RPA: RESIDUAL PATCH-TOKEN ADAPTER FOR IMAGE RETRIEVAL FROM EEG AND MEG

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual decoding from non-invasive MEG and EEG (M/EEG) aims to identify viewed images through representation learning that aligns brain signals with corresponding images. Most existing MEG and EEG (M/EEG) visual decoding methods align brain signals with a single global embedding extracted from a pretrained visual encoder, leaving open whether intermediate patch representations, which preserve richer and more granular rich visual information, can improve representation learning. To address this question, we introduce the Residual Patch Adapter (RPA), a lightweight, modular adapter that leverages all patch tokens from an intermediate layer of a ViT visual encoder for alignment. Through extensive ablation analyses, we first show that pooling or masking patch tokens degrades the learned representation, demonstrating that retaining the full set of patch tokens is important for EEG alignment, while the CLS token provides little unique information. We then use a series of six quantitative feature analyses to show that both higher-level semantics and lower-level visual features, including color and texture, are essential for this EEG-to-image alignment. Under current protocols, our system achieves Top-1 accuracies of 95.4% within-subject and 35.5% cross-subject on THINGS-EEG2, and 65.2% and 6.7%, respectively, on THINGS-MEG, achieving state-of-the-art (SOTA) performance across both datasets. Evaluations with alternative brain encoders, including pretrained EEG foundation models, demonstrate that the approach extends beyond the projection-based EEG encoder. Furthermore, we provide a plug-and-play interface that allows RPA to be replaced by convolution, attention, or ConvNeXt alternatives. Together, these findings provide significant insight into M/EEG-to-image representation learning by establishing design principles for leveraging the latent space of visual encoders, and open new directions for brain–image alignment and non-invasive brain–computer interface (BCI).

1 INTRODUCTION

Non-invasive visual decoding aims to identify viewed images from EEG or MEG responses. Contrastive approaches learn a shared embedding space in which a brain response retrieves its corresponding image from a held-out gallery (Song et al., 2024; Li et al., 2024). The visual embedding plays a central role in this framework: it provides the image-side target during training and represents each candidate image during retrieval. How this embedding is constructed therefore shapes the visual information that the brain encoder learns to recover.

Most existing systems ultimately represent each image with a single global embedding from a pretrained visual encoder. Recent work has enriched this image-side representation by transforming the stimulus, combining complementary visual targets, fusing multiple visual encoders, or adapting visual features to brain signals (Wu et al., 2025; Liu et al., 2026b; Zhang et al., 2026; Liu et al., 2026a; Jo et al., 2026; Wu et al., 2026; Zheng et al., 2026). Other studies show that intermediate visual-encoder layers, patch aggregation, and combinations across layers can improve alignment (Du et al., 2026; Tang et al., 2026; Jiang et al., 2026). However it remains unclear whether intermediate patch tokens should be compressed into a fixed global embedding or retained for learned adaptation. We therefore ask whether learning from the complete patch-token set can produce a more effective visual embedding for M/EEG-to-image alignment.

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.

To address this question, we introduce the *Residual Patch Adapter* (RPA), a lightweight, modular visual adapter applied to all patch tokens from a selected intermediate layer of a frozen visual encoder. RPA transforms the complete token set into a single visual embedding. A plug-and-play interface also allows RPA to be replaced by alternative visual adapters without changing the visual encoder or the brain–image alignment objective.

Our contributions are summarized as follows:

- We introduce the Residual Patch Adapter (RPA), a lightweight, modular visual adapter that processes all patch tokens from a selected intermediate layer before global pooling while keeping the pretrained visual encoder frozen. We further provide a plug-and-play interface for fixed-pooling, conv-based, attention-based, and ConvNeXt visual adapters, allowing the adapter to be changed without modifying the frozen visual encoder or alignment objective; RPA performs best in this comparison.
- Within the selected visual-encoder output, we find that visual information recoverable from EEG is distributed across multiple patch tokens and is not captured as effectively by the CLS token or fixed patch mean. This evidence empirically favors passing the patch tokens through an adaptive network before global pooling.
- Our system achieves 95.4% within-subject and 35.5% leave-one-subject-out(LOSO) Top-1 accuracy on THINGS-EEG2, and 65.2% 6.7% on THINGS-MEG, establishing SOTA across all settings. Further evaluation with pretrained EEG foundation models and other brain encoders shows that strong results are not limited to the a specific EEG encoder. Our EEG-to-image reconstruction pipeline also achieves SOTA on THINGS-EEG2 across all seven reported metrics.

2 Related Work

EEG and MEG visual decoding. Visual decoding aims to recover information about viewed images from recorded brain activity. EEG and MEG capture visually evoked responses with millisecond-level temporal resolution. Two common tasks are image retrieval, which identifies the viewed stimulus from a set of candidate images, and image reconstruction, which generates an image from the recorded brain response. Recent approaches address these tasks by aligning brain representations with features from pretrained visual encoders. Many existing methods use global representations from pretrained visual encoders, typically a final-layer CLS token or a pooled feature vector (Song et al., 2024; Li et al., 2024; Zheng et al., 2026). NICE uses contrastive EEG–image alignment for zero-shot object recognition, comparing each EEG embedding with one averaged image template for each unseen class (Song et al., 2024). ATM uses aligned EEG and CLIP embeddings for image retrieval and conditions a diffusion model on the EEG embedding for image reconstruction (Li et al., 2024). Reconstruction methods use several forms of visual supervision. CognitionCapturer learns separate EEG embeddings aligned with image, text, and depth embeddings (Zhang et al., 2025). NEED transfers a video-reconstruction model to unseen subjects and to static-image reconstruction without retraining (Huang et al., 2025). AVDE instead predicts multiscale visual tokens autoregressively from EEG (Dai et al., 2026). For cross-subject retrieval, SATTC uses statistics from the unlabeled test batch to recalibrate the EEG–image similarity matrix (Huang & Zhu, 2026). For MEG, Csaky et al. (2023) study many-class image decoding, and Benchetrit et al. (2024) align MEG with pretrained visual embeddings for image retrieval and reconstruction.

Constructing the visual embedding. Several methods construct the visual embedding from transformed images or multiple pretrained representations. UBP chooses the blur applied to each image from the estimated uncertainty of its paired EEG response, while Blur Perception extracts CLIP features at multiple blur levels and learns to select and adapt them (Wu et al., 2025; Liu et al., 2026b). NeuroBridge averages CLIP features from several augmented views of each image (Zhang et al., 2026). ViEEG constructs three CLIP embeddings from a binary object mask, a foreground-masked image, and the full scene, to align with three EEG streams (Liu et al., 2026a). HyFI interpolates two CLIP embeddings produced from foveally and Gaussian-blurred images in hyperbolic space (Jo et al., 2026). BrainHIVE combines pooled embeddings from multiple CLIP encoders with a flattened VAE representation (Zheng et al., 2026). Our comparison focuses on a single frozen layer and varies how its patch tokens are transformed before pooling.

Intermediate visual representations. Recent work examines visual-encoder layer selection and patch-token aggregation for EEG-to-image alignment. (Du et al., 2026) compares pooling methods over intermediate patch tokens. (Tang et al., 2026) averages patch tokens and combines embeddings from two intermediate layers. (Jiang et al., 2026) combines intermediate-layer embeddings using subject-specific weights. Building on these studies, we focus on adapting the complete patch-token set in to a visual embedding for alignment.

3 METHOD

3.1 TASK FORMULATION

Let $\mathcal{D} = \{(b_i, I_i)\}_{i=1}^{n}$ contain paired visually evoked brain responses and stimulus images, where $b_i \in \mathbb{R}^{C_b \times T}$ and $I_i \in \mathbb{R}^{H \times W \times 3}$. Here C_b is the number of brain-signal channels, T is the number of time samples, and H and W are the image height and width. The brain encoder and visual adapter produce d-dimensional vectors $h_{b,i}$ and $h_{v,i}$. We call these the M/EEG embedding and the visual embedding. For image retrieval, we rank the set of test images by cosine similarity to the brain embedding. Appendix A specifies the datasets and retrieval protocols.

3.2 LEARNING A VISUAL EMBEDDING FROM INTERMEDIATE PATCH TOKENS

Visual-encoder layer and visual adapter. We define a visual embedding by selecting an intermediate layer ℓ from a frozen visual encoder F and applying a trainable visual adapter G_θ to its output. Thus $h_{v,i} = G_\theta(F_\ell(I_i))$ is the visual embedding used for alignment. Throughout the paper, *visual adapter* refers to the complete trainable image-side mapping G_θ, whereas *visual embedding* refers to the normalized vector $z_{v,i}$ formed from that adapter's output. At a selected intermediate layer, a VIT based frozen visual encoder returns a CLS token and a grid of patch tokens,

$$F_\ell(I) = [c; X] \in \mathbb{R}^{(N+1)\times D}, \tag{1}$$

where c is the CLS token and $X = (x_1, \ldots, x_N)$ contains N patch tokens arranged on an $H_\ell \times W_\ell$ grid. The visual encoder is kept frozen. Appendix A gives the full configuration for M/EEG-to-Image retrieval task .

Visual adapters. The visual adapter G_θ maps the selected visual encoder features to a learned representation for brain–image alignment. In our approach, the adapter transforms the complete patch-token set before pooling, allowing the visual representation to be learned from individual patch tokens. A learned linear projection q maps the aggregated features to a d-dimensional vector h_v, which is normalized to obtain the visual embedding z_v.

We implement this mapping with the *Residual Patch Adapter* (RPA). We discard the CLS token and reshape the patch tokens into a $H_\ell \times W_\ell$ grid. RPA applies a residual convolutional branch A_θ to this grid and adds its output to the original patch-token features before pooling and projection. Let B_1, B_2, and B_3 denote three blocks, each applying a convolution followed by GroupNorm and GELU. Their kernel sizes are 1×1, 3×3, and 1×1, respectively, with channel dimensions $D \to d_a \to d_a \to D$, where d_a is the bottleneck width. The residual branch and complete visual adapter are

$$\begin{aligned} A_\theta(X) &= B_3\big(B_2(B_1(X))\big), \\ G_\theta(X) &= q\big(\mathrm{GAP}(X + A_\theta(X))\big). \end{aligned} \tag{2}$$

Here GAP denotes global average pooling over the patch grid, and q is the learned linear projection. Appendix C provides a set of alternative adapters.

