MUCCA: MULTILEVEL HUMAN COLOR REPRESENTATION-GUIDED COMPLEMENTARY ALIGNMENT FOR NEURAL VISUAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural visual decoding links brain activity to visual content and supports the development of brain–computer interfaces. However, existing approaches to visual supervision adaptation do not explicitly account for the multilevel organization of color processing in human vision, potentially limiting neural–visual alignment. To address this limitation, we propose Multilevel Human Color Representation-guided Complementary Alignment (MuCCA) for visual decoding. MuCCA constructs RGB, LMS, DKL-inspired, and CIELUV views of each stimulus to provide supervision informed by cone responses, opponent relationships, and perceptual color attributes. Four independently trained Residual Projection Encoders align neural recordings with frozen CLIP ResNet-50 features of these views through symmetric contrastive learning. A separate original-image branch jointly trains an ATM neural encoder and a ViT-H-14 image encoder for generation. For retrieval and classification, the four RPE branches produce individual rankings, and their candidate union quantifies complementary decoding successes. For reconstruction, a diffusion prior predicts the primary condition from the ATM embedding, while three linear projections map the LMS-, DKL-, and CIELUV-supervised neural embeddings into the same conditioning space. Feature supervision and cross-space consistency regularize these projections. Four IP-Adapters integrate these conditions into a pretrained SDXL generator for reconstruction. Extensive experiments on THINGS-EEG and THINGS-MEG demonstrate that MuCCA outperforms state-of-the-art methods across retrieval, classification, and reconstruction. The code is available at `https://anonymous.4open.science/r/MuCCA-8471-842D`.

1 INTRODUCTION

Neural visual decoding identifies or reconstructs viewed stimuli from the associated brain activity (Kay et al., 2008; Miyawaki et al., 2008), providing insights into visual representations and supporting brain–computer interfaces. Latent diffusion models enable detailed reconstruction from functional magnetic resonance imaging (fMRI) (Takagi & Nishimoto, 2023), but its slow hemodynamic response limits the temporal precision of visual decoding (Benchetrit et al., 2024b). 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 therefore increased interest in EEG- and MEG-based classification, retrieval, and reconstruction (Song et al., 2024; Li et al., 2024; Benchetrit et al., 2024b).

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

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

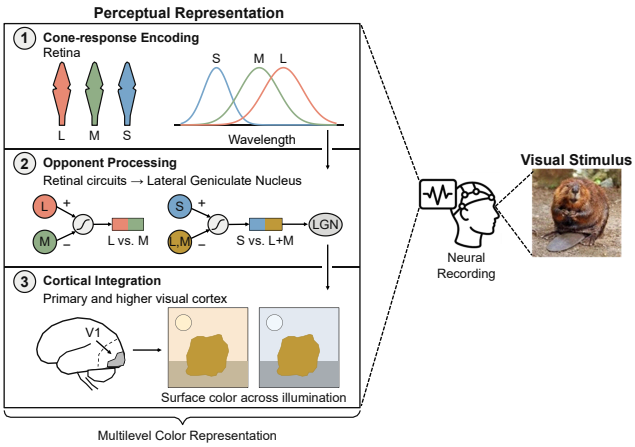

Figure 1: Multilevel color representation in human vision.

Recent work addresses this mismatch by adapting visual supervision. The uncertainty-aware blur prior (UBP) (Wu et al., 2025) adjusts stimulus detail using cross-modal uncertainty to improve retrieval. Human-aligned image encoders (Rajabi et al., 2025) incorporate perceptual judgments into visual targets, improving retrieval across encoders and recording modalities. ShrinkAdapter (Wu et al., 2026) adapts the visual teacher through a task-guided bottleneck. These findings establish target design as an effective means of improving neural–visual alignment. Psychophysical and neurophysiological evidence further shows that color perception emerges through successive transformations of retinal input (Fig. 1). Long-, middle-, and short-wavelength-sensitive cones first convert incident light into neural responses according to their spectral sensitivities (Stockman & Sharpe, 2000; Stockman et al., 1999). Retinal circuits combine these responses into opponent signals, which reach the visual cortex through the lateral geniculate nucleus (Derrington et al., 1984; Gegenfurtner, 2003). Primary and higher visual cortex further transform and integrate chromatic signals with spatial context to support surface-color perception under varying illumination (Gegenfurtner, 2003). This progression connects receptor encoding, opponent processing, and perceptual interpretation. However, existing approaches to visual supervision adaptation do not explicitly incorporate this multilevel organization, potentially limiting their ability to capture complementary color representations relevant to neural decoding.

To address this limitation, we propose **Mu**ltilevel Human **C**olor Representation-guided **C**omplementary **A**lignment (MuCCA), which translates multilevel human color representations into complementary supervision targets. Four independent Residual Projection Encoders (RPEs) align EEG or MEG recordings with frozen CLIP ResNet-50 features of RGB, LMS, DKL-inspired, and CIELUV views. Their rankings support retrieval and classification, and the candidate union captures complementary successes. A separate ATM–ViT-H-14 branch learns original-image alignment for reconstruction. Its diffusion prior provides the primary condition, while three linear projections map the transformed-view embeddings into auxiliary conditions. Feature supervision and cross-space consistency guide these projections, and four IP-Adapters jointly condition pretrained SDXL. Our contributions are threefold.

- To our knowledge, we are the first to jointly use computational descriptions of cone responses, opponent relationships, and perceptual color attributes as complementary supervision for EEG and MEG visual decoding.
- We propose MuCCA, integrating independent color-specific alignment with a primary diffusion prior and consistency-regularized auxiliary projections for complementary visual decoding.
- Extensive experiments on THINGS-EEG and THINGS-MEG demonstrate that MuCCA outperforms state-of-the-art methods in most evaluated settings across retrieval, classification, and reconstruction.

2 Related Work

2.1 Neural Visual Decoding and Visual Supervision

Early fMRI studies demonstrated image identification and reconstruction (Kay et al., 2008; Miyawaki et al., 2008), with latent diffusion 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) develop contrastive alignment, generative decoding, and multimodal supervision,

respectively. MEG decoding likewise combines pretrained visual features with generation and examines the effects of representation choice and response timing (Benchetrit et al., 2024b). Subsequent work adapts visual targets through uncertainty-dependent image processing in UBP (Wu et al., 2025), human-aligned encoders (Rajabi et al., 2025), and task-adaptive teaching in ShrinkAdapter (Wu et al., 2026). These improvements show that both image content and feature geometry influence neural–visual alignment. The joint contribution of complementary supervision informed by receptor encoding, opponent processing, and cortical integration remains insufficiently explored.

2.2 Multistage color processing and computational color representations

Color-coordinate transformations offer computational descriptions of different aspects of color processing. LMS coordinates approximate long-, middle-, and short-wavelength-sensitive cone responses (Stockman & Sharpe, 2000; Stockman et al., 1999). DKL coordinates organize stimulus variation along luminance and cone-opponent directions, emphasizing relationships among cone signals (Derrington et al., 1984). CIELUV describes lightness and chromatic attributes through L^*, u^*, and v^* in an approximately perceptually uniform space (Commission Internationale de l'Eclairage, n.d.). These complementary descriptions provide a basis for exploring visual supervision informed by receptor encoding, opponent processing, and perceptual color attributes.

3 Method

MuCCA comprises multilevel color representation construction, independent contrastive alignment, and complementary decoding (Fig. 2). Four RPE branches use RGB, LMS, DKL-inspired, and CIELUV supervision for retrieval and classification. A separate Adaptive Thinking Mapper (ATM)–ViT-H-14 branch provides the primary reconstruction condition through a diffusion prior, complemented by three projected embeddings from the transformed-view branches.

3.1 Problem Definition

Let $\mathcal{D} = \{(\mathbf{X}_i, I_i, y_i)\}_{i=1}^{N}$ contain N neural recordings, paired RGB stimuli, and category labels. Each recording $\mathbf{X}_i \in \mathbb{R}^{C_s \times T}$ contains T sampled time points from C_s EEG channels or MEG sensors. Each image $I_i \in [0,1]^{H \times W \times 3}$ has height H and width W. EEG and MEG are evaluated separately. Define the discriminative branch set $\mathcal{B} = \{R, L, D, U\}$ and auxiliary generation set $\mathcal{A} = \{L, D, U\}$. Their visual inputs are

$$I^R = I^G = I, \qquad I^L = \Phi_L(I), \qquad I^D = \Phi_D(I), \qquad I^U = \Phi_U(I). \tag{1}$$

Here, R, L, D, U index RGB, LMS, DKL-inspired, and CIELUV supervision, and G denotes the separate RGB generation branch. The functions Φ_b construct three-channel encoder inputs from the corresponding color coordinates. Given $\mathbf{X}$, retrieval ranks image candidates, classification ranks categories, and generation produces a reconstruction $\widehat{I}$.

3.2 Multilevel Color Representation Construction

Fig 2(a) constructs computational views motivated by cone responses, opponent relationships, and perceptual color attributes. All views describe the same stimulus and supply different inputs to the pretrained visual encoder.

Cone-response and opponent views. At pixel p, let $\mathbf{r}(p) = [R(p), G(p), B(p)]^\top$ contain sRGB values in $[0,1]$. Linearize each channel value a using

$$\Gamma^{-1}(a) = \begin{cases} a/12.92, & a \le 0.04045, \\ ((a+0.055)/1.055)^{2.4}, & a > 0.04045. \end{cases} \tag{2}$$

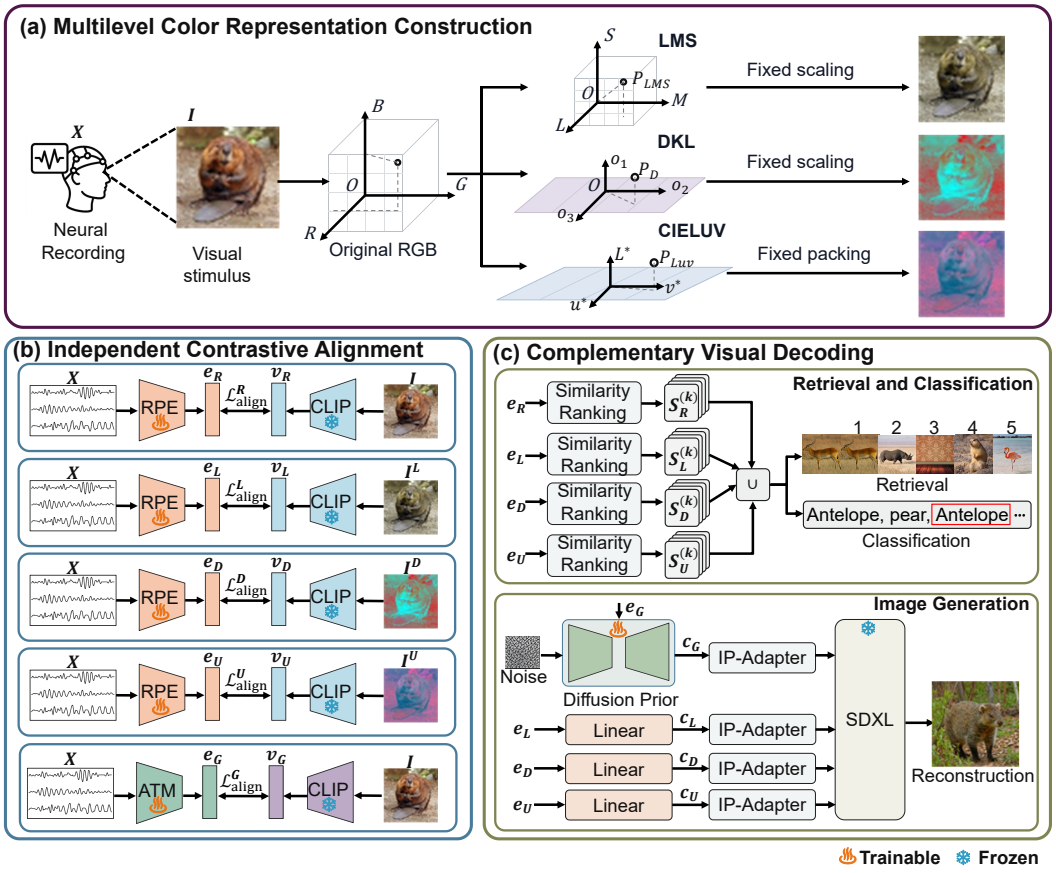

Figure 2: Overview of MuCCA. **(a)** Multilevel color representation construction. **(b)** Independent contrastive alignment. **(c)** Complementary visual decoding.

