75	2.70	5.75	1.05	2.90	0.1624	0.4909	0.7797	0.8382	0.6844	0.7720	0.5493
500–1000	1.65	6.65	3.00	9.50	2.45	5.10	0.85	2.65	0.1378	0.4929	0.7568	0.8125	0.7053	0.7379	0.5651
600–1000	1.10	4.10	2.60	8.35	2.10	4.45	0.65	2.10	0.1527	0.4970	0.7374	0.8167	0.7138	0.7522	0.5508
700–1000	0.95	3.55	1.47	6.50	1.85	3.40	0.40	1.75	0.1595	0.4933	0.7627	0.8371	0.7223	0.7720	0.5506
800–1000	0.90	2.50	1.41	3.35	1.30	2.50	0.20	1.10	0.1422	0.4930	0.7707	0.8114	0.6839	0.7594	0.5557
900–1000	0.50	3.15	0.09	1.25	0.85	1.30	0.15	0.65	0.1546	0.4920	0.7665	0.8268	0.6882	0.7404	0.5612

PixCorr, both AlexNet metrics, and SwAV, while other subsets perform better on SSIM, Inception, or CLIP. Adding all available channels does not improve retrieval or classification over the occipital–parietal subset. This supports selecting informative posterior channels rather than assuming that more recorded channels necessarily improve alignment. The reconstruction differences nevertheless indicate that useful information is distributed across channel groups and depends on the evaluation metric.

Quantitative spectral results. Table 14 shows that the full passband performs best on all reported metrics. Delta provides the strongest retrieval and classification among the individual bands, while the

Table 12: Retrieval and classification with EEG channel subsets. Accuracies are percentages. O and P denote occipital and parietal EEG channels.

Channel subset	Retrieval				Classification			
	Intra-subject		Inter-subject		Intra-subject		Inter-subject	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
O	48.90	81.40	16.25	44.35	6.95	19.70	1.80	6.80
P	22.95	54.60	11.50	34.75	4.05	16.30	1.35	4.60
O+P	**54.50**	**84.00**	**17.55**	**46.45**	**7.45**	**20.95**	**1.95**	**7.05**
Others	12.65	35.70	7.95	26.50	2.60	11.65	0.50	2.60
All	44.05	75.80	14.75	42.95	6.00	17.70	1.75	5.55

Table 13: Reconstruction with EEG channel subsets. O+P denotes occipital and parietal EEG channels.

Channel subset	Low-level		High-level				
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
Central	0.1662	0.4707	0.7909	0.8543	0.7815	0.8098	0.5391
Frontal	0.1554	**0.4953**	0.7687	0.8312	0.7005	0.7510	0.5536
Occipital	0.1848	0.4808	0.8257	0.8979	0.7617	0.8089	0.5375
Parietal	0.1664	0.4814	0.7831	0.8808	**0.7873**	0.8060	0.5343
Temporal	0.1707	0.4726	0.7794	0.8584	0.7602	0.8060	0.5486
Others	0.1626	0.4648	0.7697	0.8763	0.7669	0.8192	0.5428
O+P	**0.1944**	0.4862	**0.8398**	**0.9041**	0.7871	0.8319	**0.5223**
All	0.1856	0.4834	0.8173	0.8928	0.7793	**0.8353**	0.5257

Table 14: Decoding from individual EEG bands and the full passband. Accuracies are percentages.

Band	Retrieval				Classification				Reconstruction						
	Intra-subject		Inter-subject		Intra-subject		Inter-subject		Low-level		High-level				
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
Delta	29.25	65.10	12.40	35.05	5.70	17.30	1.75	6.20	0.1857	0.4725	0.8205	0.8819	0.7631	0.8245	0.5470
Theta	27.85	62.20	11.70	27.45	5.25	16.85	1.60	5.65	0.1707	0.4638	0.8266	0.8651	0.7605	0.8186	0.5500
Alpha	22.85	51.85	10.20	22.15	3.90	11.65	0.95	3.10	0.1720	0.4602	0.8279	0.8623	0.7503	0.8049	0.5510
Beta	23.80	54.55	10.60	23.90	4.25	12.40	1.05	3.75	0.1684	0.4617	0.8038	0.8579	0.7639	0.8059	0.5654
Gamma	9.45	25.65	7.35	12.35	1.95	4.65	0.55	1.65	0.1484	0.4457	0.7828	0.8365	0.7316	0.7860	0.5837
Full passband	**54.50**	**84.00**	**17.55**	**46.45**	**7.45**	**20.95**	**1.95**	**7.05**	**0.1944**	**0.4862**	**0.8398**	**0.9041**	**0.7871**	**0.8319**	**0.5223**