3.3 BRAIN–IMAGE ALIGNMENT

Brain encoder. We adopt EEGProjection brain encoder to map each M/EEG response to a d-dimensional representation for alignment with the visual embedding. The brain encoder is trained jointly with the visual adapter through contrastive learning, while the pretrained visual encoder remains frozen. The implementation details are provided in the appendix A.

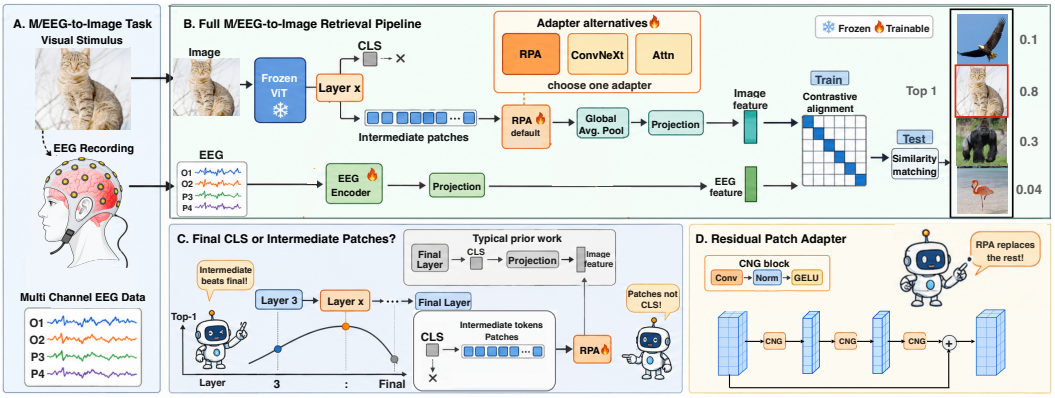

Figure 1: Overview of the EEG-to-image retrieval framework. **(A)** EEG is recorded while a participant views an image. **(B)** A frozen visual encoder and a trainable visual adapter produce the visual embedding, which is aligned with the brain embedding for retrieval. **(C)** The schematic contrasts a final-layer CLS route with an intermediate-patch route; the controlled comparison in Table 3 uses CLS and patch tokens from the same intermediate layer. **(D)** The residual adapter branch transforms the patch-token grid with convolution–normalization–GELU layers. Pooling and projection, shown separately in panel B, complete the visual adapter.

MEG channel projection. THINGS-MEG records C channels, making the flattened input to the brain encoder large. For an MEG trial $b \in \mathbb{R}^{C\times T}$, we learn a bias-free matrix $M \in \mathbb{R}^{K\times C}$ and compute

$$\widetilde{b} = Mb. \tag{3}$$

This produces $\widetilde{b} \in \mathbb{R}^{K\times T}$. We pass $\widetilde{b}$ to the brain encoder; EEG inputs do not use this operation. Appendix H reports the effect of the projection output dimension.

Contrastive learning objective. We use a CLIP-style symmetric InfoNCE objective (Radford et al., 2021) to align the brain and visual embeddings. For a minibatch of B paired normalized embeddings, the scaled cosine-similarity logits are

$$s_{ij} = \exp(\gamma) h_{b,i}^{\top} h_{v,i}, \tag{4}$$

where γ is a trainable logit scale. The alignment objective is

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ij})} + \log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ji})}\right]. \tag{5}$$

The two terms perform brain-to-image and image-to-brain matching, respectively. In each direction, the paired sample is the positive and the remaining batch elements serve as negatives. We jointly optimize the brain encoder, visual adapter and for MEG, the learned sensor projection. The supplement reports the detailed settings.

3.4 EEG-to-Image Reconstruction

Beyond retrieval, we evaluate EEG-to-image reconstruction with a separately trained dual-stream generation pipeline. Its brain encoder is aligned with the RPA visual embedding and a pooled CLIP-H embedding; two conditional priors then predict semantic and structural conditions for the image generator. Appendix I describes the pipeline, compares alignment objectives, and reports quantitative and qualitative results.

4 Experiments

We evaluate 200-way image retrieval on THINGS-EEG2 and THINGS-MEG under within-subject and leave-one-subject-out (LOSO) protocols. Unless stated otherwise, each test input averages 80

repeated EEG responses or 12 repeated MEG responses to the same image. EEG training averages four responses per image; MEG training uses single presentations. Appendix A describes the data, training configurations, and evaluation procedures. In this section, we first provide our quantitative result for retrievals and reconstructions. Then We examine visual-encoder depth, and compare adapters at a fixed image encoder layer and test patch-token ability for EEG decoding.We next provide results for reconstruction and visualization .

4.1 Quantitative Evaluation

We evaluate two tasks: M/EEG-to-image retrieval and EEG-to-image reconstruction. For retrieval, we report 200-way zero-shot Top-1 and Top-5 accuracy on THINGS-EEG2 and THINGS-MEG under within-subject and leave-one-subject-out (LOSO) protocols. For reconstruction, we evaluate low-level similarity using PixCorr and SSIM, and feature-level agreement using AlexNet-2/5, Inception, CLIP, and SwAV. Higher values are better for all metrics except SwAV. Results are averaged across participants.

Our method achieves the highest Top-1 and Top-5 accuracies across all four retrieval settings in Table 1. On THINGS-EEG2, within-subject Top-1/Top-5 accuracy reaches $95.4\%/99.8\%$, compared with $84.6\%/98.2\%$ for HCF, while LOSO accuracy reaches $35.5\%/67.4\%$, compared with $23.4\%/54.9\%$. On THINGS-MEG, our method achieves $65.2\%/85.7\%$ within subject and $6.7\%/19.2\%$ under LOSO, outperforming the reported baselines in both settings. These improvements extend across EEG and MEG and across both evaluations. Subject-level results and additional baselines are provided in Appendix B.

For EEG-to-image reconstruction, we evaluate the separately trained EEG encoder aligned with both the RPA visual embedding and the pooled CLIP-H embedding, using the pipeline described in Appendix I.1. Table 2 reports results across 10 THINGS-EEG2 participants. Our system obtains the best reported value on all seven metrics. Together, these results show higher reconstruction similarity at both the pixel and feature levels.

We further examine whether RPA can be paired with alternative brain encoders, including pretrained EEG foundation models. Keeping the frozen visual encoder and RPA architecture unchanged, we replace the EEGProjection encoder with conventional EEG foundation models. EEG-Conformer (without pretraining) and CBraMod (pretrained) achieve within-subject/LOSO Top-1 accuracies of $94.90\%/34.45\%$ and $94.55\%/35.63\%$, respectively, comparable to those of the EEGProjection encoder. Full results are reported in Appendix F. The results show that strong retrieval performance with RPA is not limited to the EEGProjection and extends to a set of EEG foundation model.

4.2 From Images to Intermediate Visual Representations

In Figure 2, we compare seven frozen visual encoders, OpenCLIP ViT-H/14, CLIP ViT-B/32 and RN50 (Radford et al., 2021; Ilharco et al., 2021), I-JEPA (Assran et al., 2023), DINOv2-S/L (Oquab et al., 2024), and SDXL-VAE (Podell et al., 2024), at six relative depths under within-subject and LOSO on M/EEG. Scores are standardized within each encoder. Although the best depth varies across image encoders, the broadly inverted-U-shaped depth shape appears in all image encoders. This trend aligns with the hierarchical progression of visual representations across encoder layers, in which intermediate layers combine low- and high-level image information. In contrast, deeper layers become more specialized for pretraining objectives, such asimage–text alignment, self-supervised learning, or reconstruction. Appendix D reports absolute scores and procedures.

4.3 From Patch Tokens to a Visual Embedding

We next examine whether adapting the complete patch-token set produces a more effective visual embedding than using the CLS token or the patch mean. We compare the CLS-token and patch-mean baselines with pointwise RPA and RPA. As shown in Table 3, the patch-mean achieves higher accuracy than the CLS-token , while pointwise RPA provides a further substantial improvement. These results show that adapting the complete patch-token set before pooling yields higher retrieval accuracy than using the CLS token or patch mean.

Table 1: Average Top-1/Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG2 and THINGS-MEG. Values are participant averages; "–" indicates not reported.

	EEG				MEG			
	Within		Cross		Within		Cross	
Method	T1	T5	T1	T5	T1	T5	T1	T5
NICE (Song et al., 2024)	13.8	39.5	6.2	21.4	12.8	36.0	–	–
ATM (Li et al., 2024)	28.6	58.5	11.8	33.7	–	–	–	–
UBP (Wu et al., 2025)	50.9	79.7	12.4	33.4	26.7	55.2	2.2	10.4
BrainHIVE (Zheng et al., 2026)	75.7	94.6	20.0	44.1	33.7	60.5	5.4	15.2
Blur Perception (Liu et al., 2026b)	80.0	96.9	20.0	48.0	44.0	72.0	5.3	15.9
Shallow Alignment (Du et al., 2026)	82.6	97.7	21.8	49.4	48.0	74.4	3.7	13.0
HCF (Tang et al., 2026)	84.6	98.2	23.4	54.9	–	–	–	–
Ours	$\mathbf{95.4 \pm 0.1}$	$\mathbf{99.8 \pm 0.0}$	$\mathbf{35.5 \pm 0.3}$	$\mathbf{67.4 \pm 0.7}$	$\mathbf{65.2 \pm 1.0}$	$\mathbf{85.7 \pm 0.8}$	$\mathbf{6.7 \pm 0.4}$	$\mathbf{19.2 \pm 1.0}$

Table 2: Average image-reconstruction performance on THINGS-EEG2.