The fixed coordinate transformations are

$$\mathbf{x}(p) = \mathbf{M}_X \Gamma^{-1}(\mathbf{r}(p)), \quad \mathbf{M}_X = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix},$$

$$\mathbf{q}_L(p) = \mathbf{M}_L \mathbf{x}(p), \qquad \mathbf{M}_L = \begin{bmatrix} 0.3897 & 0.6890 & -0.0787 \\ -0.2298 & 1.1834 & 0.0464 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

$$\mathbf{q}_D(p) = \mathbf{M}_D \mathbf{q}_L(p), \qquad \mathbf{M}_D = \begin{bmatrix} 22.0454 & 22.0454 & 0 \\ 13.8336 & -32.4902 & 0 \\ -12.7279 & -12.7279 & 9.6048 \end{bmatrix}.$$

The vectors $\mathbf{x} = [X, Y, Z]^\top$, $\mathbf{q}_L = [\ell, m, s]^\top$, and $\mathbf{q}_D = [o_1, o_2, o_3]^\top$ represent tristimulus, cone-response-like, and opponent coordinates, respectively. The opponent rows encode $\ell + m$, a weighted ℓ–m difference, and s opposed to $\ell + m$. We call this fixed transform DKL-inspired to distinguish it from calibrated DKL coordinates, which specify the adapting background, cone contrasts, and stimulus calibration (Derrington et al., 1984).

Perceptual color view. The CIELUV branch uses the same tristimulus values, with $X_U = X$, $Y_U = Y$, and $Z_U = Z$. Its D65 reference white has $X_n = 0.950456$, $Y_n = 1$, and $Z_n = 1.088754$ (Commission Internationale de l'Eclairage, n.d.). With $d_U = X_U + 15Y_U + 3Z_U$,

$$\begin{aligned} L^* &= 116 f(Y_U/Y_n) - 16, \\ u^* &= 13L^*(u' - u'_n), \qquad u' = 4X_U/d_U, \\ v^* &= 13L^*(v' - v'_n), \qquad v' = 9Y_U/d_U, \end{aligned} \tag{4}$$

where $\delta = 6/29$, $f(t) = t^{1/3}$ for $t > \delta^3$, and $f(t) = t/(3\delta^2) + 4/29$ otherwise. The white-point coordinates are $u'_n = 4X_n/(X_n + 15Y_n + 3Z_n)$ and $v'_n = 9Y_n/(X_n + 15Y_n + 3Z_n)$. Black pixels are assigned $(L^*, u^*, v^*) = (0, 0, 0)$. Denote this coordinate map by Q_U, and the maps formed by $\mathbf{q}_L$ and $\mathbf{q}_D$ by Q_L and Q_D.

Three-channel target construction. For $b \in \{L, D\}$, let $q_{b,c}(p)$ be channel $c \in \{1, 2, 3\}$ of Q_b. Compute fixed bounds $l_{b,c}$ and $h_{b,c}$ as the minimum and maximum of that channel over the eight vertices of the unit linear-RGB cube. Since both transforms are linear, these vertices determine the extrema over the entire cube. Define

$$I_c^b(p) = \mathrm{clip}_{[0,1]}\left(\frac{q_{b,c}(p) - l_{b,c}}{h_{b,c} - l_{b,c}}\right), \qquad b \in \{L, D\}. \tag{5}$$

For CIELUV, use fixed channel packing

$$I^U = \mathrm{clip}_{[0,1]}\left(\frac{L^*}{100}, \frac{u^* + 134}{354}, \frac{v^* + 140}{262}\right). \tag{6}$$

Clipping acts component-wise. The floating-point targets preserve channel order (ℓ, m, s), (o_1, o_2, o_3), or (L^*, u^*, v^*), with fixed bounds across images. They enter the visual encoder directly after its pretrained resizing, cropping, and channel normalization. Training stimuli and retrieval candidates undergo the same branch-specific transformation Φ_b.

3.3 Independent Contrastive Alignment

Fig 2(b) shows five branches with separate neural encoders. Each receives the same recording $\mathbf{X}$.

Branch	Neural embedding	Visual embedding
R	$\mathbf{e}_R = f_R^{\mathrm{RPE}}(\mathbf{X})$	$\mathbf{v}_R = F_{\mathrm{RN50}}(I)$
L	$\mathbf{e}_L = f_L^{\mathrm{RPE}}(\mathbf{X})$	$\mathbf{v}_L = F_{\mathrm{RN50}}(I^L)$
D	$\mathbf{e}_D = f_D^{\mathrm{RPE}}(\mathbf{X})$	$\mathbf{v}_D = F_{\mathrm{RN50}}(I^D)$
U	$\mathbf{e}_U = f_U^{\mathrm{RPE}}(\mathbf{X})$	$\mathbf{v}_U = F_{\mathrm{RN50}}(I^U)$
G	$\mathbf{e}_G = f_G^{\mathrm{ATM}}(\mathbf{X})$	$\mathbf{v}_G = F_{\mathrm{H14}}(I)$

(7)

For $b \in \mathcal{B}$, neural and visual embeddings satisfy $\mathbf{e}_b, \mathbf{v}_b \in \mathbb{R}^d$, and all four RPE branches share the frozen CLIP ResNet-50 encoder F_{RN50} (Radford et al., 2021). Branch G independently trains ATM f_G^{ATM} (Li et al., 2024) and ViT-H-14 F_{H14} on original-image pairs, producing $\mathbf{e}_G, \mathbf{v}_G \in \mathbb{R}^{d_g}$. Here, d and d_g are the respective joint embedding dimensions.

Residual Projection Encoder. For $b \in \mathcal{B}$, flatten the recording into $\mathbf{h}_{b,0} = \mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{C_s T}$. Define $\mathcal{A}_{b,j}(\mathbf{h}) = \mathrm{LN}_{b,j}(\mathrm{Dropout}(\mathrm{GELU}(\mathbf{h})))$ for $j \in \{1, \mathrm{res}\}$. The RPE computes

$$\begin{aligned} \mathbf{h}_{b,1} &= \mathcal{A}_{b,1}(\mathbf{W}_{b,1}\mathbf{h}_{b,0} + \mathbf{q}_{b,1}), \\ \mathbf{h}_{b,\mathrm{res}} &= \mathbf{h}_{b,1} + \mathcal{A}_{b,\mathrm{res}}(\mathbf{W}_{b,\mathrm{res}}\mathbf{h}_{b,1} + \mathbf{q}_{b,\mathrm{res}}), \\ \mathbf{e}_b &= \mathrm{LN}_{b,\mathrm{out}}(\mathbf{W}_{b,\mathrm{out}}\mathbf{h}_{b,\mathrm{res}} + \mathbf{q}_{b,\mathrm{out}}). \end{aligned} \tag{8}$$

GELU is the Gaussian error linear unit, Dropout regularizes hidden activations, and LN denotes a separate LayerNorm at each stage. Both hidden vectors have dimension 1024. The matrices $\mathbf{W}_{b,1}$, $\mathbf{W}_{b,\mathrm{res}}$, and $\mathbf{W}_{b,\mathrm{out}}$ have dimensions $1024 \times C_s T$, 1024×1024, and $d \times 1024$, respectively. Their biases $\mathbf{q}_{b,1}$, $\mathbf{q}_{b,\mathrm{res}}$, and $\mathbf{q}_{b,\mathrm{out}}$ have the corresponding output dimensions.

Symmetric contrastive objective. For branch $b \in \mathcal{B} \cup \{G\}$ and a minibatch of B paired samples, define $s_{ij}^b = \overline{\mathbf{e}}_{b,i}^\top \overline{\mathbf{v}}_{b,j}/\tau_b$, where $i, j \in \{1, \ldots, B\}$, $\overline{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$, and $\tau_b > 0$ is the branch's learned temperature. Each branch independently minimizes

$$\mathcal{L}_{\mathrm{align}}^b = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(s_{ii}^b)}{\sum_{j=1}^{B}\exp(s_{ij}^b)} + \log\frac{\exp(s_{ii}^b)}{\sum_{j=1}^{B}\exp(s_{ji}^b)}\right]. \tag{9}$$

The two terms perform neural-to-image and image-to-neural contrastive alignment. They update each RPE against frozen RN50 targets and jointly update ATM and ViT-H-14 in branch G.

3.4 Complementary Visual Decoding

Fig 2(c) shows the task-specific use of the learned embeddings.

3.4.1 Retrieval and Classification

For each $b \in \mathcal{B}$, set $\Phi_R(I) = I$ and compute

$$s_b^{\mathrm{ret}}(\mathbf{X}, I_j) = \overline{\mathbf{e}}_b^\top \overline{F_{\mathrm{RN50}}(\Phi_b(I_j))}, \tag{10}$$

$$s_b^{\mathrm{cls}}(\mathbf{X}, y) = \overline{\mathbf{e}}_b^\top \overline{F_T(p(y))}. \tag{11}$$

Here, I_j is a candidate image, y is a candidate category, $p(y)$ is its text prompt, and F_T is the frozen CLIP text encoder paired with RN50. Each branch ranks candidates by cosine similarity. Let $\mathcal{S}_{b,i}^{(k)}$ contain its top-k image identities or category labels for test recording i. We evaluate the union

$$\mathcal{U}_i^{(k)} = \bigcup_{b \in \mathcal{B}} \mathcal{S}_{b,i}^{(k)}, \qquad \mathrm{Hit}_\cup @k = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \mathbf{1}[a_i \in \mathcal{U}_i^{(k)}]. \tag{12}$$

Here, N_{test} is the number of test recordings, a_i is the correct image identity or category, and $\mathbf{1}[\cdot]$ is the indicator function. The union contains at most $4k$ candidates, and its hit rate measures complementary successes across branches.

3.4.2 Image Reconstruction

Reconstruction uses branch G for primary conditioning and $\mathcal{A} = \{L, D, U\}$ for auxiliary conditioning. Branch R supports discriminative decoding. After alignment, all neural and image encoders are frozen while the diffusion prior and linear projections learn from their embeddings. SDXL and the IP-Adapters remain fixed.

Primary diffusion prior. We use the unnormalized ViT-H-14 output as the feature target and set the additional feature scale $\rho = 1$, giving $\mathbf{z}_0 = \rho \mathbf{v}_G$. For diffusion step $t \in \{1, \ldots, T_p\}$,

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}_p, \tag{13}$$

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

Here, T_p is the number of steps, $\bar{\alpha}_t = \prod_{j=1}^{t}(1 - \beta_j)$ for noise variances $\beta_j \in (0, 1)$, and $\boldsymbol{\epsilon}_p \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_g})$ has identity covariance. The conditioned denoiser has parameters θ. For classifier-free guidance, $\widetilde{\mathbf{e}}_G$ equals the null condition $\varnothing$ with probability $p_{\mathrm{drop}} = 0.1$ and $\mathbf{e}_G$ otherwise. The expectation averages over paired samples, steps, noise, and condition dropout. At inference, the guided prediction is

$$\widehat{\boldsymbol{\epsilon}}_\theta = \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \varnothing) + g\left[\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{e}_G) - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \varnothing)\right], \tag{15}$$

where g is the guidance scale, set to 1 in the base configuration. Reverse diffusion from Gaussian noise yields $\widehat{\mathbf{z}}_0$, and the primary adapter condition is $\mathbf{c}_G = \widehat{\mathbf{z}}_0/\rho$.