best individual band varies across reconstruction metrics. Delta alone remains well below the full passband in retrieval, despite being the strongest individual band for discrimination. Conversely, Alpha has the highest AlexNet(2) score among individual bands. These differences indicate that no single frequency band captures all information useful for decoding and support retaining complementary spectral content.

A.4 Hyperparameter Analysis

A.4.1 Contour-Mask Thickness

To assess how much image content should be protected from smoothing, we vary contour-mask thickness from one to ten pixels. Table 15 shows that a two-pixel mask achieves the best results across all reported metrics. A one-pixel mask performs worse, suggesting that a very narrow protected region provides insufficient support around the estimated boundaries. Increasing the thickness beyond two pixels also reduces performance, with fluctuations across settings. Wider masks preserve more neighboring texture and therefore weaken the intended simplification. The results favor a narrow contour region that retains useful boundary information while allowing the surrounding context to be smoothed.

Table 15: Contour-mask thickness on THINGS-EEG. Accuracies are percentages.

κ (px)	Retrieval				Classification				Reconstruction						
	Intra-subject		Inter-subject		Intra-subject		Inter-subject		Low-level		High-level				
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
1	50.50	83.30	17.50	45.45	6.20	20.90	1.40	5.40	0.1685	0.4582	0.8115	0.8758	0.7569	0.8032	0.5507
2	**54.50**	**84.00**	**17.55**	**46.45**	**7.45**	**20.95**	**1.95**	**7.05**	**0.1944**	**0.4862**	**0.8398**	**0.9041**	**0.7871**	**0.8319**	**0.5223**
3	53.10	83.60	16.40	44.50	6.50	20.50	1.40	5.50	0.1897	0.4745	0.8195	0.8822	0.7680	0.8118	0.5353
4	51.25	82.80	16.50	45.20	6.50	20.45	1.50	5.30	0.1852	0.4698	0.8142	0.8765	0.7627	0.8065	0.5408
5	51.25	82.80	16.50	45.25	5.55	19.20	1.20	5.30	0.1861	0.4712	0.8158	0.8781	0.7643	0.8081	0.5392
6	51.15	83.55	16.30	44.85	5.50	19.50	1.50	4.90	0.1837	0.4675	0.8126	0.8749	0.7611	0.8049	0.5425
7	51.15	83.55	15.70	43.55	6.10	19.95	1.60	5.30	0.1845	0.4689	0.8139	0.8762	0.7624	0.8062	0.5411
8	50.40	82.55	15.85	43.95	6.10	20.60	1.80	4.70	0.1793	0.4631	0.8087	0.8708	0.7572	0.8010	0.5463
9	50.40	82.55	15.25	44.15	6.30	20.60	1.75	5.00	0.1778	0.4615	0.8072	0.8693	0.7557	0.7995	0.5478
10	50.50	83.30	15.15	44.45	6.30	20.10	1.60	4.50	0.1659	0.4567	0.8103	0.8746	0.7541	0.8004	0.5519

Table 16: Gaussian kernel size on THINGS-EEG. Accuracies are percentages.