	Low-level		High-level				
Method	PixCorr ↑	SSIM ↑	AlexNet-2 ↑	AlexNet-5 ↑	Inception ↑	CLIP ↑	SwAV ↓
CognitionCapturer (All) (Zhang et al., 2025)	0.150	0.347	0.754	0.623	0.669	0.715	0.590
AVDE (Dai et al., 2026)	0.147	0.366	0.766	0.835	0.724	0.747	0.587
BrainHIVE (Zheng et al., 2026)	0.195	0.336	0.843	0.905	0.756	0.808	0.554
Ours	**0.225**	**0.454**	**0.865**	**0.909**	**0.783**	**0.811**	**0.526**

We further examine how RPA compares with alternative visual adapters. Using the same frozen patch-token input, we compare local RPA with attention pooling, transformer-based adapter, ConvNeXt, and depthwise-separable convolution. As reported in Appendix, local RPA achieves the highest mean accuracy in both within-subject and LOSO evaluation. The alternative learned adapters have more trainable parameters but lower accuracies, supporting RPA as a lightweight and effective visual adapter for EEG-to-image alignment. The adapters evaluated in these ablations also form a plug-and-play toolkit for frozen visual encoders.

Table 3: Controlled within-subject THINGS-EEG2 comparison at a fixed visual-encoder output.

Visual input	Visual adapter	Top-1 (%)
CLS token	LayerNorm + linear	69.50 ± 0.45
patch tokens	mean, LayerNorm + linear	77.03 ± 0.7
patch tokens	pointwise RPA	92.98 ± 0.2
patch tokens	local RPA	$\mathbf{95.4 \pm 0.1}$

4.4 Patch-Token Interventions and Visual-Feature Probes

To understand why RPA benefits from the complete patch-token set, we first examine whether simply averaging more frozen tokens improves their predictability from EEG. The plateau in this linear probe motivates a further question: does learned adaptation still benefit from retaining individual tokens when averaging more tokens provides little improvement? We therefore use pooling and masking to test whether merging or replacing tokens limits RPA, while permutation examines the additional contribution of spatial arrangement. Finally, we probe six image attributes to help interpret these token-level findings in terms of the visual information predictable from EEG.

Averaging frozen patch tokens. We next examine whether averaging more patch tokens produces visual representations that are more predictable from EEG. At a fixed visual-encoder layer, we average patch-token subsets of different sizes. For each subset size, we fit a ridge regression model to predict the averaged features from EEG and use the predictions to retrieve the paired images. This experiment uses frozen visual features without RPA. As shown in Figure 5**C**, averaging larger subsets yields higher Top-1 accuracy than using a single patch token. These results suggest that aggregating multiple patch tokens better captures visual information predictable from EEG than using a single token. Although accuracy in this linear probe plateaus as more patch tokens are averaged, our pooling and masking ablations in the follow section show that RPA benefits from retaining and

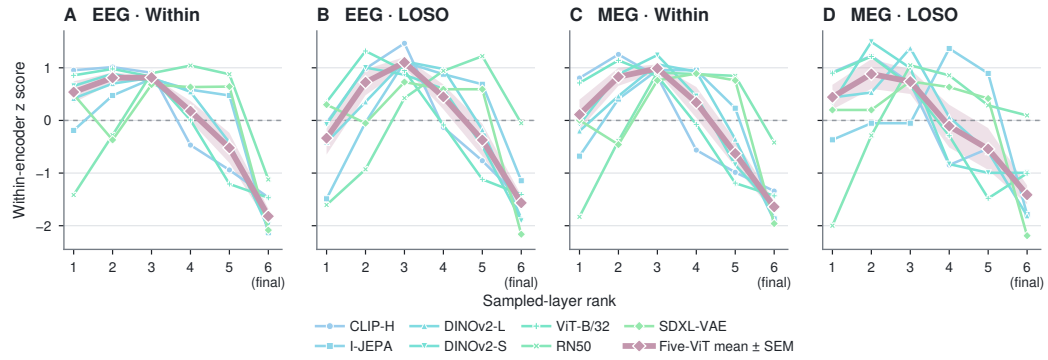

Figure 2: Visual-encoder depth profiles for EEG and MEG under within-subject and LOSO evaluation. The x-axis shows six uniformly spaced relative depths. Curves show mean 200-way Top-1 scores, averaged across three independent runs and standardized within each encoder.

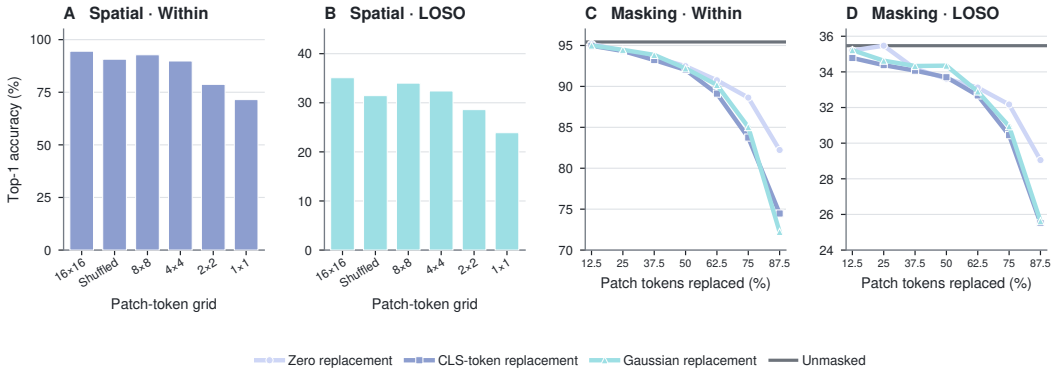

Figure 3: **RPA diagnostics. (A–B)** Top-1 accuracy for native, shuffled, and spatially averaged patch-token grids. **(C–D)** Top-1 accuracy under increasing zero, CLS-token, or Gaussian replacement. Solid gray lines mark unmasked performance.

adapting the complete patch-token set before pooling. Full experimental details are provided in Appendix E.1.

Patch-token pooling and spatial arrangement. At a fixed visual-encoder layer, we apply local average pooling to the original 16×16 patch-token grid before training RPA. We use pooling blocks of 2×2,..., 16×16, with the last condition corresponding to global mean pooling. Larger pooling blocks merge more patch tokens, reducing the number of distinct token representations available to RPA when forming the visual embedding. As shown in Figure 3**A–B**, Top-1 accuracy decreases as the pooling block size increases in both within-subject and LOSO evaluation. These results indicate retaining the full patch-token set for adaptation provide enougn information to form visual embedding.

To examine the contribution of spatial arrangement, we also permute the original patch tokens on the grid and retrain RPA using the permuted inputs during both training and evaluation. This preserves every token's values while disrupting the original spatial arrangement presented to the adapter. Within-subject Top-1 accuracy decreases from 95.40% to 90.70%, indicating that preserving the original arrangement provides an additional benefit for EEG-to-image retrieval.

Masking patch tokens. To complement the pooling experiment, we examine how retrieval accuracy changes when RPA has access to fewer original patch tokens. At the same fixed visual-encoder layer, we independently mask each patch token with probability p, replacing it with the image's CLS

token, Gaussian noise, or a zero vector. Unlike pooling, this operation preserves the patch-token grid while replacing selected token values. We retrain RPA for each condition, using the same masking rate and replacement strategy during training and evaluation.

As shown in Figure 3**C–D**, none of the masking conditions outperforms the unmasked baseline, and Top-1 accuracy generally decreases as the masking rate increases in both within-subject and LOSO evaluation. At high masking rates, zero replacement outperforms CLS-token replacement, indicating that filling masked positions with the CLS token provides no advantage over leaving them zero. Together with the pooling experiment, these experiments show that preserving the complete set of patch-token for RPA provide all information needed for EEG decoding.

Image attributes predicted from EEG. We next examine which image attributes can be predicted from EEG. We represent each stimulus image using features describing six image attributes: (1) object-category semantics, (2) color, (3) texture, (4) objectness and two-dimensional layout, (5) oriented edges, and (6) object silhouette. For each attribute, we fit a separate ridge regression model to predict the corresponding image features from EEG, then use the predicted features to retrieve the paired images. As reported in Appendix E.2, object-category semantics, color, and texture yield the highest Top-1 accuracies. These results indicate that image attributes predictable from EEG extend beyond object-category semantics to lower-level visual properties, particularly color and texture. These findings may help explain the strong EEG-to-image decoding performance achieved using the complete set of patch tokens from an intermediate layer.

4.5 EEG-to-Image Reconstruction and Alignment Diagnostics

EEG-to-image reconstruction. Figure 4 shows the viewed image and all 10 participant reconstructions for three animal concepts. Additional examples are reported in the Appendix I.3.

Visualizing the aligned EEG embedding space. We extract EEG embeddings for all test images from each subject's within-subject model. For visualization, we align them across participants using orthogonal Procrustes transformations fitted to the corresponding visual embeddings. We align EEG embeddings from four categories images and project them to two dimensions using UMAP (McInnes et al., 2018). Each point in Figure 5**A** represents one subject–image pair. The details of this experiment goes into the appendix E.3. The plot shows category-level organization in this selected subset. Although the model is not explicitly trained for four-category classification, the aligned EEG embeddings tend to cluster by category in the visualization. This suggests that EEG embeddings learned through RPA capture category-level semantic structure.

Visual-encoder depth and EEG latency. We next examine whether visual representations from deeper encoder layers are best predicted from later EEG activity. We use ridge regression probes to predict frozen visual features from successive EEG windows. As shown in Figure 5**B**, all selected layers reach their highest decoding accuracy in windows centered between 110 and 230 ms, with most peaking at 170 ms, which temporally overlaps with the N170, an occipitotemporal component associated with early face processing (Bentin et al., 1996). Accuracy then declines and approaches the chance level after 400 ms. These results indicate that visual representations are most predictable from EEG within early intervals. Full experimental details are provided in Appendix E.3.

Grad-CAM visualization. We next examine which image regions RPA highlights under EEG alignment and image-only training using Grad-CAM (Selvaraju et al., 2017). For image-only RPA, the maps are computed from similarity to the final CLIP image embedding. Figure 5**E** compares the normalized Grad-CAM maps for eight selected test images. In these examples, EEG-aligned RPA produces maps concentrated on foreground objects, whereas image-only RPA shows less foreground focus. This contrast suggests that EEG alignment helps RPA emphasize foreground-related visual information. The visualization procedure are provided in Appendix E.3.