Auxiliary projections and joint conditioning. For $b \in \mathcal{A}$, a bias-free linear projection produces

$$\mathbf{a}_b = \mathbf{A}_b \mathbf{e}_b, \qquad \mathbf{A}_b \in \mathbb{R}^{d_g \times d}, \qquad \mathbf{c}_b = \mathbf{a}_b/\rho. \tag{16}$$

The projected vector $\mathbf{a}_b$ uses the scaled target space, and $\mathbf{c}_b$ supplies the corresponding adapter condition. Separate IP-Adapters (Ye et al., 2023) inject all four conditions into pretrained SDXL (Podell et al., 2024) using the same feature-scale convention. At each adapted attention layer,

$$\mathbf{H}_{\mathrm{out}} = \mathbf{H}_{\mathrm{base}} + 0.8\,\mathrm{Attn}(\mathbf{Q}, \mathbf{K}_G, \mathbf{V}_G) + 0.5 \sum_{b \in \mathcal{A}} \mathrm{Attn}(\mathbf{Q}, \mathbf{K}_b, \mathbf{V}_b),$$
$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d_h})\mathbf{V}. \tag{17}$$

The matrices $\mathbf{H}_{\mathrm{base}}$ and $\mathbf{H}_{\mathrm{out}}$ are the base and conditioned attention outputs. Queries $\mathbf{Q}$ come from generator features, while $\mathbf{K}_b$ and $\mathbf{V}_b$ are keys and values derived from the adapter tokens for $\mathbf{c}_b$. The scalar d_h is the attention-head key/query dimension. The coefficients 0.8 and 0.5 are fixed attention-output scales.

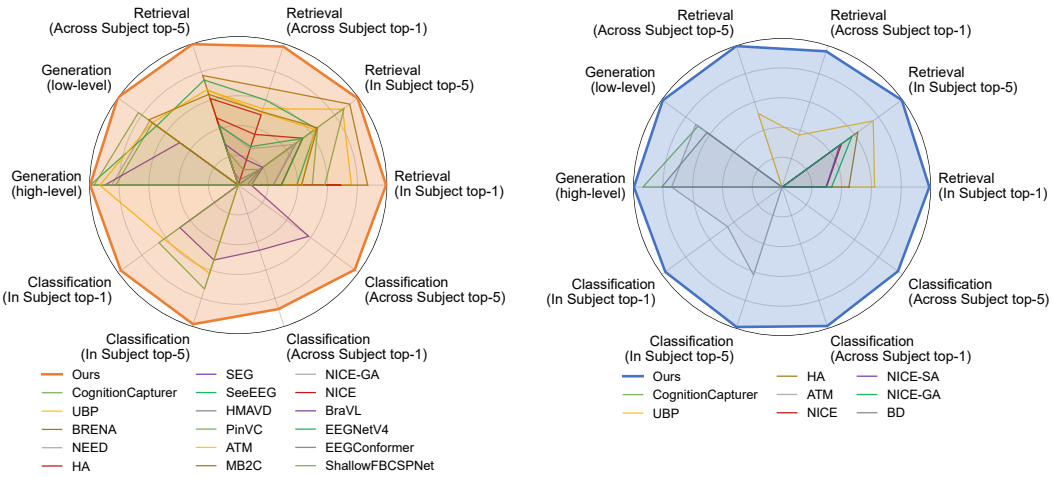

Figure 3: Decoding performance on THINGS-EEG (left) and THINGS-MEG (right).

Generation objective. Auxiliary projections are supervised by the scaled original-image target $\mathbf{z}_0 = \rho \mathbf{v}_G$. With pair set $\mathcal{P} = \{(L, D), (L, U), (D, U)\}$, define

$$\mathcal{L}_{\text{sup}} = \frac{1}{3} \sum_{b \in \mathcal{A}} \mathbb{E} \|\mathbf{a}_b - \mathbf{z}_0\|_2^2, \tag{18}$$

$$\mathcal{L}_{\text{cons}} = \frac{1}{3} \sum_{(b,c) \in \mathcal{P}} \mathbb{E} \|\mathbf{a}_b - \mathbf{a}_c\|_2^2, \tag{19}$$

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{prior}} + \lambda_{\text{proj}} (\mathcal{L}_{\text{sup}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}}). \tag{20}$$

Expectations are taken over paired training examples. We set $\lambda_{\text{proj}} = 0.5$ and $\lambda_{\text{cons}} = 0.1$. Feature supervision aligns auxiliary projections with the common target, while consistency regularizes their pairwise differences. The prior loss updates θ, and the projection losses update $\{\mathbf{A}_b\}_{b \in \mathcal{A}}$. Global gradient clipping limits the combined norm to $\kappa_{\text{clip}} = 1$.

At inference, $\widehat{I} = \mathcal{G}_{\text{SDXL}}(\boldsymbol{\xi}, \mathbf{c}_G, \mathbf{c}_L, \mathbf{c}_D, \mathbf{c}_U)$, where $\boldsymbol{\xi}$ is standard Gaussian noise in the SDXL latent space and $\mathcal{G}_{\text{SDXL}}$ includes conditional denoising and VAE decoding. All four conditions are computed from the test neural recording.

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 MuCCA on THINGS-EEG and THINGS-MEG across retrieval, classification, and reconstruction. MuCCA outperforms state-of-the-art methods in the vast majority of evaluated settings as shown in Fig 3. For discriminative decoding, MuCCA reports candidate-union hit rates, which quantify complementary branch successes. Appendix A.2 presents retrieval in Tables 3 and 4, classification in Table 5, and reconstruction in Table 6. Retrieval examples in Appendix A.4 (Figs. 8–13) show cases where the target ranks first and cases where it ranks second through fifth. In the latter cases, higher-ranked candidates often share visual or semantic similarities with the target. The detailed reconstruction comparison shows gains in low-level fidelity alongside metric-dependent differences in semantic similarity. MuCCA leads on five of seven EEG reconstruction metrics and all seven MEG metrics (Table 6). The subject-wise results also show a substantial gap between intra-subject and inter-subject retrieval, identifying cross-participant transfer as a remaining challenge.

Table 1: Color-representation combinations on THINGS-EEG. L, D, and U denote LMS, DKL-inspired, and CIELUV views. RGB supervision is retained in every row. Retrieval and classification values are union-hit percentages.

Views	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 ↓
L	58.0	86.7	14.1	37.4	7.8	23.7	2.2	7.6	0.171	0.438	0.818	0.896	0.785	0.813	0.547
D	53.6	83.7	15.3	40.3	8.3	25.5	2.6	9.0	0.161	0.438	0.826	0.886	**0.810**	0.819	0.543
U	59.2	88.7	15.6	40.3	9.4	25.5	2.5	8.4	0.177	0.439	0.836	0.880	0.800	0.824	0.546
L+D	58.0	87.2	18.9	44.8	9.7	29.2	3.1	10.2	0.174	0.465	0.829	0.904	0.800	0.815	**0.522**
L+U	65.4	**91.0**	19.1	45.1	10.4	29.0	2.9	9.7	0.189	0.469	0.837	0.905	0.780	0.821	0.528
D+U	63.4	89.6	19.6	46.5	10.9	29.5	3.1	10.6	0.197	0.471	0.822	0.894	0.796	**0.831**	0.526
L+D+U	**66.9**	**91.0**	**22.3**	**49.8**	**11.7**	**32.4**	**3.5**	**11.6**	**0.2112**	**0.515**	**0.8441**	**0.911**	0.785	0.819	0.526

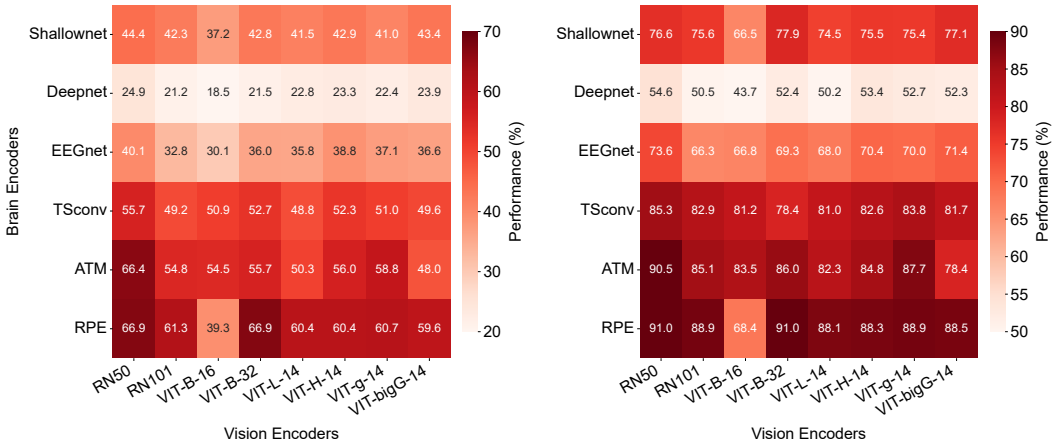

Figure 4: Neural–image encoder comparison on THINGS-EEG. Panels show Top-1 and Top-5 retrieval performance.

4.2 Ablation Study

Complementary color representations. To assess the contribution of complementary visual targets, we vary the LMS, DKL-inspired, and CIELUV branches while retaining RGB supervision. Table 1 shows that removing branches reduces the proportion of targets retrieved or classified correctly by at least one branch. The complete configuration reaches 66.9% for intra-subject retrieval and 11.7% for classification when combining the branches' first-ranked predictions. It also achieves the best PixCorr, SSIM, and both AlexNet scores, while some subsets perform better on Inception, CLIP, or SwAV. The strongest two-view combination varies across metrics, suggesting that the transformed targets provide complementary information. Combining all three auxiliary views yields the strongest overall discriminative coverage and improves several structural reconstruction measures.

Neural and image encoders. To select the alignment architecture, we compare neural encoders across visual backbones in Fig. 4. The heatmap favors RPE with RN50 for retrieval and shows that the relative suitability of visual backbones depends on their alignment with neural recordings. To test whether reconstruction benefits from a different pairing, Table 2 varies the primary and auxiliary encoder assignments. The ATM–ViT-H-14 primary branch with RPE–RN50 auxiliary branches achieves the best results on all seven reconstruction metrics. This supports using original-image features for primary conditioning while retaining representation-specific auxiliary conditions. The advantage over both homogeneous assignments indicates that reconstruction benefits from assigning different encoder pairs to the primary and auxiliary roles. Reversing these roles reduces performance across all seven metrics.