k_B (px)	Retrieval				Classification				Reconstruction						
	Intra-subject		Inter-subject		Intra-subject		Inter-subject		Low-level		High-level				
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
33	53.80	83.25	16.55	46.35	6.65	19.25	1.80	5.60	0.1721	0.4615	0.8147	0.8789	0.7602	0.8065	0.5478
35	54.20	83.70	17.00	45.75	6.50	19.70	1.70	6.15	0.1786	0.4683	0.8212	0.8854	0.7667	0.8130	0.5413
37	53.80	**84.00**	17.10	45.90	6.15	19.55	1.50	5.90	0.1753	0.4651	0.8179	0.8821	0.7634	0.8097	0.5446
39	53.85	83.50	17.25	46.10	6.95	20.10	1.60	6.50	0.1827	0.4724	0.8253	0.8895	0.7708	0.8171	0.5372
41	54.25	83.60	17.05	45.40	7.25	19.85	1.90	6.80	0.1802	0.4699	0.8228	0.8870	0.7683	0.8146	0.5397
43	54.05	83.25	17.35	46.25	7.10	20.70	1.75	6.95	0.1875	0.4796	0.8321	0.8963	0.7776	0.8239	0.5304
45	**54.50**	**84.00**	**17.55**	**46.45**	**7.45**	**20.95**	**1.95**	**7.05**	**0.1944**	**0.4862**	**0.8398**	**0.9041**	**0.7871**	**0.8319**	**0.5223**
47	53.80	83.60	17.10	46.10	6.90	20.90	1.65	6.70	0.1768	0.4666	0.8194	0.8836	0.7649	0.8112	0.5431
49	53.15	83.20	17.30	45.80	6.65	20.55	1.60	6.40	0.1735	0.4629	0.8161	0.8803	0.7616	0.8079	0.5464
51	52.75	83.80	17.00	45.80	6.25	19.50	1.70	5.95	0.1692	0.4586	0.8118	0.8760	0.7573	0.8036	0.5507
53	52.80	83.75	16.95	45.70	6.30	19.65	1.50	5.80	0.1657	0.4551	0.8083	0.8725	0.7538	0.8001	0.5542

A.4.2 Gaussian Kernel Size

To evaluate context smoothing, we vary the Gaussian kernel side length from 33 to 53 pixels and determine σ using the rule in Appendix A.1. This experiment changes both spatial support and smoothing strength. Table 16 favors a 45×45 kernel, which leads across the reported metrics and ties with the 37-pixel kernel for intra-subject Top-5 retrieval. Smaller kernels provide less effective reconstruction, while larger kernels reduce both discriminative and reconstruction performance. These results suggest that context should be simplified sufficiently to improve supervision without excessively suppressing useful image information.

A.4.3 Contour Threshold

To assess contour selection, we vary the threshold applied to peak-normalized EGNet responses from 0.1 to 0.9. Table 17 shows that 0.4 achieves the best results across all reported metrics. Lower thresholds retain weaker responses and can protect irrelevant detail, reducing the selectivity of the prior. Higher thresholds remove more predicted contour pixels and can weaken boundary preservation. Performance fluctuates away from the optimum, so the results support an intermediate selection threshold rather than a monotonic preference for denser or sparser masks.

A.4.4 Minimum Contour Length

To evaluate fragment removal, we vary the minimum retained contour length from 50 to 160 pixels. Table 18 favors 100 pixels across the reported metrics, with intra-subject Top-5 retrieval tied at 120 pixels. Shorter thresholds retain more fragments and produce weaker reconstruction results. Higher thresholds do not improve performance and can remove useful boundaries along with short fragments.

Table 17: Contour threshold on THINGS-EEG. Accuracies are percentages.

Threshold	Retrieval				Classification				Reconstruction						
	Intra-subject		Inter-subject		Intra-subject		Inter-subject		Low-level		High-level				
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
0.1	49.40	82.00	15.20	44.00	5.50	20.10	1.30	5.85	0.1672	0.4571	0.8102	0.8745	0.7553	0.8016	0.5527
0.2	52.70	82.60	15.70	44.50	5.95	20.65	1.25	6.00	0.1815	0.4723	0.8256	0.8901	0.7708	0.8172	0.5371
0.3	51.60	83.05	16.70	45.90	6.50	20.55	1.75	6.65	0.1889	0.4798	0.8331	0.8976	0.7784	0.8248	0.5296
0.4	**54.50**	**84.00**	**17.55**	**46.45**	**7.45**	**20.95**	**1.95**	**7.05**	**0.1944**	**0.4862**	**0.8398**	**0.9041**	**0.7871**	**0.8319**	**0.5223**
0.5	52.50	82.65	16.40	45.75	7.20	20.50	1.80	6.55	0.1853	0.4765	0.8297	0.8942	0.7751	0.8215	0.5330
0.6	52.50	82.30	15.85	45.05	6.35	20.20	1.90	6.70	0.1796	0.4702	0.8235	0.8878	0.7687	0.8151	0.5394
0.7	52.00	82.55	15.50	44.80	6.65	19.95	1.55	6.25	0.1821	0.4734	0.8268	0.8913	0.7722	0.8186	0.5359
0.8	50.75	82.00	15.90	44.80	6.25	20.55	1.40	5.85	0.1714	0.4627	0.8154	0.8796	0.7605	0.8069	0.5476
0.9	50.55	81.25	15.70	43.95	6.45	20.70	1.65	5.00	0.1648	0.4556	0.8087	0.8729	0.7538	0.8002	0.5543