Test-repetition averaging Recall that for test data, each data clip is the mean over 80 test trial given a fixed image stimulus. In Figure 5**D** we do a experiemnt for a test count k, repetitions are partitioned into disjoint groups of size k, each group is averaged, and group accuracies are pooled . Single-test-trial accuracy exceeds chance, and accuracy increases as more responses are averaged.

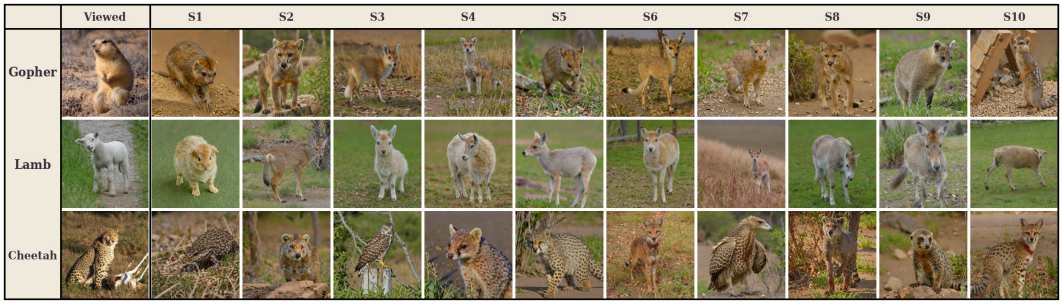

Figure 4: **Qualitative EEG-to-image reconstructions.** Each row shows one animal concept: the viewed image followed by reconstructions from Subjects 1–10.

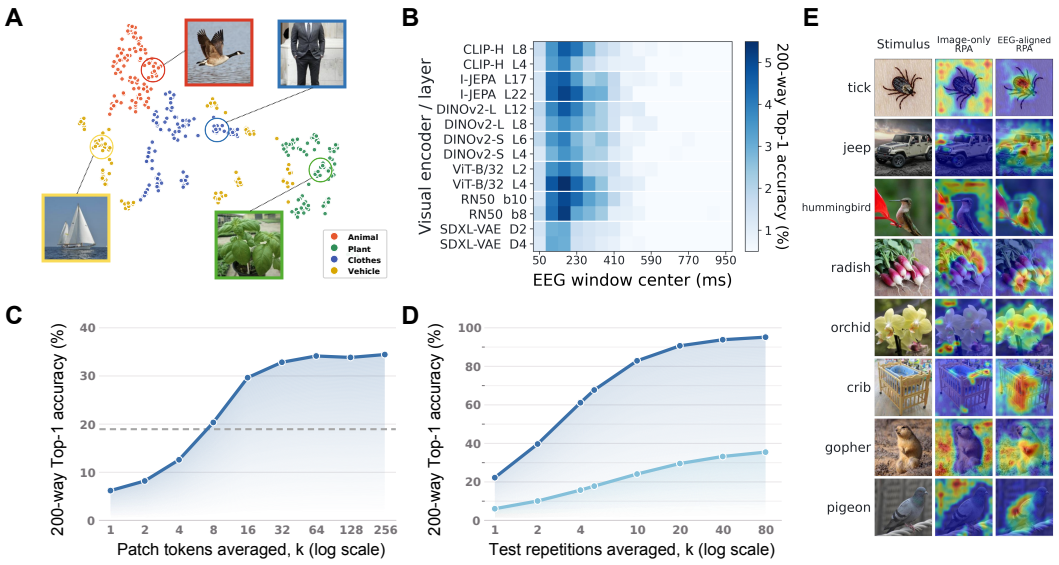

Figure 5: **Alignment, timing, averaging, and attribution diagnostics. (A)** UMAP projection of EEG embeddings; colors indicate categories, and insets show example stimuli. **(B)** Given layers of visual encoder and 100-ms EEG windows advanced in 60-ms steps, we using a ridge probe to obtain a decoding accuracy. **(C)** Retrieval with ridge-predicted means of frozen patch-token subsets; the dashed line marks the CLS-token result. **(D)** Top-1 accuracy versus the number of EEG test trials averaged per stimulus; dark and light blue denote within-subject and LOSO evaluation. **(E)** Grad-CAM maps; columns show the stimulus, image-only RPA, and RPA.

5 Conclusion

This study proposed the Residual Patch Adapter (RPA), a novel lightweight visual adapter that moves beyond global embeddings to learn visual representations from the complete patch-token set. Ablation studies involving intermediate layer analysis, patch-token pooling and masking demonstrate the effectiveness of RPA, which achieves state-of-the-art performance in both within-subject and leave-one-subject-out M/EEG-to-image retrieval. The contributions of semantics, color, and texture to EEG decoding further support RPA's use of diverse visual features. We further develop a suite of adapters that provides a plug-and-play interface for various visual encoders. Together, our work offers an effective and flexible approach to representation learning for brain-to-image decoding, with significant implications for brain–computer interface applications and future research on brain–image alignment.

AI USE STATEMENT

We used large language models (LLMs) to assist with manuscript drafting and revision, including grammar checking, language editing, and feedback on clarity and paragraph organization. The authors remain responsible for evaluating AI-assisted suggestions and verifying the text, references, and numerical claims against original sources and experimental records. The authors take full responsibility for the final manuscript, research decisions, and reported results.

ETHICS STATEMENT

This study analyzes previously released THINGS-EEG2 and THINGS-MEG data and introduces no new data collection or intervention with human participants. We follow the datasets' terms of use and report aggregate and participant-level retrieval results, together with image reconstructions. The evaluation concerns responses to controlled visual stimuli, and the strongest results use averages over repeated presentations. These conditions limit conclusions about unrestricted or real-time brain decoding.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: Toward real-time reconstruction of visual perception. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3y1K6buO8c.

Shlomo Bentin, Truett Allison, Aina Puce, Erik Perez, and Gregory McCarthy. Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, 8(6):551–565, 1996. doi: 10.1162/jocn.1996.8.6.551.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.

Richard Csaky, Mats WJ van Es, Oiwi Parker Jones, and Mark Woolrich. Interpretable many-class decoding for meg. *NeuroImage*, 282:120396, 2023.

Sicheng Dai, Hongwang Xiao, Shan Yu, and Qiwei Ye. Autoregressive visual decoding from EEG signals. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TKjfzuVLX4.

Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2023. doi: 10.1109/TAFFC.2022.3169001.

Yang Du, Siyuan Dai, Yonghao Song, Paul M. Thompson, Haoteng Tang, and Liang Zhan. Deep models, shallow alignment: Uncovering the granularity mismatch in neural decoding, 2026. URL https://arxiv.org/abs/2601.21948.

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

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

Qunjie Huang and Weina Zhu. SATTC: Structure-aware label-free test-time calibration for cross-subject EEG-to-image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16887–16896, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Huang_SATTC_Structure-Aware_Label-Free_Test-Time_Calibration_for_Cross-Subject_EEG-to-Image_Retrieval_CVPR_2026_paper.html.

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, pp. 173134–173173, 2025. doi: 10.52202/085713-5761. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fca83589e85cb061631b7ebc5db5d6bd-Abstract-Conference.html.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021.

Lin Jiang, Qingshan She, Jiale Xu, Haiqi Xu, Duanpo Wu, and Zhenzhong Kuang. Subject-aware multi-granularity alignment for zero-shot EEG-to-image retrieval, 2026. URL https://arxiv.org/abs/2604.17782.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Sangmin Jo, Wootaek Jeong, Da-Woon Heo, Yoohwan Hwang, and Heung-Il Suk. HyFI: Hyperbolic feature interpolation for brain-vision alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(7):5575–5583, 2026. doi: 10.1609/aaai.v40i7.37476. URL https://ojs.aaai.org/index.php/AAAI/article/view/37476.

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.

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. doi: 10.52202/079017-3266. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ba5f1233efa77787ff9ec015877dbd1f-Abstract.html.

Minxu Liu, Donghai Guan, Chuhang Zheng, Chunwei Tian, Jie Wen, and Qi Zhu. ViEEG: Hierarchical visual neural representation for EEG brain decoding. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=DkK7GUr8n3.

Wenchao Liu, Hongwei Li, Zhouyang Xu, Lin Ma, and Haifeng Li. Leveraging visual blur perception characteristics for EEG decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17580–17588, 2026b. doi: 10.1609/aaai.v40i21.38813. URL https://ojs.aaai.org/index.php/AAAI/article/view/38813.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

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 *The Twelfth International Conference on Learning Representations*, 2024.

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*, pp. 8748–8763, 2021.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, 2024.

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.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74. URL https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

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.

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

Jingyi Tang, Shuai Jiang, Fei Su, and Zhicheng Zhao. Aligning what EEG can see: Structural representations for brain-vision matching, 2026. URL https://arxiv.org/abs/2603.07077.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

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. doi: 10.1109/CVPR52734.2025.00215. 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.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract.html.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human EEG signal with multimodal information. *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.

Wenjiang Zhang, Sifeng Wang, Yuwei Su, Xinyu Li, Chen Zhang, and Suyu Zhong. NeuroBridge: Bio-inspired self-supervised EEG-to-image decoding via cognitive priors and bidirectional semantic alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):18028–18036, 2026. doi: 10.1609/aaai.v40i21.38863. URL https://ojs.aaai.org/index.php/AAAI/article/view/38863.

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IEq71qS8B7.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2915. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7e199ad8ae40eb19b2980f61f659cb07-Abstract-Conference.html.

A Datasets, Training, and Evaluation Protocols

This appendix describes the datasets, training configurations, and evaluation procedures used in Section 3 and the experiments. The matched adapter comparison, benchmark system, and supplementary sweeps are reported separately; their reference accuracies should be interpreted within the configurations and replication schemes stated for each experiment.

A.1 Datasets and preprocessing

THINGS-EEG2. (Gifford et al., 2022) The primary dataset contains recordings from 10 participants. The distributed data are sampled at 250 Hz, preprocessed, and whitened. The training split has 16,540 images from 1,654 concepts; each image has four repetitions. The test split has 200 images from 200 unseen concepts; each image has 80 repetitions. There is no train–test concept overlap. Unless an ablation states otherwise, the four training repetitions and 80 test repetitions are averaged before the model receives them. Each standard test input is therefore the mean of 80 responses to one image. The default EEG input comprises 17 anatomically specified posterior channels: P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, and O2. This set was selected by scalp location rather than retrieval performance.