Table 2: 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 ↓
ATM–ViT-H-14	ATM–ViT-H-14	0.146	0.473	0.789	0.852	0.756	0.788	0.540
RPE–RN50	ATM–ViT-H-14	0.136	0.342	0.576	0.524	0.561	0.602	0.733
RPE–RN50	RPE–RN50	0.129	0.303	0.536	0.500	0.547	0.507	0.750
ATM–ViT-H-14	RPE–RN50	**0.2112**	**0.515**	**0.8441**	**0.911**	**0.785**	**0.819**	**0.526**

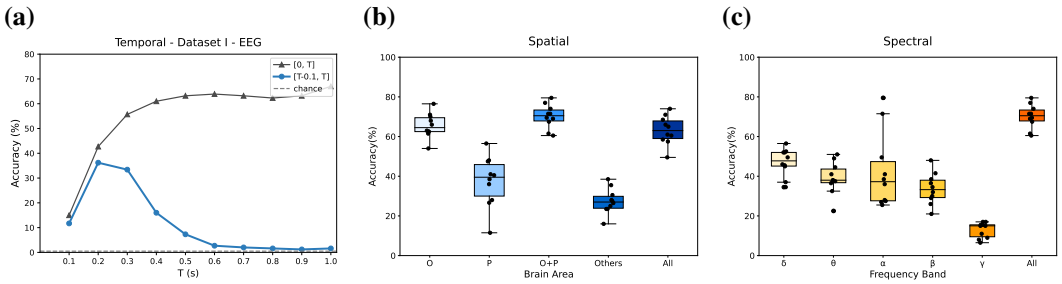

Figure 5: EEG retrieval across **(a)** temporal windows, **(b)** channel subsets, and **(c)** frequency bands.

4.3 Temporal, Spatial, and Spectral Analysis

To identify informative neural measurements, we vary temporal windows, EEG channel subsets, and frequency bands. Fig. 5 summarizes these comparisons. Early post-stimulus responses during 100–200 ms support strong retrieval, occipital–parietal channels perform well across tasks, and full-passband EEG outperforms individual bands. Among individual bands, Delta achieves the best average performance. Appendix A.3 presents detailed visualizations in Figs. 6 and 7, followed by quantitative results in Tables 7–13. The temporal results show that useful information accumulates over the response interval, with the best window depending on the metric. The spatial and spectral comparisons favor posterior channels and broadband inputs while revealing task-dependent benefits from other channels. In particular, all-channel input improves SSIM, CLIP, and SwAV, whereas posterior channels favor PixCorr and both AlexNet measures (Table 12). This variation supports selecting neural inputs according to the desired reconstruction properties.

5 Conclusion and Future Work

MuCCA constructs complementary visual supervision from multiple color representations and learns independent neural alignments for EEG and MEG decoding. The branches reveal complementary discriminative successes and provide joint conditioning for image reconstruction. Cross-subject decoding remains challenging, and the transformations approximate aspects of color processing and could be extended through individual perceptual measurements. Future work will develop learned ranking fusion, subject adaptation, and adaptive conditioning that balances structural and semantic reconstruction quality.

6 AI Use Statement

A large language model assisted with language editing and manuscript organization. The authors reviewed the revisions and take responsibility for all scientific content.

7 Ethics Statement

This study uses publicly available THINGS-EEG and THINGS-MEG data and involves no new participant recruitment. Applications of neural visual decoding should protect neural-data privacy and respect informed consent and restrictions on secondary use.

8 Reproducibility Statement

The datasets are publicly available, and the code link is provided in the abstract. Section 3 specifies the model and objectives, Appendix A.1 describes the experimental protocol, and Appendices A.2–A.4 provide detailed results and visualizations.

References

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. In *International Conference on Learning Representations*, volume 2024, pp. 7846–7858, 2024a.

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

Commission Internationale de l'Eclairage. CIE 1976 L*u*v* colour space. CIE International Lighting Vocabulary, n.d. URL https://cie.co.at/eilv/159. Accessed 18 September 2026.

Andrew M Derrington, John Krauskopf, and Peter Lennie. Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology*, 357:241–265, 1984.

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.

Karl R. Gegenfurtner. Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, 4 (7):563–572, 2003. doi: 10.1038/nrn1138. URL https://www.nature.com/articles/nrn1138.

Shuai Huang, Huan Luo, Haodong Jing, Qixian Zhang, Litao Chang, Yating Feng, Xiao Lin, Chendong Qin, Han Chen, Shuwen Jia, et al. Need: Cross-subject and cross-task generalization for video and image reconstruction from eeg signals. *Advances in Neural Information Processing Systems*, 38:173134–173173, 2026.

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*, pp. 114212, 2025.

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 *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4883–4892. IEEE, 2025.

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

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, 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.

Xingyu Miao, Bing Zhai, Haoran Duan, Yang Long, et al. Decoding visual neural representations by multimodal with dynamic balancing. *arXiv preprint arXiv:2509.03433*, 2025.

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.

Yonghao Song, Qingqing Zheng, Bing-Chuan 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.

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.

A. Stockman, L. T. Sharpe, and C. Fach. The spectral sensitivity of the human short-wavelength sensitive cones derived from thresholds and color matches. *Vision Research*, 39(17):2901–2927, 1999. doi: 10.1016/S0042-6989(98)00225-9. URL https://pubmed.ncbi.nlm.nih.gov/10492818/.

Andrew Stockman and Lindsay T. Sharpe. The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research*, 40(13):1711–1737, 2000. doi: 10.1016/S0042-6989(00)00021-3. URL https://pubmed.ncbi.nlm.nih.gov/10814758/.

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.

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.

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.

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 63-channel recordings from 10 participants, sampled at 1,000 Hz. The training set contains 16,540 images with four retained repetitions per image, and the test set contains 200 images with 80 repetitions per image. Recordings are filtered to 0.1–100 Hz and downsampled to 250 Hz. We apply baseline correction over -200–0 ms, and analyze 0–1,000 ms after stimulus onset. The default input uses 17 posterior electrodes, comprising O1, Oz, O2, PO7, PO3, POz, PO4, PO8, P7, P5, P3, P1, Pz, P2, P4, P6, and P8. Training repetitions remain separate, and all 80 test repetitions are averaged for each image. The channel analysis additionally evaluates other subsets and all 63 recorded channels. Multivariate noise normalization uses training data. THINGS-MEG contains recordings from four participants over 12 sessions. The main stimulus set includes 22,248 images, and the repeated set contains 200 images presented once per session. The decoding input uses 271 channels. We filter the recordings to 0.1–40 Hz, resample to 200 Hz, baseline-correct each epoch, and analyze 0–1,000 ms after stimulus onset. The 12 test repetitions are averaged for each image. Test concepts are excluded from training for zero-shot evaluation.

Evaluation metrics. Retrieval ranks 200 candidate images by neural–image cosine similarity, and classification compares neural embeddings with category-text embeddings. Intra-subject evaluation trains and tests on the same participant, whereas inter-subject evaluation holds out one participant. Each branch produces Top-1 and Top-5 candidate sets. MuCCA combines the corresponding sets across branches by taking their union and counts a prediction as correct when the target belongs to this union. Reconstruction uses PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, CLIP, and SwAV. Higher values indicate better performance except for SwAV. Bold values indicate the best results within each comparison block, including ties.

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., 2026), human-aligned encoders (Rajabi et al., 2025), SEG (Meng & Yang, 2024), SeeEEG (Kim et al., 2025), HMAVD (Miao et al., 2025), 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., 2024a). Each result table identifies the available task-specific comparisons.

Implementation details. Experiments use a single NVIDIA RTX 4090 GPU and Adam. The learning rate is 10^{-4} for intra-subject alignment and 10^{-5} for inter-subject alignment. Alignment is trained for 50 epochs, and the reconstruction conditioning predictors are trained for 100 epochs. The contrastive temperature is initialized to 0.07. Four independent RPEs align RGB, LMS, DKL-inspired, and CIELUV views with frozen RN50 features. A separate ATM–ViT-H-14 branch learns original-image supervision for reconstruction. After alignment, all neural and image encoders are frozen. The subsequent objective trains the diffusion prior and auxiliary linear projections, while SDXL and the IP-Adapters remain fixed. Contrastive similarities use unit-normalized embeddings. The generation target is the unnormalized ViT-H-14 embedding with feature scale $\rho = 1$. The projection inputs and prior conditions retain the embeddings specified in Section 3. We use $\lambda_{\text{proj}} = 0.5$, $\lambda_{\text{cons}} = 0.1$, and global gradient clipping at norm 1. The primary adapter scale is 0.8, and each auxiliary adapter scale is 0.5. The effective consistency coefficient in the total objective is 0.05. Adapter scales are fixed attention-output multipliers. The diffusion prior uses condition dropout 0.1 and guidance scale $g = 1$. All inference conditions are computed from the neural recording.

A.2 Overall Performance

Retrieval. To assess participant variability, Tables 3 and 4 report subject-wise results. MuCCA has the highest reported mean union-hit values in both evaluation settings on both datasets. On THINGS-EEG, the reported means are 66.9% and 90.9% for intra-subject retrieval and 22.3% and 46.8% for inter-subject retrieval. On THINGS-MEG, they are 42.5% and 72.5%, and 5.75% and 19.9%, respectively. The reduction in inter-subject performance on both datasets shows that cross-participant transfer remains challenging. Subject-wise variation further indicates that the benefit of alignment depends on the available neural signal. Appendix A.4 illustrates successful matches and remaining retrieval errors.

Table 3: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG.

Method	Subject-wise accuracy (%)																					
	S1		S2		S3		S4		S5		S6		S7		S8		S9		S10		Mean	
	t1	t5	t1	t5	t1	t5	t1	t5	t1	t5	t1	t5	t1	t5	t1	t5	t1	t5	t1	t5	t1	t5
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
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
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
UBP	41.2	70.5	51.2	80.9	51.2	82.0	51.1	76.9	42.2	72.8	57.5	83.5	49.0	79.9	58.6	85.8	45.1	76.2	61.5	88.2	50.9	79.7
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	33.0	69.3	27.5	59.3	36.2	70.3	28.2	60.3
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	62.3	21.5	46.7	26.1	55.8	23.7	50.9
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	88.6	34.6	79.4	35.5	81.2
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	88.7	37.7	87.8	35.8	80.9	35.9	81.7
MuCCA	**55.0**	**83.5**	**67.0**	**92.5**	**67.5**	**93.5**	**63.0**	**92.5**	**59.0**	**83.0**	**71.0**	**92.5**	**66.0**	**90.0**	**75.5**	**97.0**	**68.5**	**90.5**	**77.0**	**95.0**	**66.9**	**90.9**
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
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
UBP	11.5	29.7	15.5	40.0	9.8	27.0	13.0	32.3	8.8	33.8	11.7	31.0	10.2	23.8	12.2	32.2	15.5	40.5	16.0	43.5	12.4	33.4
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	10.5	31.8	12.0	32.8	13.1	32.5	13.0	33.0	11.6	31.3
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	26.5	8.2	23.8	8.3	24.2	8.1	21.9
MuCCA	**17.0**	**41.0**	**23.5**	**52.5**	**12.0**	**36.5**	**24.0**	**44.5**	**17.0**	**45.0**	**16.5**	**49.0**	**17.0**	42.0	**20.0**	**43.0**	**25.0**	**58.5**	**29.5**	**56.0**	**22.3**	**46.8**

Table 4: Subject-wise retrieval on THINGS-MEG. Values are percentages. MuCCA reports candidate-union hit rates.

Method	Subject 1		Subject 2		Subject 3		Subject 4		Mean	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Intra-subject										
NICE	9.6	27.8	18.5	47.8	14.2	41.6	9.0	26.6	12.8	36.0
NICE-SA	9.8	27.8	18.6	46.4	10.5	38.4	11.7	27.2	12.7	35.0
NICE-GA	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.3	42.3
UBP	15.0	38.0	46.0	80.5	27.3	59.0	18.5	43.5	26.7	55.2
MuCCA	**23.5**	**58.0**	**65.0**	**90.0**	**50.5**	**79.5**	**31.0**	**62.5**	**42.5**	**72.5**
Inter-subject										
UBP	2.0	5.7	1.5	17.2	2.7	10.5	2.5	8.0	2.2	10.4
MuCCA	**5.0**	**14.5**	**5.0**	**24.5**	**7.5**	**24.5**	**5.5**	**16.0**	**5.75**	**19.9**

Classification. To assess category-level decoding, Table 5 compares MuCCA with available baselines. On THINGS-EEG, MuCCA reaches 11.7% and 32.4%, exceeding ATM by 4.3 and 11.8 percentage points. On THINGS-MEG, its 3.5% and 11.6% remain below ATM. The EEG improvements show that the aligned visual representations support category-text matching as well as image retrieval. The MEG comparison identifies category-level transfer as an area requiring further improvement, even when reconstruction and image retrieval are effective.