Table 18: Minimum contour length on THINGS-EEG. Accuracies are percentages.

δ_{len} (px)	Retrieval				Classification				Reconstruction						
	Intra-subject		Inter-subject		Intra-subject		Inter-subject		Low-level		High-level				
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
50	53.15	83.00	13.80	37.50	6.45	20.40	1.50	5.10	0.1652	0.4563	0.8095	0.8738	0.7558	0.8011	0.5531
60	53.50	83.10	17.00	45.15	6.40	20.45	1.80	6.20	0.1827	0.4735	0.8268	0.8913	0.7720	0.8184	0.5359
70	53.65	83.55	16.80	45.45	6.90	20.80	1.50	4.60	0.1859	0.4768	0.8301	0.8946	0.7753	0.8217	0.5326
80	54.00	83.70	17.25	45.75	7.00	20.75	1.00	4.00	0.1883	0.4792	0.8325	0.8970	0.7777	0.8241	0.5302
90	54.10	83.55	17.10	45.90	7.30	20.90	1.80	5.20	0.1916	0.4825	0.8358	0.9003	0.7810	0.8274	0.5269
100	**54.50**	**84.00**	**17.55**	**46.45**	**7.45**	**20.95**	**1.95**	**7.05**	**0.1944**	**0.4862**	**0.8398**	**0.9041**	**0.7871**	**0.8319**	**0.5223**
110	54.05	83.70	16.90	46.40	7.10	20.20	1.20	5.10	0.1897	0.4806	0.8339	0.8984	0.7791	0.8255	0.5288
120	54.05	**84.00**	17.15	46.30	6.55	20.50	1.30	5.20	0.1921	0.4830	0.8363	0.9008	0.7815	0.8279	0.5264
130	54.20	83.80	16.65	46.30	6.60	19.80	1.70	5.20	0.1875	0.4784	0.8317	0.8962	0.7769	0.8233	0.5310
140	54.15	83.75	16.70	45.95	6.55	19.55	1.50	5.80	0.1889	0.4798	0.8331	0.8976	0.7783	0.8247	0.5296
150	54.15	83.50	16.80	46.00	6.95	20.10	1.33	5.44	0.1867	0.4776	0.8309	0.8954	0.7761	0.8225	0.5318
160	54.00	83.10	17.15	45.70	6.50	19.75	1.70	5.70	0.1879	0.4788	0.8321	0.8966	0.7773	0.8237	0.5306

Table 19: Neural–image encoder pairs on THINGS-EEG. Accuracies are percentages.

Encoder pair	Retrieval				Classification			
	Intra-subject		Inter-subject		Intra-subject		Inter-subject	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
ATM–ViT-H-14	40.20	73.70	13.30	37.50	6.00	19.40	1.40	4.90
RPE–RN50	**54.50**	**84.00**	**17.50**	**46.40**	**7.45**	**20.95**	**2.00**	**7.10**

The selected threshold therefore balances contour selectivity with structural coverage. This optimum applies to the image resolution and contour-processing pipeline used in these experiments.

A.5 Additional Encoder Comparison

Table 19 compares RPE–RN50 with ATM–ViT-H-14. RPE–RN50 leads across retrieval and classification in both subject settings, consistent with the heatmap in Fig. 4. Its intra-subject Top-1 retrieval reaches 54.50%, compared with 40.20% for ATM–ViT-H-14. However, the reconstruction comparison in Table 4 favors ATM–ViT-H-14 for primary conditioning when paired with RPE–RN50 auxiliary conditioning. Together, these results support assigning encoder pairs according to their roles rather than using one pair throughout the pipeline.