THINGS-MEG. (Hebart et al., 2023) The MEG experiments use four participants, 271 sensors, 19,848 single-presentation training images, and 200 test images with 12 repetitions. Test repetitions are averaged. The learned sensor projection is fit only from the training split: within subject it is fit on that participant's training data; under LOSO it is fit on the other participants. Both benchmark MEG settings use projection output dimension $K = 64$.

A.2 WITHIN-SUBJECT AND LOSO TRAINING CONFIGURATIONS

Table 4 gives the within-subject and LOSO training configurations. They differ in the data split, input duration, and optimization settings. The difference in their accuracies therefore reflects these configuration changes as well as the subject-generalization task.

Table 4: Within-subject and LOSO EEG training configurations.

Setting	Within subject	LOSO
Training responses	16,540 from one subject	148,860 from nine subjects
Test subject	same subject, unseen concepts	held-out subject, unseen concepts
Default window	0–700 ms	0–1,000 ms
Epochs	25	15
Peak learning rate	5×10^{-4}	3×10^{-4}
Projection dropout	0.5	0.7
Batch size	512	512
Validation fraction	3%	3%
Checkpoint rule	best validation retrieval	best validation retrieval

A.3 MODEL AND OPTIMIZATION DETAILS

We adopt the EEGProjection encoder from UBP (Wu et al., 2025) to map each M/EEG response to a d-dimensional representation for alignment with the visual embedding.

Visual input and adapter configuration. The default visual input is the cached OpenCLIP ViT-H/14 layer-11 output, containing one CLS token and a 16×16 grid of patch tokens with feature dimension $D = 1{,}280$. These tokens have already passed through transformer self-attention and can encode information beyond their corresponding image patches. RPA uses a bottleneck width of $d_a = 256$, 32 groups in each GroupNorm layer, and an output dimension of $d = 1{,}024$.

The pointwise RPA control replaces the middle 3×3 convolution with a 1×1 convolution, giving kernel sizes $(1, 1, 1)$. Although this removes explicit local mixing, GroupNorm still couples patch positions by aggregating statistics across them.

Optimization. We use AdamW with $\beta = (0.9, 0.995)$, weight decay 0.05, gradient-norm clipping at 1.0, and cosine learning-rate decay. The logit scale is initialized as $\gamma_0 = \log(1/0.07)$, corresponding to an initial temperature of 0.07.

A.4 RETRIEVAL EVALUATION AND MODEL SELECTION

For each brain query, evaluation ranks the 200 held-out stimulus images by cosine similarity. Within-subject evaluation trains a separate model for each subject and evaluates it on that subject's unseen test concepts. Leave-one-subject-out (LOSO) evaluation trains on the remaining participants and evaluates the held-out participant. For each fixed configuration, we evaluate the test set at the validation-selected checkpoint. For benchmark results, accuracy is first averaged across participants within each training run; $\pm$ denotes the SD across the three participant-averaged run means.

A.5 TRAINING AND RETRIEVAL ALGORITHM

Algorithm 1 summarizes RPA training and M/EEG-to-image retrieval. We first extract and cache the complete patch-token grid from a fixed layer of the frozen visual encoder, discarding the CLS token. RPA transforms this grid before global average pooling and projection. We jointly train RPA and the brain encoder with symmetric contrastive learning, together with the logit scale and, for MEG, the sensor projection. After restoring the checkpoint with the highest validation retrieval accuracy, we encode each candidate image and rank the candidates by cosine similarity to the query's brain embedding.

Let f_ϕ denote the brain encoder and $\nu(a) = a/\|a\|_2$. The trainable parameters are $\Phi = \{\theta, \phi, \gamma\}$ for EEG, with M added for MEG; θ includes the residual branch and projection q. $\mathrm{PatchGrid}$ discards

the CLS token, retains all patch tokens, and reshapes them to $D \times H_\ell \times W_\ell$. Each $\mathrm{CNG}_{k\times k}^{a\leftarrow b}$ maps b channels to a using a convolution followed by GroupNorm(32) and GELU; d_a is the bottleneck width. GAP denotes global average pooling, and CE denotes mean row-wise cross-entropy. The procedure uses the preprocessing, regularization, and optimization settings of the corresponding within-subject or LOSO protocol.

B Subject-Level Benchmark Comparisons

These tables report subject-level Top-1/Top-5 retrieval. Baselines are published values; ours is mean ± SD across three repeated training runs, "–" is unreported.

Table 5: Subject-level within-subject retrieval on THINGS-EEG2.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Mean
Top-1 accuracy (%)											
NICE (Song et al., 2024)	12.3	10.4	13.1	16.4	8.0	14.1	15.2	20.0	13.3	14.9	13.8
UBP (Wu et al., 2025)	41.2	51.2	51.2	51.1	42.2	57.5	49.0	58.6	45.1	61.5	50.9
ViEEG (Liu et al., 2026a)	34.1	38.4	40.6	50.1	28.9	44.3	38.6	54.0	37.3	42.8	40.9
NeuroBridge (Zhang et al., 2026)	50.0	63.2	61.6	61.4	54.8	69.7	62.7	71.2	64.0	73.6	63.2
BrainHIVE (Zheng et al., 2026)	64.3	76.3	74.0	67.0	68.0	81.5	76.8	84.8	76.8	87.3	75.7
Blur (Liu et al., 2026b)	81.9	81.7	78.3	76.9	71.4	84.3	78.2	84.5	78.7	84.1	80.0
Shallow Align. (Du et al., 2026)	75.0	87.5	83.2	79.5	74.6	89.9	78.5	86.9	81.3	89.3	82.6
HCF (Tang et al., 2026)	81.9	86.0	84.3	82.2	74.5	90.4	82.8	90.9	79.7	93.1	84.6
Ours	**96.8** ±0.5	**96.3** ±0.2	**97.2** ±0.5	**91.5** ±0.7	**92.5** ±1.1	**95.7** ±0.6	**96.2** ±0.6	**97.7** ±1.4	**93.7** ±0.2	**96.8** ±0.2	**95.4** ±0.1
Top-5 accuracy (%)											
NICE	36.6	33.9	39.0	47.0	26.9	40.6	42.1	49.9	37.1	41.9	39.5
UBP	70.5	80.9	82.0	76.9	72.8	83.5	79.9	85.8	76.2	88.2	79.7
ViEEG	71.3	67.9	74.7	80.8	61.5	76.5	75.2	82.5	74.9	79.8	74.5
NeuroBridge	77.6	90.6	91.1	90.0	85.0	92.9	88.8	95.1	91.0	97.1	89.9
BrainHIVE	88.8	95.3	95.0	91.8	91.5	96.3	96.8	98.5	95.8	99.3	94.6
Blur	96.8	96.8	96.0	97.2	93.7	98.7	96.7	98.1	96.8	98.1	96.9
Shallow Align.	94.3	98.9	98.2	96.1	96.4	99.3	97.3	99.4	97.8	99.1	97.7
HCF	98.3	98.9	97.7	98.8	94.1	98.9	97.2	99.4	98.4	99.8	98.2
Ours	**99.7** ±0.2	**99.7** ±0.2	**100.0** ±0.0	**99.5** ±0.4	**99.8** ±0.2	**99.8** ±0.2	**100.0** ±0.0	**100.0** ±0.0	**99.8** ±0.2	**100.0** ±0.0	**99.8** ±0.0

Algorithm 1 RPA training and M/EEG-to-image retrieval

Require: Training and validation pairs $\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{val}}$; frozen visual encoder F and fixed layer ℓ
Require: Epoch count E; candidate images $\{I_j^{\mathrm{cand}}\}_{j=1}^{J}$; query response $b^\star$
Ensure: Trained parameters Φ and ranked candidate-image indices π

function RPAEMBED($X;\theta$)
  $U_1 \leftarrow \mathrm{CNG}_{1\times1}^{d_a\leftarrow D}(X)$
  $U_2 \leftarrow \mathrm{CNG}_{3\times3}^{d_a\leftarrow d_a}(U_1)$
  $U_3 \leftarrow \mathrm{CNG}_{1\times1}^{D\leftarrow d_a}(U_2)$
  return $\nu(q(\mathrm{GAP}(X+U_3)))$
end function
Frozen patch-token extraction
Cache $X(I) \leftarrow \mathrm{PatchGrid}(F_\ell(I))$ for all training and validation images
Joint training
Initialize Φ with $\gamma = \log(1/0.07)$; keep F frozen throughout
for $e = 1, \dots, E$ **do**
  Set trainable modules to training mode
  for each minibatch $\{(b_i, I_i)\}_{i=1}^{B}$ from $\mathcal{D}_{\mathrm{tr}}$ **do**
    $\widetilde{b}_i \leftarrow Mb_i$ for MEG, and $\widetilde{b}_i \leftarrow b_i$ for EEG
    $z_{b,i} \leftarrow \nu(f_\phi(\widetilde{b}_i)), \quad z_{v,i} \leftarrow \text{RPAEMBED}(X(I_i);\theta)$ for each i
    $S_{ij} \leftarrow \exp(\gamma)\, z_{b,i}^\top z_{v,j}$ for $i, j = 1, \dots, B$
    $\mathbf{y} \leftarrow (1, \dots, B)$
    $\mathcal{L} \leftarrow \frac{1}{2}\big[\mathrm{CE}(S, \mathbf{y}) + \mathrm{CE}(S^\top, \mathbf{y})\big]$
    Update Φ with AdamW after clipping the gradient norm to 1.0
  end for
  Evaluate retrieval on $\mathcal{D}_{\mathrm{val}}$ in evaluation mode; save the best checkpoint
end for
Restore the best-validation checkpoint and set all modules to evaluation mode
Image retrieval
for $j = 1, \dots, J$ **do**
  $X_j^{\mathrm{cand}} \leftarrow \mathrm{PatchGrid}(F_\ell(I_j^{\mathrm{cand}}))$
  $z_{v,j}^{\mathrm{cand}} \leftarrow \text{RPAEMBED}(X_j^{\mathrm{cand}};\theta)$
end for
$\widetilde{b}^\star \leftarrow Mb^\star$ for MEG, and $\widetilde{b}^\star \leftarrow b^\star$ for EEG
$z_b^\star \leftarrow \nu(f_\phi(\widetilde{b}^\star))$
$\pi \leftarrow \mathrm{argsort}_j\big[(z_b^\star)^\top z_{v,j}^{\mathrm{cand}}\big]$ in descending order
return Φ, π