Reconstruction. To assess complementary conditioning, Table 6 compares reconstruction against the available baselines. MuCCA leads on five of seven metrics on THINGS-EEG and all seven on THINGS-MEG. The EEG PixCorr and SSIM reach 0.2112 and 0.515, respectively. Inception and CLIP favor other EEG baselines, indicating a remaining trade-off between low-level fidelity and feature-level similarity. The combination improves pixel-level fidelity and several learned feature measures, supporting the use of complementary conditions in the shared generator. Reconstruction examples under temporal and spatial perturbations appear in Figs. 6 and 7, illustrating how the available neural input influences recovered structure and appearance.

Table 5: 200-way zero-shot classification on THINGS-EEG and THINGS-MEG. Values are percentages.

Dataset	Method	Accuracy	
		Top-1	Top-5
THINGS-EEG	BraVL	5.8	17.5
	ATM	7.4	20.6
	MuCCA	**11.7**	**32.4**
THINGS-MEG	ATM	**5.9**	**19.3**
	MuCCA	3.5	11.6

Table 6: Reconstruction on THINGS-EEG and THINGS-MEG.

Methods	Datasets	Low-level		High-level				
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
ATM	THINGS-EEG	0.160	0.345	0.776	0.866	0.734	0.786	0.582
PinVC	THINGS-EEG	0.189	0.390	0.794	0.881	**0.806**	0.809	0.563
CognitionCapturer	THINGS-EEG	0.150	0.347	0.754	0.623	0.669	0.715	0.590
SeeEEG	THINGS-EEG	0.130	0.344	0.809	0.904	0.784	**0.869**	0.539
MuCCA	THINGS-EEG	**0.2112**	**0.515**	**0.8441**	**0.911**	0.785	0.819	**0.526**
ATM	THINGS-MEG	0.104	0.340	0.613	0.672	0.619	0.603	0.651
MuCCA	THINGS-MEG	**0.152**	**0.461**	**0.815**	**0.913**	**0.793**	**0.835**	**0.528**

Table 7: Decoding with non-overlapping 100-ms EEG windows. Retrieval and classification values 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	11.70	33.10	4.05	18.10	2.11	11.78	0.63	4.20	0.1560	0.4510	0.7780	0.8340	0.6880	0.7060	0.5720
100–200	**36.20**	**69.80**	**12.53**	**38.16**	**6.52**	**24.85**	**1.95**	**8.86**	0.1720	0.4560	**0.8030**	**0.8710**	**0.7440**	**0.8060**	**0.5390**
200–300	33.40	66.50	11.56	36.36	6.01	23.68	1.80	8.44	0.1730	0.4730	0.7820	0.8450	0.7370	0.7790	0.5430
300–400	16.00	44.00	5.54	24.05	2.88	15.67	0.86	5.58	**0.1733**	**0.4900**	0.7590	0.8300	0.7060	0.7680	0.5550
400–500	7.30	23.50	2.53	12.85	1.31	8.37	0.39	2.98	0.1544	0.4790	0.7200	0.7700	0.7030	0.7390	0.5650
500–600	2.70	12.10	0.93	6.62	0.49	4.31	0.14	1.54	0.1613	0.4690	0.7490	0.7960	0.6980	0.7240	0.5720
600–700	2.00	10.40	0.69	5.69	0.36	3.70	0.11	1.32	0.1510	0.4700	0.7444	0.7690	0.6520	0.7150	0.5830
700–800	1.60	7.80	0.55	4.26	0.29	2.78	0.09	0.99	0.1590	0.4680	0.7270	0.7720	0.6340	0.6670	0.5840
800–900	1.20	6.70	0.42	3.66	0.22	2.39	0.07	0.85	0.1380	0.4740	0.7470	0.7900	0.6750	0.7210	0.5690
900–1000	1.60	7.10	0.55	3.88	0.29	2.53	0.09	0.90	0.1500	0.4720	0.7380	0.7600	0.6290	0.6670	0.5900

A.3 Temporal, Spatial, and Spectral Analysis

To examine which neural measurements support decoding, we compare temporal windows, channel subsets, and frequency bands. Figs 6 and 7 present the visual results before the corresponding tables.

Temporal and spatial visualizations. Fig 6 shows retrieval across EEG and MEG windows and reconstructions from cumulative EEG responses. Extending the early response interval improves the available visual information. Fig 7 compares channel subsets, frequency bands, and channel-dependent reconstructions, illustrating that decoding quality depends on the selected neural input.

Temporal windows. Tables 7 and 8 evaluate non-overlapping 100-ms windows. The 100–200 ms interval produces the highest retrieval values in both modalities, while classification and reconstruction peaks depend on the modality and metric. Tables 9 and 10 evaluate cumulative responses. Longer windows generally improve decoding over the shortest intervals, but the best endpoint is not identical across metrics. The advantage of early windows indicates that informative visual responses emerge soon after stimulus onset. Cumulative inputs improve performance over isolated short windows,

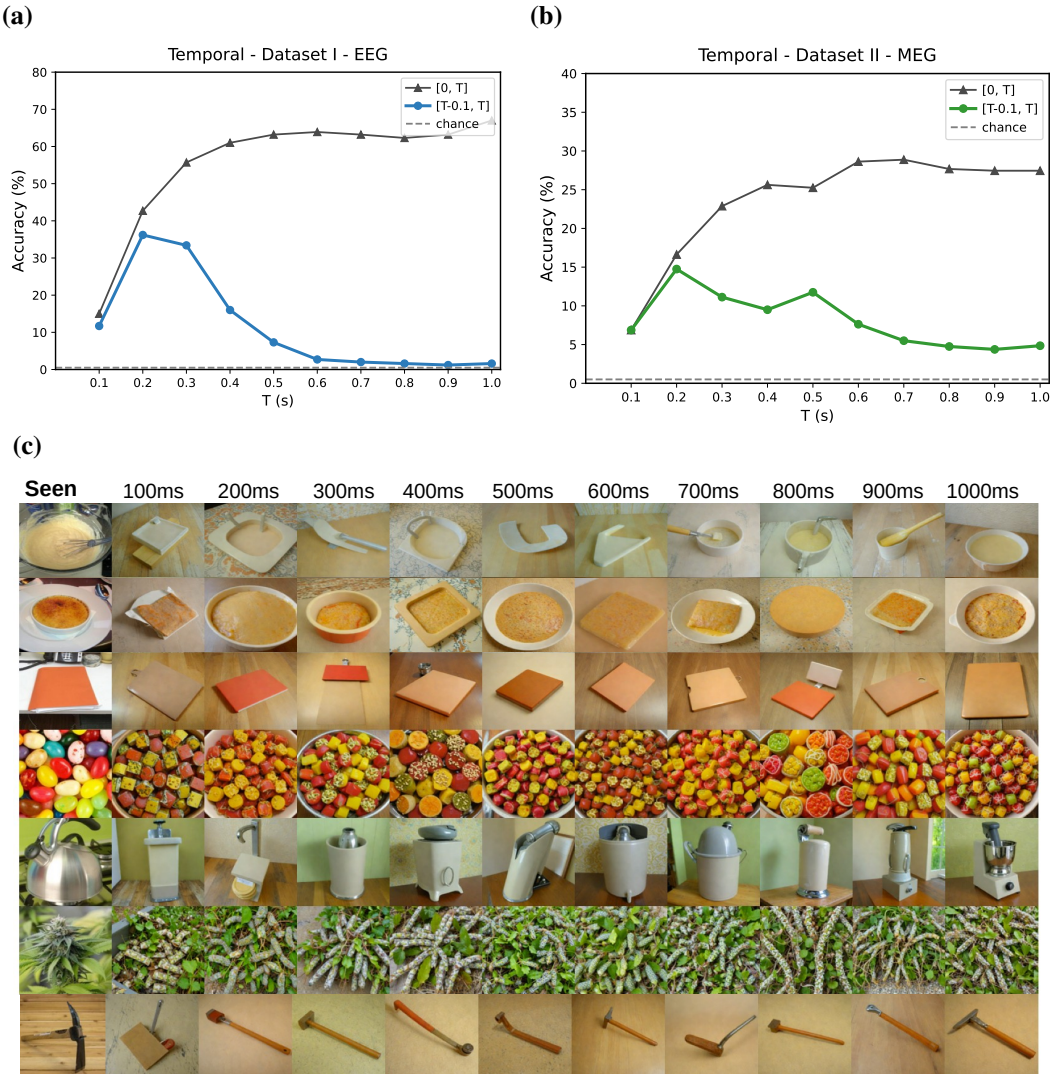

Figure 6: Temporal analysis. **(a, b)** Retrieval across EEG and MEG windows. **(c)** Reconstructions with cumulative EEG windows.

supporting integration across the response interval. Differences among the reconstruction optima indicate that pixel-level and feature-level information benefit from partly different temporal inputs.

Channel subsets. Table 11 compares occipital, parietal, combined, other, and all-channel inputs. Occipital–parietal recordings yield the strongest Top-1 results, while occipital-only recordings lead on several Top-5 measures. Table 12 compares reconstruction. Occipital–parietal inputs lead on PixCorr, both AlexNet metrics, and Inception, whereas all-channel inputs perform best on SSIM, CLIP, and SwAV. The combined posterior subset is particularly effective for Top-1 discrimination and several reconstruction measures. However, the gains from all-channel inputs on SSIM, CLIP, and SwAV show that additional channels can contribute information useful to specific reconstruction criteria. These results support task-aware channel selection.

Frequency bands. Table 13 compares individual bands with the full passband. Full-passband inputs achieve the highest performance across the reported metrics, and delta is the strongest individual band. The full passband improves both discriminative performance and reconstruction, indicating that the useful information extends beyond the strongest individual band. The results support retaining complementary spectral content under the evaluated protocol.

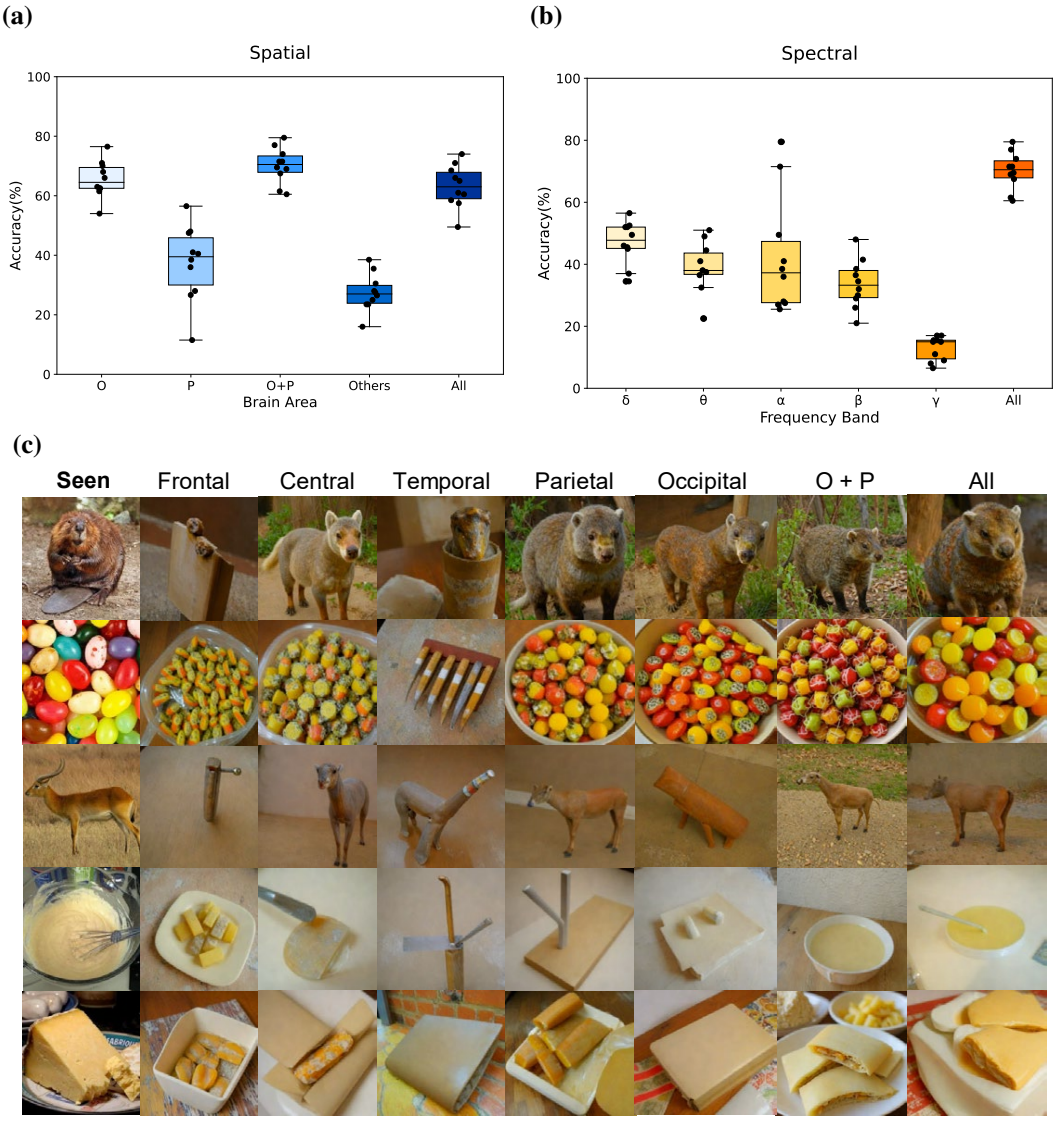

Figure 7: Spatial and spectral analysis. **(a)** EEG channel subsets. **(b)** EEG frequency bands. **(c)** Reconstructions with different channel subsets.

A.4 Retrieval Visualizations

To examine instance-level retrieval, Figs. 8, 9, and 10 show examples where the target ranks first. Figs. 11, 12, and 13 show examples where the target ranks second through fifth. The Top-1 matches demonstrate that the learned embeddings distinguish individual stimuli across diverse appearances. When the target appears at ranks two through five, higher-ranked candidates often share visual or semantic attributes with it, indicating difficulty distinguishing similar images. These examples complement the aggregate results by illustrating both accurate first-choice retrieval and correct targets recovered within a broader candidate set.

Table 8: Decoding with non-overlapping 100-ms MEG windows. Retrieval and classification values 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	6.88	19.25	2.63	12.00	3.00	13.25	2.88	10.00	0.0837	0.4134	**0.7120**	**0.7640**	**0.6650**	**0.6610**	0.6180
100–200	**14.75**	**43.50**	**5.13**	**16.38**	5.88	22.25	**3.13**	9.25	0.0820	0.4150	0.6070	0.6350	0.5630	0.5690	0.6790
200–300	11.13	36.38	4.00	15.75	6.88	**22.75**	**3.13**	10.25	0.0820	0.4020	0.6000	0.6120	0.5430	0.5600	0.6500
300–400	9.50	33.88	4.25	14.25	**7.50**	21.63	3.00	8.75	0.0790	0.4200	0.6120	0.6420	0.5330	0.5700	0.6600
400–500	11.75	31.88	3.00	13.38	5.25	19.50	2.25	9.38	0.0680	0.3900	0.5900	0.6200	0.5620	0.5480	0.6190
500–600	7.63	28.50	2.50	9.38	5.13	17.50	1.88	7.38	0.0750	**0.4600**	0.6520	0.6830	0.5420	0.5530	0.6340
600–700	5.50	20.50	2.00	9.38	3.50	11.88	1.25	6.25	0.0730	0.4300	0.6240	0.6270	0.5360	0.5640	0.6370
700–800	4.75	14.50	2.13	10.63	3.00	11.88	2.00	7.13	**0.0850**	0.4260	0.5820	0.5930	0.5820	0.5810	0.6210
800–900	4.38	13.70	1.95	9.13	2.83	11.15	1.88	6.65	0.0620	0.3820	0.6210	0.6630	0.5830	0.5310	**0.6110**
900–1000	4.85	14.30	2.13	10.33	3.23	11.40	2.00	**19.50**	0.0790	0.4260	0.6280	0.6650	0.5560	0.5720	0.6240

Table 9: Decoding with cumulative EEG windows. Retrieval and classification values 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	15.00	43.00	5.19	23.51	2.70	15.31	0.81	5.46	0.1560	0.4510	0.7780	0.8340	0.6880	0.7060	0.5720
0–200	42.70	75.00	14.78	41.00	7.69	26.70	2.30	9.52	0.1757	0.4530	0.7830	0.8600	0.7390	0.7940	0.5580
0–300	55.70	86.40	19.28	47.24	10.03	30.76	3.00	10.97	0.1880	0.4690	0.8210	0.8830	0.7520	0.8260	0.5380
0–400	61.00	88.50	21.12	48.38	10.98	31.51	3.29	11.23	0.1880	0.4690	0.8210	0.8830	0.7520	0.8260	0.5380
0–500	63.20	90.30	21.88	49.37	11.38	32.15	3.40	11.46	0.1870	0.4600	0.8450	0.8970	**0.8000**	0.8200	0.5280
0–600	63.90	89.20	22.12	48.77	11.50	31.76	3.44	11.32	0.1820	0.4660	0.8320	0.8950	0.7780	0.8480	0.5310
0–700	63.20	89.90	21.88	49.15	11.38	32.01	3.40	11.41	0.1824	0.4610	0.8340	0.8960	**0.8000**	0.8330	0.5220
0–800	62.30	90.10	21.57	49.26	11.21	32.08	3.36	11.44	0.1850	0.4630	**0.8490**	0.8890	0.7660	0.8220	0.5370
0–900	63.20	88.60	21.88	48.44	11.38	31.55	3.40	11.25	0.1890	0.4700	0.8340	0.9020	0.7950	**0.8490**	**0.5200**
0–1000	**66.90**	**91.00**	**22.30**	**49.75**	**11.70**	**32.40**	**3.50**	**11.55**	**0.2112**	**0.5150**	0.8441	**0.9110**	0.7850	0.8190	0.5260

Table 10: Decoding with cumulative MEG windows. Retrieval and classification values 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	6.88	19.25	2.63	12.00	3.00	13.25	2.88	10.00	0.0837	0.4134	0.7120	0.7640	0.6650	0.6610	0.6180
0–200	16.63	44.25	5.75	17.88	8.25	24.75	2.50	9.50	0.0920	**0.4676**	0.6840	0.7160	0.6350	0.6410	0.6020
0–300	22.88	54.38	6.25	20.75	11.00	30.50	3.38	11.88	0.1220	0.4230	0.6880	0.8820	0.6720	0.7000	0.5720
0–400	25.63	59.63	**6.88**	21.13	**12.75**	33.25	3.75	12.50	0.1300	0.4300	0.7200	0.8900	0.7000	0.7350	0.5560
0–500	25.25	60.13	6.25	**21.75**	12.63	33.88	3.50	**13.50**	0.1350	0.4350	0.7150	0.8850	0.6950	0.7300	0.5580
0–600	28.63	62.63	5.50	18.74	12.38	33.88	3.00	12.25	0.1400	0.4400	0.7600	0.9000	0.7500	0.7850	0.5380
0–700	28.88	61.25	6.38	21.63	10.75	32.88	**3.88**	12.63	0.1450	0.4450	0.7800	0.9050	0.7700	0.8050	0.5370
0–800	27.68	62.38	6.00	**21.75**	11.73	33.38	3.25	13.25	0.1480	0.4500	0.7900	0.9080	0.7750	0.8100	0.5350
0–900	27.45	61.75	5.30	20.45	12.38	32.80	3.33	12.10	0.1500	0.4550	0.8000	0.9100	0.7800	0.8200	0.5320
0–1000	**42.50**	**72.50**	5.90	20.75	12.63	**35.80**	3.63	12.50	**0.1520**	0.4608	**0.8150**	**0.9130**	**0.7930**	**0.8350**	**0.5280**

Table 11: Retrieval and classification with EEG channel subsets. Values are percentages. O and P denote occipital and parietal electrodes.

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	62.50	88.50	21.63	**53.85**	11.13	**35.07**	3.36	**12.50**
P	33.40	62.80	11.56	34.33	6.01	22.36	1.80	7.97
O+P	**66.90**	**91.00**	**22.30**	49.75	**11.17**	32.40	**3.50**	11.55
Others	23.70	55.10	8.20	30.12	4.27	19.62	1.28	6.99
All	59.60	87.77	20.63	47.95	10.73	31.22	3.21	11.13

Table 12: Reconstruction with EEG channel subsets. O+P denotes occipital and parietal electrodes.

Channel subset	Low-level		High-level				
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
Frontal	0.142	0.514	0.761	0.828	0.673	0.730	0.564
Central	0.160	0.511	0.751	0.823	0.694	0.770	0.542
Parietal	0.184	0.498	0.778	0.873	0.751	0.799	0.531
Temporal	0.167	0.514	0.720	0.773	0.643	0.688	0.568
Occipital	0.184	0.498	0.830	0.870	0.680	0.770	0.547
All	0.188	**0.521**	0.822	0.893	0.780	**0.844**	**0.516**
O+P	**0.2112**	0.515	**0.8441**	**0.911**	**0.785**	0.819	0.526

Table 13: Decoding with individual EEG bands and the full passband. Retrieval and classification values 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	43.50	75.90	15.06	41.49	7.83	27.02	2.34	9.63	0.1650	0.4900	0.7570	0.8360	0.7400	0.7580	0.5520
Theta	36.11	68.10	12.50	37.23	6.50	24.25	1.94	8.64	0.1360	0.3270	0.5110	0.5110	0.4770	0.5110	0.7360
Alpha	38.50	66.90	13.33	36.57	6.93	23.82	2.07	8.49	0.1070	0.2917	0.5090	0.5243	0.4830	0.5160	0.7400
Beta	30.20	59.40	10.45	32.47	5.44	21.15	1.63	7.54	0.1139	0.3010	0.5080	0.5460	0.4740	0.4930	0.7200
Gamma	11.20	31.30	3.88	17.11	2.02	11.14	0.60	3.97	0.1020	0.3080	0.5320	0.5230	0.5110	0.4900	0.7200
Full passband	**66.9**	**91.0**	**22.30**	**49.75**	**11.70**	**32.40**	**3.50**	**11.56**	**0.2112**	**0.5150**	**0.8441**	**0.9110**	**0.7850**	**0.8190**	**0.5260**