Table 6: Subject-level LOSO retrieval on THINGS-EEG2.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Mean
Top-1 accuracy (%)											
NICE (Song et al., 2024)	7.6	5.9	6.0	6.3	4.4	5.6	5.6	6.3	5.7	8.4	6.2
UBP (Wu et al., 2025)	11.5	15.5	9.8	13.0	8.8	11.7	10.2	12.2	15.5	16.0	12.4
ViEEG (Liu et al., 2026a)	22.7	24.7	19.0	25.5	19.8	20.7	20.9	20.8	23.8	31.2	22.9
NeuroBridge (Zhang et al., 2026)	23.2	21.2	13.2	17.0	14.5	25.0	15.3	20.1	13.7	27.2	19.0
BrainHIVE (Zheng et al., 2026)	–	–	–	–	–	–	–	–	–	–	20.0
Blur (Liu et al., 2026b)	27.3	32.0	11.1	18.6	16.9	16.1	16.0	18.3	13.6	30.1	20.0
Shallow Align. (Du et al., 2026)	23.5	30.6	10.0	19.5	18.1	22.7	18.6	17.3	23.0	34.4	21.8
HCF (Tang et al., 2026)	28.5	22.5	20.0	20.5	18.5	**29.5**	20.0	20.0	21.0	34.0	23.4
Ours	**49.3** ±1.3	**42.2** ±2.5	**24.2** ±2.4	**31.5** ±0.7	**34.2** ±1.2	27.0 ±1.5	**35.2** ±0.9	**29.8** ±1.2	**35.0** ±0.4	**46.3** ±1.5	**35.5** ±0.3
Top-5 accuracy (%)											
NICE	22.8	20.5	22.3	20.7	18.3	22.2	19.7	22.0	17.6	28.3	21.4
UBP	29.7	40.0	27.0	32.3	33.8	31.0	23.8	32.2	40.5	43.5	33.4
ViEEG	53.5	52.5	48.4	54.1	47.3	49.3	49.4	46.8	52.7	60.3	51.4
NeuroBridge	52.4	49.3	36.5	45.3	37.7	55.0	45.1	44.9	36.5	56.3	45.9
BrainHIVE	–	–	–	–	–	–	–	–	–	–	44.1
Blur	54.9	63.0	34.5	46.0	40.7	43.2	45.4	49.0	41.2	62.1	48.0
Shallow Align.	53.2	60.4	28.1	48.3	45.2	49.8	46.0	46.1	54.8	62.0	49.4
HCF	61.0	56.5	45.5	54.0	46.5	**60.5**	53.5	53.5	48.5	69.0	54.9
Ours	**79.5** ±0.8	**73.0** ±0.8	**50.7** ±1.3	**63.0** ±1.9	**69.7** ±1.3	59.2 ±1.3	**63.7** ±0.5	**62.5** ±0.7	**72.2** ±1.5	**80.8** ±1.7	**67.4** ±0.7

B.1 MEG

Table 7: Subject-level within-subject retrieval on THINGS-MEG.

Method	S1	S2	S3	S4	Mean
Top-1 accuracy (%)					
NICE (Song et al., 2024)	9.6	18.5	14.2	9.0	12.8
UBP (Wu et al., 2025)	15.0	46.0	27.3	18.5	26.7
ViEEG (Liu et al., 2026a)	16.6	37.4	30.6	17.2	25.5
NeuroBridge (Zhang et al., 2026)	16.5	53.7	40.4	18.1	32.2
BrainHIVE (Zheng et al., 2026)	14.0	63.8	41.0	17.0	33.9
Blur (Liu et al., 2026b)	25.4	78.1	37.3	35.3	44.0
Shallow Align. (Du et al., 2026)	25.5	81.9	56.0	28.6	48.0
Ours	$\mathbf{34.7 \pm 1.7}$	$\mathbf{96.2 \pm 0.6}$	$\mathbf{82.2 \pm 0.8}$	$\mathbf{47.7 \pm 2.2}$	$\mathbf{65.2 \pm 1.0}$
Top-5 accuracy (%)					
NICE	27.8	47.8	41.6	26.6	36.0
UBP	38.0	80.5	59.0	43.5	55.2
ViEEG	50.3	79.1	73.9	49.5	63.2
NeuroBridge	41.6	85.3	73.2	43.1	60.8
BrainHIVE	31.8	91.8	78.3	41.0	60.9
Blur	–	–	–	–	72.0
Shallow Align.	54.5	97.4	87.5	58.3	74.4
Ours	$\mathbf{67.7 \pm 1.9}$	$\mathbf{99.8 \pm 0.2}$	$\mathbf{97.2 \pm 1.0}$	$\mathbf{78.0 \pm 0.4}$	$\mathbf{85.7 \pm 0.8}$

Table 8: Subject-level LOSO retrieval on THINGS-MEG.

Method	S1	S2	S3	S4	Mean
Top-1 accuracy (%)					
UBP (Wu et al., 2025)	2.0	1.5	2.7	2.5	2.2
NeuroBridge (Zhang et al., 2026)	**4.3**	3.6	3.0	2.5	3.4
BrainHIVE (Zheng et al., 2026)	–	–	–	–	5.4
Blur (Liu et al., 2026b)	2.9	7.7	5.8	**4.7**	5.3
Shallow Align. (Du et al., 2026)	1.3	6.6	5.4	1.5	3.7
Ours	3.0 ± 1.1	$\mathbf{11.0 \pm 1.4}$	$\mathbf{9.0 \pm 0.8}$	3.7 ± 0.2	$\mathbf{6.7 \pm 0.4}$
Top-5 accuracy (%)					
UBP	5.7	17.2	10.5	8.0	10.4
NeuroBridge	**13.1**	15.6	11.2	11.3	12.8
BrainHIVE	–	–	–	–	15.2
Blur	–	–	–	–	15.9
Shallow Align.	6.9	18.5	18.5	7.9	13.0
Ours	10.5 ± 1.1	$\mathbf{28.2 \pm 1.9}$	$\mathbf{26.0 \pm 3.1}$	$\mathbf{12.3 \pm 1.2}$	$\mathbf{19.2 \pm 1.0}$

C Visual Adapters at a Fixed Visual-Encoder Output

Table 3 reports the matched adapter comparison. This appendix gives parameter counts, alternative adapter definitions and results, and the bottleneck-width and kernel-placement experiments.

C.1 Visual adapter families

Shared adapter input and output. Every adapter in Table 9 receives the same frozen OpenCLIP ViT-H/14 layer-11 output. Writing

$$s = [c; t], \qquad t \in \mathbb{R}^{B\times 256\times 1280}, \qquad g = \mathrm{grid}(t) \in \mathbb{R}^{B\times 1280\times 16\times 16}, \tag{6}$$

c is the CLS token, t is the batched patch-token sequence denoted by X in Section 3, and g is its grid form. The normalized visual embedding is used in the contrastive loss. Let $q : \mathbb{R}^{1280} \to \mathbb{R}^{1024}$ denote the final learned linear projection (q in the main text), and $\mathrm{GAP}(g) = \frac{1}{256}\sum_{i=1}^{16}\sum_{j=1}^{16} g_{:,i,j}$.

For any patch-token grid x, $\mathrm{tok}(x)$ denotes the same values reshaped into a sequence of 256 patch tokens; it performs no pooling. For the attention-pooling adapters, let t_p denote the original patch token at position p. A learned query $u \in \mathbb{R}^{1280}$ and key map K define

$$\alpha_p(t) = \frac{\exp\big((Kt_p)^\top u/\sqrt{1280}\big)}{\sum_{j=1}^{256}\exp\big((Kt_j)^\top u/\sqrt{1280}\big)}, \qquad \mathrm{AttnPool}_V(t) = \sum_{p=1}^{256}\alpha_p(t)V(t_p), \tag{7}$$

where u is initialized as $\mathcal{N}(0, 0.02^2)$, K is linear, and $\mathrm{AttnPool}$ denotes learned-query attention pooling. The value map V distinguishes the three patch-token query-pool adapters below. MHA denotes multihead attention. $\mathcal{B}_{131}$ denotes RPA's 1×1, 3×3, 1×1 residual branch and $\mathcal{B}_{111}$ replaces its middle kernel by 1×1; every convolution is followed by GroupNorm and GELU. Optional output normalization, external layer scale, and drop path are disabled; ConvNeXt's internal layer scale is specified in its row

Table 9: Comparison of visual adapters (200-way Top-1, %).

Family	Adapter	Within	LOSO
RPA	local $(1, 3, 1)$	**95.4**	**35.5**
RPA	pointwise $(1, 1, 1)$	93.0	34.9
RPA ablation	no final GELU	90.9	32.4
RPA ablation	pointwise + query pool	90.7	34.0
scalar attention	residual-MLP values	89.7	32.5
scalar attention	MLP values	89.2	31.5
scalar attention	linear values	89.0	29.5
attention	8-head MHA query pool	88.1	29.6
transformer	block + MHA query pool	83.0	32.3
modern convolution	full-width ConvNeXt	88.7	31.8
modern convolution	depthwise-separable residual conv	77.8	27.1
fixed	patch mean + linear (no LN)	79.2	26.6

Table 10: Trainable visual-adapter parameters, each maps visual embedding of 1,024 dimensions. Counts are for the fixed image encoder used in Table 9.

Adapter	Parameters	Adapter	Parameters
CLS-token / patch-mean controls	1.314M	patch mean + linear (no LN)	1.312M
pointwise RPA $(1, 1, 1)$	2.038M	local RPA $(1, 3, 1)$	2.562M
local RPA, no final GELU	2.562M	depthwise-separable conv	2.969M
query pool + MLP values	3.610M	pointwise RPA + query pool	3.679M
query pool + linear values	4.592M	8-head MHA query pool	7.872M
full-width ConvNeXt	14.493M	transformer + MHA pool	20.993M

Table 11: Definitions of the visual adapters evaluated in Table 9.

.