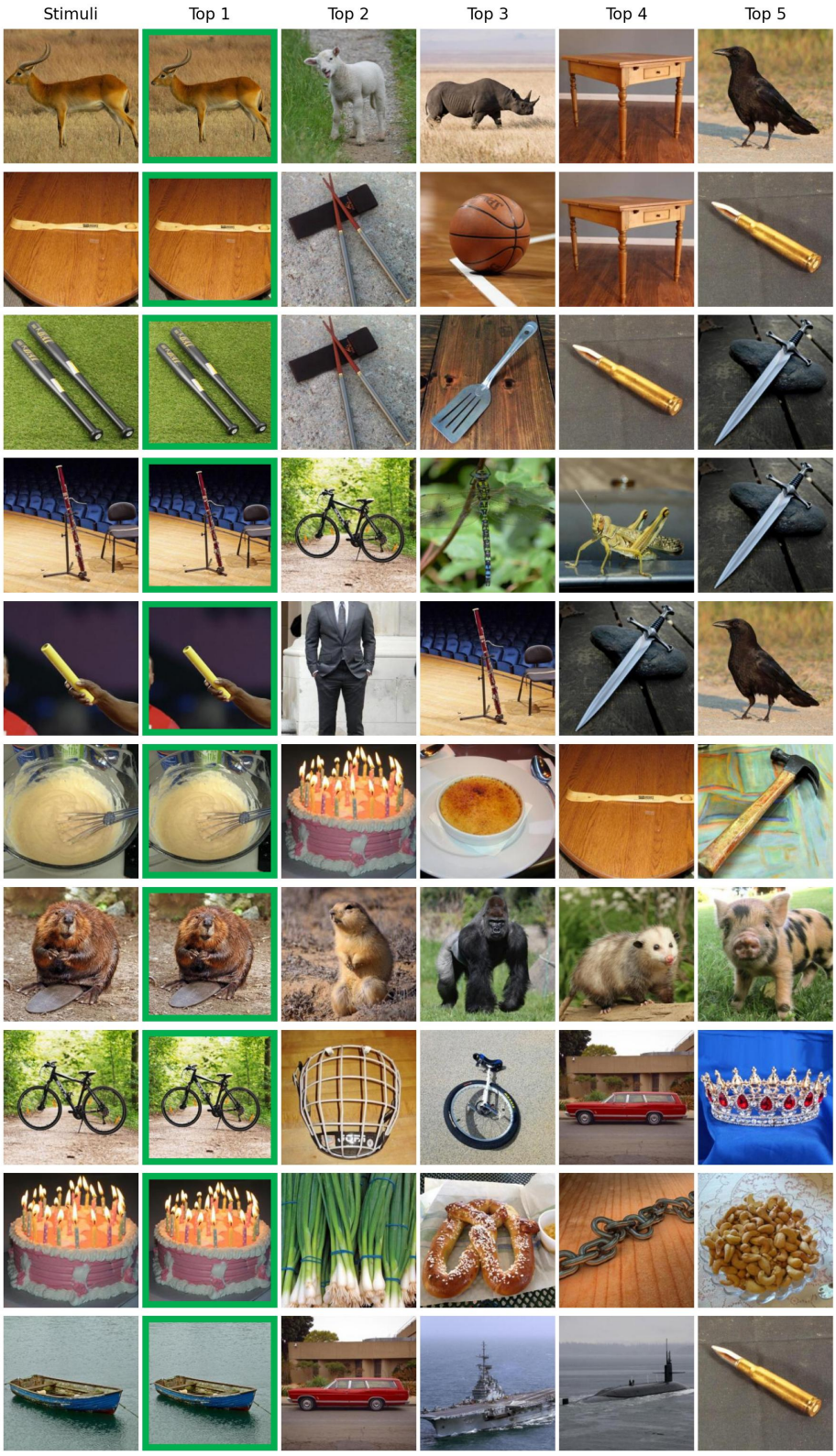

Figure 8: Successful Top-1 retrieval examples. Example group 1.

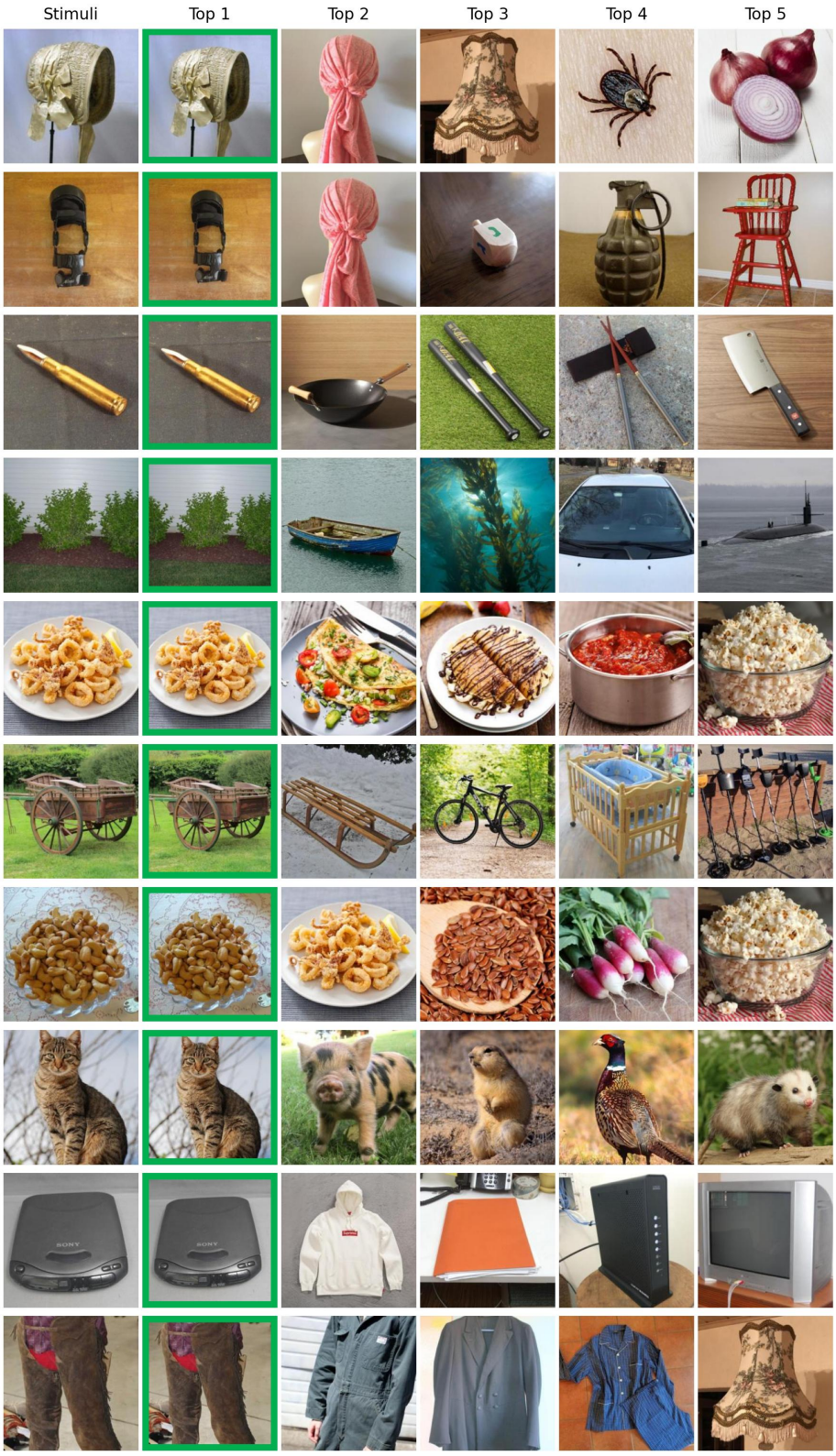

Figure 9: Successful Top-1 retrieval examples. Example group 2.

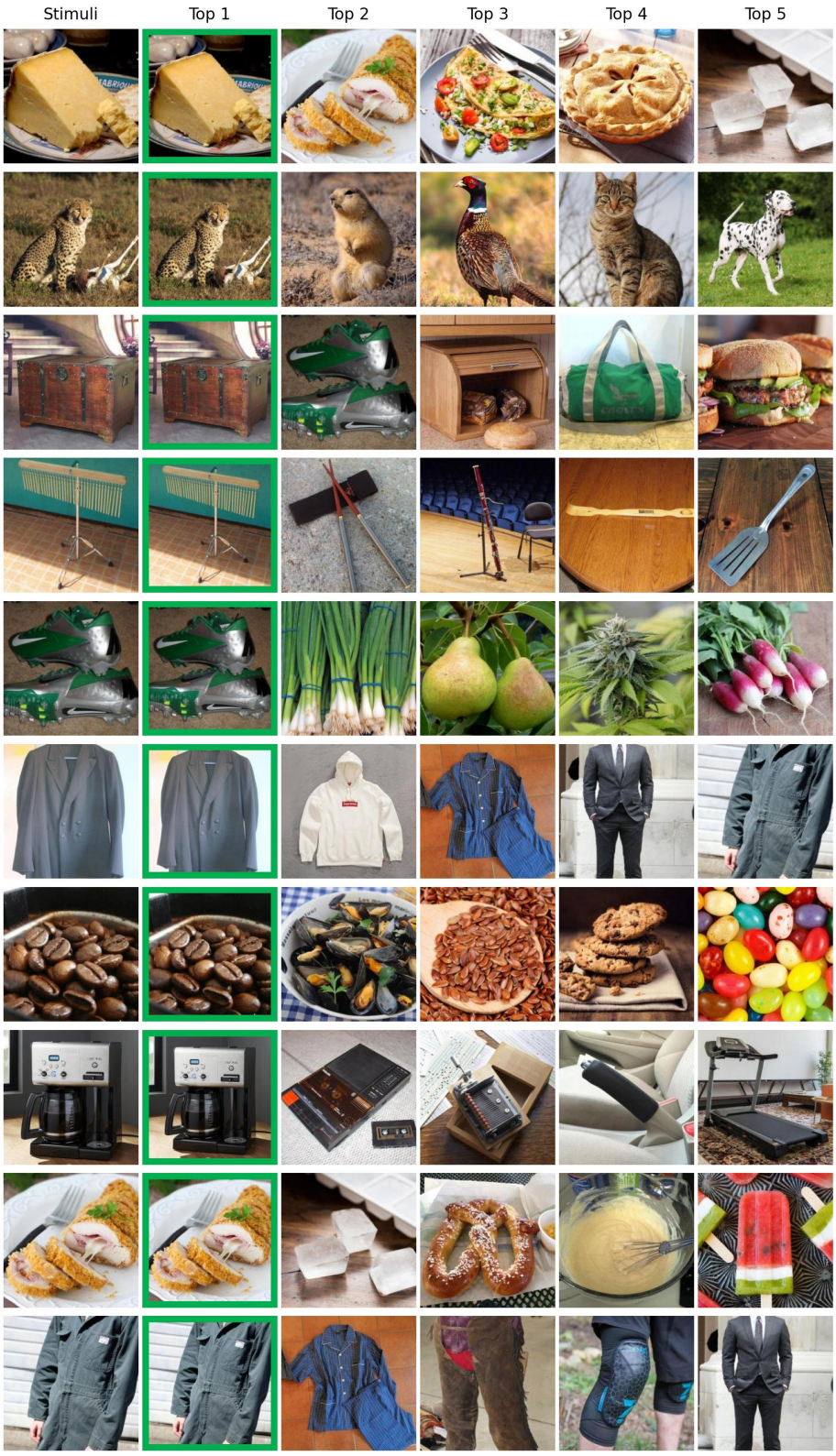

Figure 10: Successful Top-1 retrieval examples. Example group 3.