Adapter	Adapter output h_v before L2 normalization
local RPA $(1, 3, 1)$	$Q(\mathrm{GAP}[g + \mathcal{B}_{131}(g)])$
pointwise RPA $(1, 1, 1)$	$Q(\mathrm{GAP}[g + \mathcal{B}_{111}(g)])$
local RPA, no final GELU	local RPA with the last GELU in $\mathcal{B}_{131}$ removed; the residual $+g$ is retained
pointwise RPA + query pool	$\tilde{t} = \mathrm{tok}[g + \mathcal{B}_{111}(g)]$; $Q\big(\mathrm{AttnPool}_{\mathrm{Id}}(\tilde{t})\big)$
query pool + residual-MLP values	$Q(\mathrm{AttnPool}_{V_{\mathrm{res}}}(t))$, $V_{\mathrm{res}}(x) = x + W_2\,\mathrm{GELU}(W_1 x)$, $W_1 : 1280 \to 256$, $W_2 : 256 \to 1280$
query pool + MLP values	$Q\Big(\mathrm{AttnPool}_{V_{\mathrm{mlp}}}(t)\Big)$, $V_{\mathrm{mlp}}(x) = W_2\,\mathrm{GELU}(W_1 x)$ with the same widths
query pool + linear values	$Q(\mathrm{AttnPool}_V(t))$ with linear $V : 1280 \to 1280$
8-head MHA query pool	$Q(\mathrm{MHA}_8(u, t, t))$ using one learned query
transformer block + MHA query pool	one post-norm 8-head transformer layer over t (FFN width 2,560, GELU, dropout 0.1), followed by the learned-query MHA above
full-width ConvNeXt	$Q(\mathrm{GAP}[g + \mathrm{CN}(g)])$; CN uses depthwise 7×7, LayerNorm, $1280 \to 5120 \to 1280$ channel mixing with GELU, and layer scale initialized to 10^{-6}
depthwise-separable residual conv	$Q(\mathrm{GAP}[g + \mathrm{DW}(g)])$; DW is depthwise 3×3 then pointwise 1×1, each followed by GroupNorm(32) and GELU
patch mean + linear (no LN)	$Q(\mathrm{GAP}(g))$

Local RPA has the highest reported mean accuracy in this comparison, despite having fewer trainable parameters than the learned non-RPA adapters. Larger adapters therefore do not yield higher accuracy in these runs.

C.2 Bottleneck width and kernel placement

Table 12: RPA bottleneck-width and kernel-placement study.

Adapter	Structure	Params	Within subject	LOSO
width 64	$r = 64$	1.52M	93.4	34.9
no local mixing	$k = (1, 1, 1)$	2.04M	93.0	34.8
reference	$r = 256,\ k = (1, 3, 1)$	2.56M	95.4	35.5
all-local	$k = (3, 3, 3)$	7.81M	91.9	33.8
width 1280	$r = 1280$	19.35M	94.5	35.3

Increasing the adapter parameter count by $7.6\times$ through bottleneck widening does not consistently improve performance. Using a 3×3 kernel in every layer gives lower accuracy despite approximately tripling the parameter count. The reference bottleneck with one local layer therefore provides a favorable parameter–accuracy trade-off in this experiment.

D Visual-Encoder Layer and Backbone Study

We examine how visual-encoder depth affects retrieval and whether layer choice changes the relative ordering of encoders. Intermediate-layer optima have also been reported in recent visual-decoding work (Du et al., 2026; Tang et al., 2026; Jiang et al., 2026).

Baseline layer-selection. We selected OpenCLIP ViT-H/14 layer 11 because it achieved the highest EEG Top-1 test accuracy after a fully depth sweep among ViT-H/14

D.1 Full visual-encoder layer sweep

Table 13: Full visual-encoder layer sweep (EEGProjection encoder, uniform six-point grid; Top-1 accuracy in %). Rank 1 is the shallowest sampled state and rank 6 is the final block.

Encoder	rank 1	rank 2	rank 3	rank 4	rank 5	rank 6 (final)
EEG within-subject						
CLIP-H	93.0 ± 0.2	94.0 ± 0.1	92.2 ± 0.2	69.3 ± 0.9	61.4 ± 0.1	52.7 ± 0.8
I-JEPA	85.7 ± 0.4	89.9 ± 0.5	91.9 ± 0.2	90.6 ± 0.6	89.9 ± 0.1	73.4 ± 0.6
DINOv2-L	89.1 ± 0.4	92.9 ± 0.2	94.4 ± 0.5	90.8 ± 0.6	77.8 ± 0.5	55.8 ± 0.6
DINOv2-S	88.7 ± 0.3	91.0 ± 0.1	89.5 ± 0.4	84.3 ± 0.4	77.0 ± 0.3	62.2 ± 1.3
ViT-B/32	91.7 ± 0.4	93.8 ± 0.2	91.3 ± 0.2	77.1 ± 0.7	56.2 ± 1.2	51.7 ± 0.8
RN50	72.2 ± 0.1	78.4 ± 0.8	84.8 ± 1.2	85.6 ± 0.6	84.7 ± 0.5	73.8 ± 0.3
SDXL-VAE	60.9 ± 0.6	42.3 ± 0.1	64.9 ± 0.9	63.9 ± 1.0	64.1 ± 1.0	5.4 ± 0.3
EEG LOSO						
CLIP-H	26.0 ± 0.1	30.7 ± 0.5	32.9 ± 0.4	25.6 ± 0.3	22.7 ± 0.8	19.3 ± 0.9
I-JEPA	24.8 ± 0.4	28.6 ± 0.5	31.6 ± 0.5	31.0 ± 0.4	30.5 ± 0.3	25.7 ± 0.6
DINOv2-L	25.5 ± 0.1	28.7 ± 0.1	31.9 ± 0.4	31.3 ± 0.5	26.5 ± 0.3	19.5 ± 0.6
DINOv2-S	26.1 ± 0.4	29.3 ± 0.8	28.9 ± 0.4	28.1 ± 0.8	24.9 ± 0.4	20.7 ± 0.3
ViT-B/32	26.9 ± 0.8	31.6 ± 0.3	29.8 ± 0.6	24.8 ± 0.3	19.8 ± 0.1	18.4 ± 0.2
RN50	21.1 ± 0.6	22.8 ± 0.4	26.2 ± 0.6	27.5 ± 0.2	28.2 ± 1.2	25.0 ± 0.3
SDXL-VAE	16.5 ± 0.2	14.7 ± 0.5	18.7 ± 0.3	18.0 ± 0.1	18.0 ± 0.4	3.9 ± 0.3
MEG within-subject						
CLIP-H	55.1 ± 1.2	60.5 ± 1.0	55.4 ± 1.1	38.6 ± 1.3	33.5 ± 1.5	29.2 ± 0.4
I-JEPA	45.4 ± 0.9	53.5 ± 2.0	57.6 ± 1.6	57.6 ± 1.4	52.2 ± 2.4	36.5 ± 1.4
DINOv2-L	46.1 ± 0.8	53.4 ± 0.6	60.0 ± 0.4	58.5 ± 0.5	44.4 ± 1.5	27.4 ± 1.1
DINOv2-S	46.0 ± 0.6	53.6 ± 1.4	56.5 ± 0.6	50.3 ± 1.2	39.9 ± 0.8	33.4 ± 0.7
ViT-B/32	54.0 ± 1.5	59.5 ± 1.5	55.7 ± 0.3	43.9 ± 0.7	29.6 ± 0.7	26.5 ± 2.2
RN50	31.2 ± 2.3	40.2 ± 0.6	48.3 ± 2.2	48.1 ± 1.2	47.9 ± 0.2	40.0 ± 1.3
SDXL-VAE	23.2 ± 0.2	18.9 ± 0.2	30.2 ± 1.0	31.4 ± 0.5	30.2 ± 1.0	5.0 ± 0.4
MEG LOSO						
CLIP-H	5.3 ± 0.4	5.6 ± 0.3	5.1 ± 0.1	3.6 ± 0.6	3.9 ± 0.3	3.0 ± 0.2
I-JEPA	5.2 ± 0.3	5.4 ± 0.4	5.4 ± 0.4	6.3 ± 0.4	6.0 ± 0.8	4.3 ± 0.2
DINOv2-L	5.3 ± 0.3	5.4 ± 0.2	6.3 ± 0.3	4.9 ± 0.7	4.2 ± 0.3	2.9 ± 0.5
DINOv2-S	5.5 ± 0.4	6.2 ± 0.7	5.9 ± 0.5	4.8 ± 0.7	4.7 ± 0.5	4.7 ± 0.2
ViT-B/32	6.1 ± 0.7	6.5 ± 0.4	5.8 ± 0.4	4.6 ± 0.2	3.1 ± 0.4	3.7 ± 0.1
RN50	3.5 ± 0.4	4.4 ± 0.4	5.1 ± 0.1	5.0 ± 0.3	4.7 ± 0.2	4.6 ± 0.4
SDXL-VAE	3.4 ± 0.1	3.4 ± 0.1	3.9 ± 0.0	3.8 ± 0.4	3.6 ± 0.0	1.2 ± 0.2

Across EEG and MEG, within-subject and LOSO evaluation, every sampled ViT reaches its maximum before the final block. Peak rank and curve shape still vary across encoders, while RN50 and the SDXL-VAE exhibit less regular layer profiles. Table 13 provides the absolute-score counterpart to the standardized trends in Figure 2.

D.2 How layer choice confounds model comparison

Table 14: LOSO EEG comparison between each encoder's final sampled state and the maximum over its six test-set scores.

Encoder	Pretraining	Final layer	Test-best sampled layer (rank)
CLIP ViT-H/14	language contrastive	19.3 ± 0.9	32.9 ± 0.4 (r3)
I-JEPA ViT-H/14	masked SSL	25.7 ± 0.6	31.6 ± 0.5 (r3)
DINOv2-L/14	SSL distillation	19.5 ± 0.6	31.9 ± 0.4 (r3)
DINOv2-S/14	SSL distillation	20.7 ± 0.3	29.3 ± 0.8 (r2)
CLIP ViT-B/32	language contrastive	18.4 ± 0.2	31.6 ± 0.3 (r2)
CLIP RN50	language contrastive	25.0 ± 0.3	28.2 ± 1.2 (r5)
SDXL VAE	pixel reconstruction	3.9 ± 0.3	18.7 ± 0.3 (r3)

At the final sampled state, I-JEPA exceeds CLIP-H by 6.4 percentage points. At each encoder's test-set maximum, CLIP-H is 1.3 points higher. The relative ordering thus changes with the layer comparison. The $35.5 \pm 0.3\%$ LOSO result for the ViT H/14 layer-11 benchmark comes from a full layer sweep over ViT H/14.