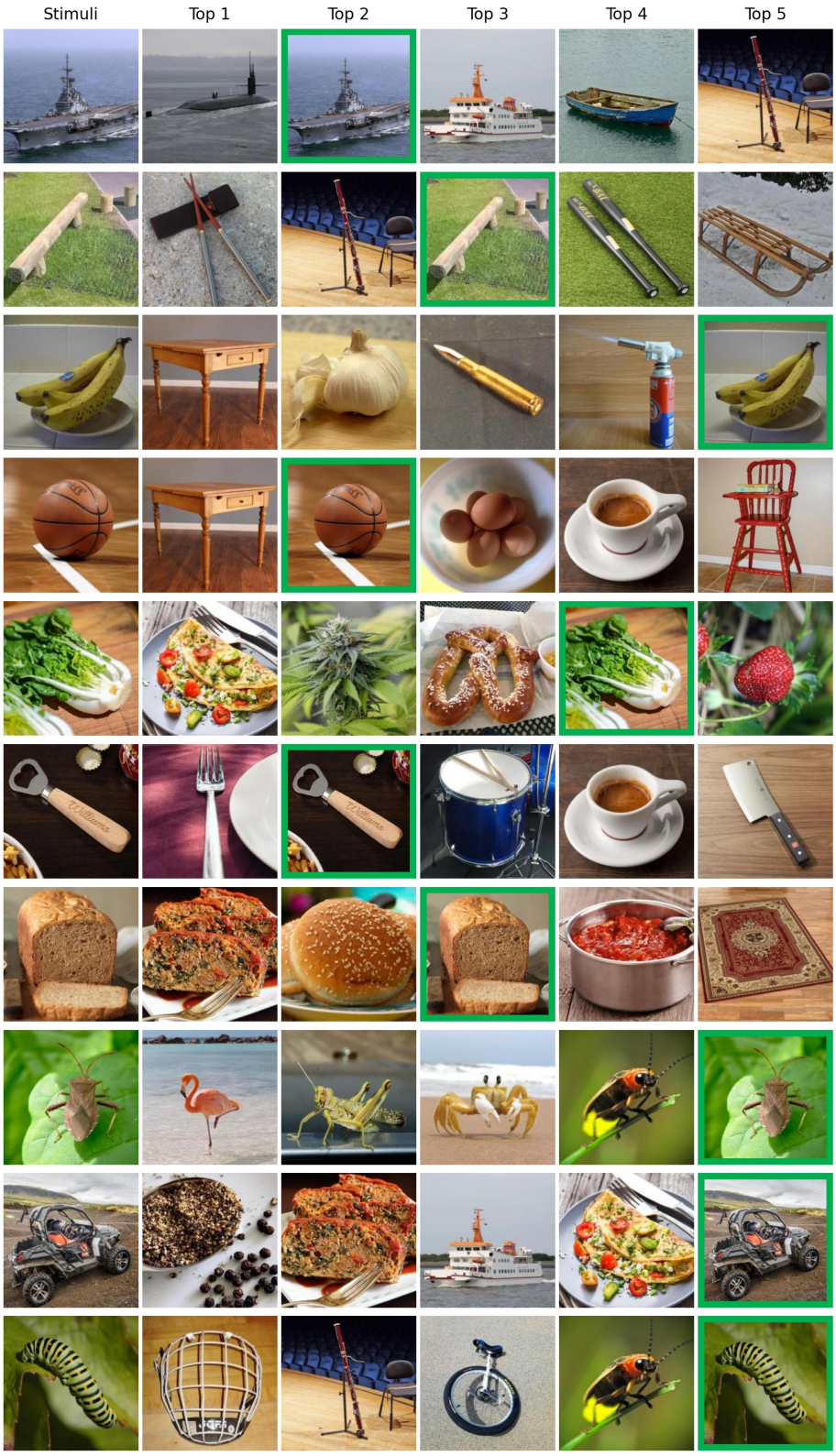

Figure 11: Top-5 retrieval hits with Top-1 misses. Example group 1.

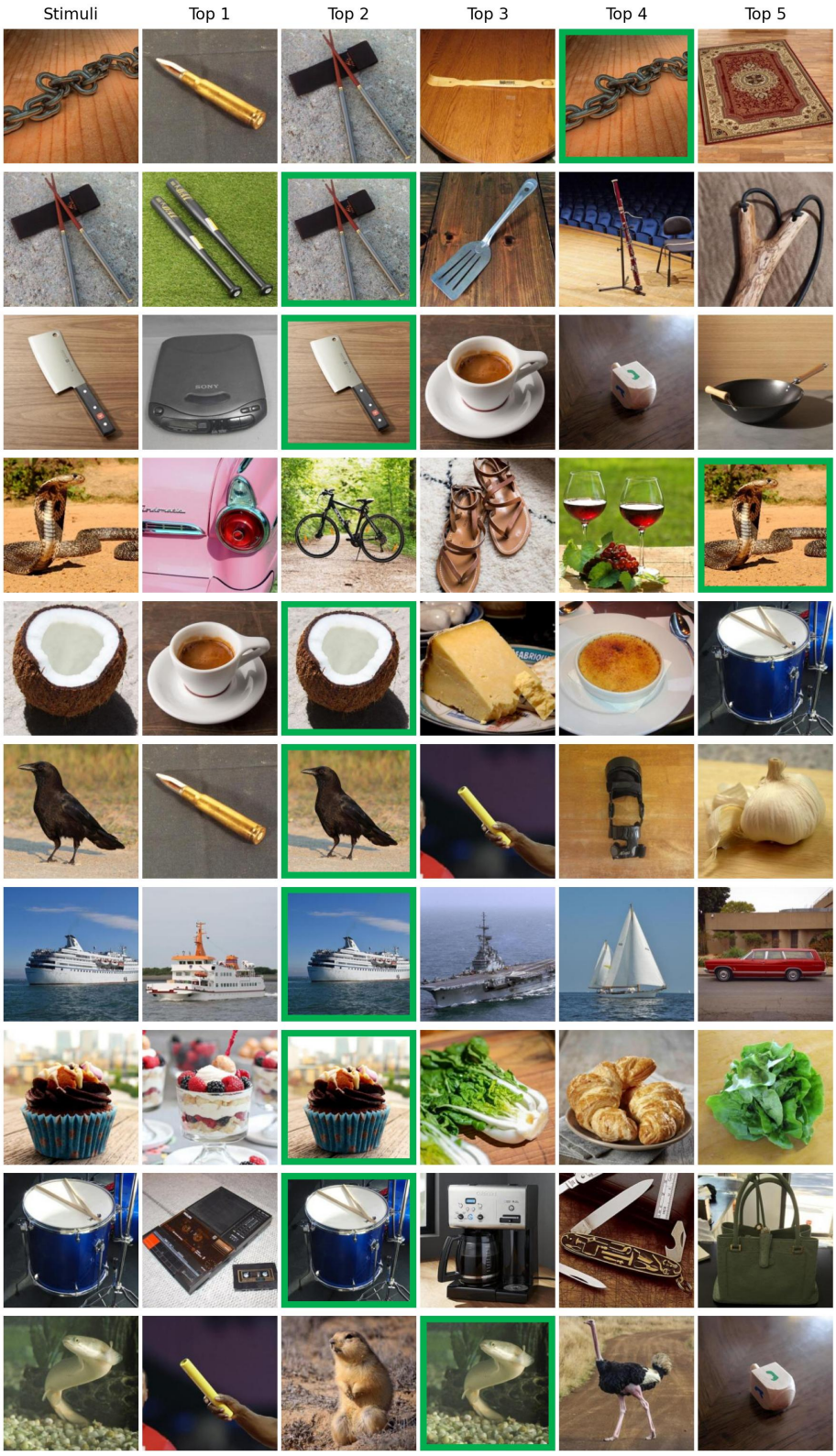

Figure 12: Top-5 retrieval hits with Top-1 misses. Example group 2.

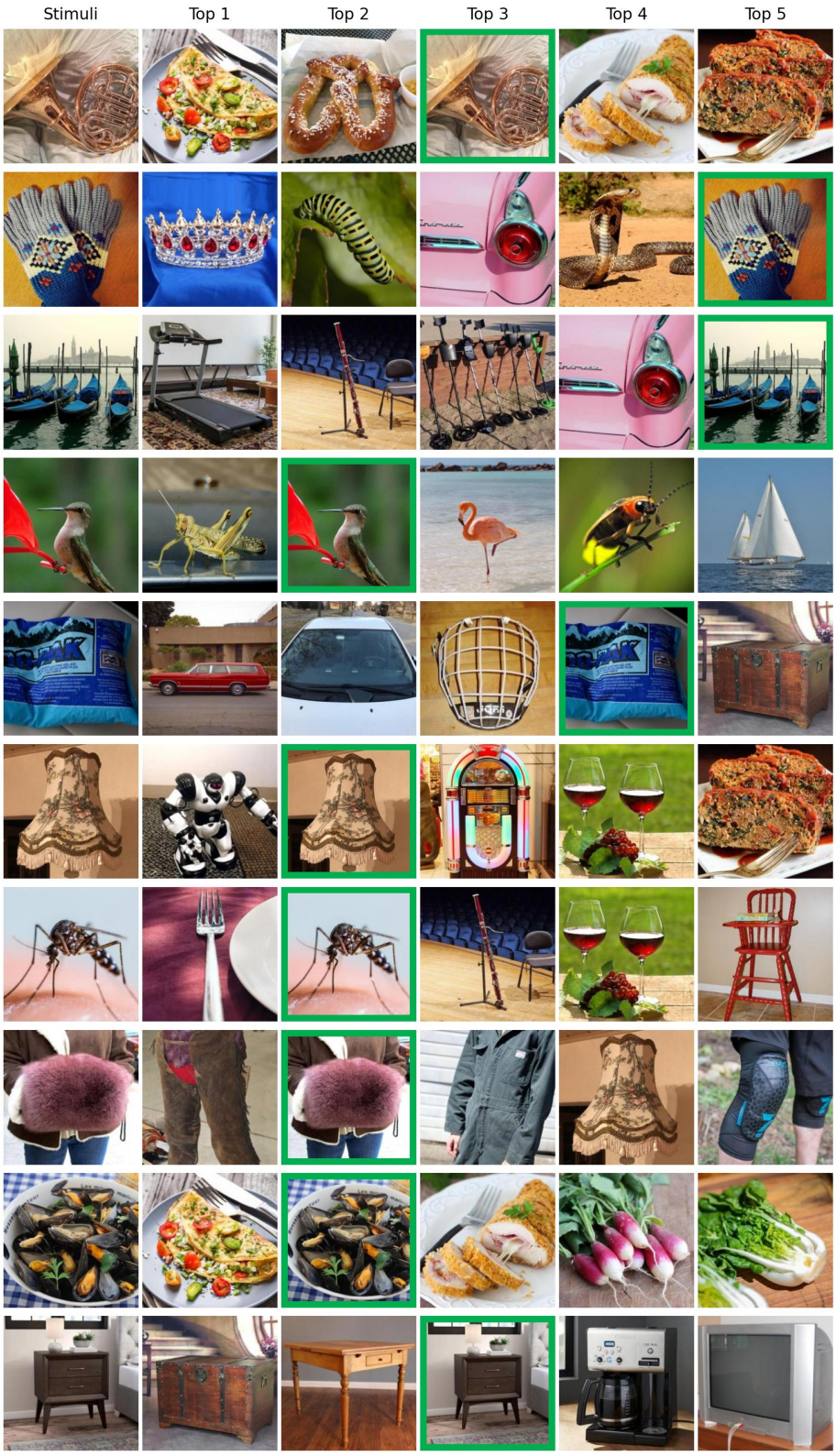

Figure 13: Top-5 retrieval hits with Top-1 misses. Example group 3.