E Patch-Token Geometry and Auxiliary Probes

Section 4.4 reports patch-token input interventions. Here, linear EEG probes compare frozen CLS features with pooled patch features and their residuals. Separate image-feature probes measure descriptor predictability, while visual-input modification evaluate retrieval after modifying the images supplied to the frozen visual encoder.

E.1 Ridge probes of frozen CLS tokens and patch tokens

Patch-token averaging probe. To test whether averaging more patch tokens produces a vector that is easier to predict from EEG, we place the 256 frozen layer-11 patch tokens in one random order shared across all images and participants. For each $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$, we average the first k patch tokens into one 1,280-dimensional *mean-pooled patch feature*. For each k and participant, we fit a ridge regression on training images to predict this feature from EEG. Retrieval compares the predicted vector with the mean-pooled patch features of the 200 test images using cosine similarity. The CLS token is evaluated with the identical procedure. Figure 5**C** shows the complete sweep: Top-1 accuracy rises from 6.20% for one patch token to 34.45% for all 256, compared with 18.95% for the CLS token. The sweep uses one random token ordering shared across images and participants. Its 1,280-dimensional predicted vectors are outputs of the ridge probe, distinct from the 1,024-dimensional M/EEG embeddings z_b in the contrastive system. The probe uses fixed pooling without RPA.

Relationship between the CLS token and patch tokens. Table 15 uses only the frozen CLIP ViT-H/14 layer-11 output for the 200 test images, without EEG data or RPA. It reports centered CLS–patch-mean cosine similarity, projection onto each image's top 128 patch-token directions, and least-squares reconstruction of the CLS token from all 256 patch tokens.

Table 15: Three comparisons between the frozen CLS token and patch tokens on the 200 test images. Before computing cosine similarity, we subtract the across-image mean separately from the CLS tokens and patch means. The projection rows report the fraction of a vector's squared norm captured by the top 128 directions fitted to the same image's patch tokens. These are deterministic visual-feature summaries; no EEG or RPA is involved.

Analysis	Vector or comparison	Result
Centered cosine similarity	CLS token versus the mean of its patch tokens	-0.05
Centered cosine reference	Means of two disjoint halves of the same patch-token set	0.89
Top-128 projection	Patch token from the same image (included when fitting the directions)	0.92
Top-128 projection	Patch token from an unrelated image	0.36
Top-128 projection	CLS token from the same image	0.17
Top-128 projection reference	Random 1,280-D vector ($128/1280$ expected)	0.10
Linear reconstruction	CLS token reconstructed from its 256 patch tokens	$R^2 = 0.21$

After centering, the CLS token has little cosine similarity with its image's patch mean (-0.05), whereas the means of two disjoint halves of the same patch-token set remain similar (0.89). The image's top 128 patch-token directions capture 0.17 of the CLS token's squared norm, compared with 0.36 for a patch token from another image and 0.10 for a random vector. A weighted sum of all 256 patch tokens reconstructs the CLS token with $R^2 = 0.21$. These visual-only comparisons distinguish the CLS token from the patch mean. The following EEG probe examines which components of each are predictable from brain responses.

Reciprocal residualization. We fit a linear map from CLS tokens to patch means on training images, subtract its predictions from the patch means, and use the resulting residuals as targets for the EEG ridge probe. We then reverse the direction: fit a map from patch means to CLS tokens, subtract the predicted CLS features, and probe the CLS residuals. Each residualization map is applied unchanged to the test images.

Table 16: Reciprocal residualization followed by the same linear EEG probe. The residualization map is fit on training images and applied unchanged to test images. Relative accuracy is Top-1 accuracy as a percentage of that for the corresponding unmodified target, without chance correction.

Vector	Top-1 (%)	Variance kept	Relative accuracy
patch mean	**34.45**	100%	—
CLS token	18.95	100%	—
patch mean, CLS-token prediction removed	**10.80**	24.2%	31.3%
CLS token, patch-mean prediction removed	2.00	35.2%	10.6%

With correctly paired residualization, the CLS residual yields lower retrieval accuracy than the patch-mean residual, despite retaining a larger fraction of variance. This experiment indicates that decodable CLS information overlaps substantially with the patch mean, while the patch-mean residual retains additional EEG-predictable information.

E.2 Auxiliary image-feature probes

Feature construction. Each probe uses a fixed image-feature vector. In addition to color, texture, and 2D layout features, we include *object-category semantics*, represented by a CLIP-based category feature; *HOG oriented edges*, represented by a histogram of local image-gradient orientations; and *SAM 3 object silhouette*, represented by an object mask produced by Meta's Segment Anything Model 3 (SAM 3) (Carion et al., 2026).

EEG decoding. For each feature family and participant, we fit a separate ridge regression from EEG to that image-feature vector. On the 200-image test set, five-fold cross-validation holds out 40

images per fold. Each EEG-predicted vector is compared by cosine similarity with the 40 held-out image vectors, and Top-1 is correct when the matching image ranks first. The training-set diagnostic instead asks whether an EEG-predicted vector is more similar to its matched image than to an unmatched image, using up to 8,000 training images. The probes evaluate prediction of their respective feature vectors independently of the RPA retrieval model.

Table 17: Exploratory linear decoding of six named image-feature targets from EEG (10 participants). Test Top-1 is five-fold, 40-way cosine retrieval on the 200-image test set (chance 2.5%). Train pairwise is matched-versus-mismatched accuracy on up to 8,000 training images (chance 50%).

Feature family	Test Top-1 (%)	Train pairwise (%)
object category / semantics	17.0	82
color	16.5	79
texture	14.5	78
objectness / 2-D layout	7.0	58
HOG oriented edges	6.0	67
SAM 3 object silhouette	6.5	55

Category/semantics, color, and texture produce the highest held-out retrieval among the named features (14.5–17.0%, versus 2.5% chance). Objectness/layout, HOG, and the SAM 3 silhouette are much lower (6.0–7.0%). These differences compare the specified descriptor vectors under a linear EEG probe. HOG and SAM 3 represent different properties—oriented-edge statistics and object silhouettes, respectively.

E.3 VISUALIZATIONS

This part is the supplementary technical detail for Section 4.3 and Section 4.4

UMAP visualization over EEG embedding . We extract normalized 1,024-dimensional EEG and visual embeddings from the corresponding within-subject model. Because the each subject's trained models use different coordinate bases, we align their embedding spaces using orthogonal Procrustes alignment with Subject 1 as the reference. For each other participant, we fit an orthogonal transformation between their visual embeddings and Subject 1's visual embeddings using the correspondence between images. We then apply this transformation to that subject's EEG embeddings. The transformation preserves norms and cosine similarities within each participant. After alignment, we combine the EEG embeddings belonging to four categories and jointly project them into two dimensions using UMAP (McInnes et al., 2018)

Grad-CAM computation and image selection. We compute Grad-CAM (Selvaraju et al., 2017) at the middle 3×3 convolution of EEG-aligned RPA and image-only RPA. Let $A \in \mathbb{R}^{C\times H\times W}$ denote the activations at this layer and z_v the normalized visual embedding produced by the corresponding RPA. We compute the cosine similarity $s = \langle \tau, z_v \rangle$, where τ is the normalized paired EEG embedding for EEG-aligned RPA and the normalized final CLIP embedding of the same image for image-only RPA. The reference embedding τ is held fixed during differentiation. We normalize the Grad-CAM maps and compare the EEG-aligned and image-only maps for each image.

Layer-wise visual decoding across EEG time windows. For each of the seven visual encoders, we select the two layers with the highest within-subject EEG Top-1 test accuracy in the exploratory depth sweep, yielding 14 selected layers. At each selected ViT layer, we average the patch tokens; for RN50 and SDXL-VAE, we globally average the spatial feature map. For each layer, we reduce the resulting visual features to 40 principal components and z-score them. We extract overlapping 100-ms windows at 60-ms intervals, flatten each window, reduce the EEG features to 40 principal components, and z-score them. For each participant and layer–window pair, we fit a ridge regression model ($\lambda = 10^3$) to predict the PCA-reduced visual features from EEG using four-fold cross-validation over the test images. This probe predicts frozen visual features directly from EEG without RPA or the contrastively trained brain encoder.

F BRAIN ENCODERS, TEMPORAL WINDOWS, AND REPETITION AVERAGING

F.1 CONVENTIONAL, AND PRETRAINED EEG FOUNDATION MODEL ENCODERS

We compare our baseline with EEG foundation model: EEG-Conformer (Song et al., 2023) and four pretrained EEG foundation model encoders: CBraMod (Wang et al., 2025), CSBrain (Zhou et al., 2025), LaBraM (Jiang et al., 2024), and BIOT (Yang et al., 2023). EEG-Conformer and CBraMod obtain accuracies close to those of the projection encoder in both protocols, while other pretrained encoders vary more widely.

Table 18: Within-subject and LOSO 200-way Top-1 accuracy (%) EEG foundation models.

Brain encoder	Pretraining	Within Top-1	LOSO Top-1
EEG-Conformer	none	94.90 ± 0.28	34.45 ± 0.26
CBraMod	EEG corpus	94.55 ± 0.40	35.63 ± 0.06
CSBrain	EEG corpus	93.83 ± 0.36	28.73 ± 0.61
LaBraM	2,500 h EEG	79.33 ± 1.10	32.22 ± 0.61
BIOT	EEG corpus	32.57 ± 0.39	11.47 ± 1.29
EEGProjection (our baseline)	none	95.4 ± 0.12	35.5 ± 0.34

A separate experiment compares the baseline encoder with conventional EEG-to-image retrieval task encoders: TSConv (TSception) (Ding et al., 2023), ShallowConvNet and DeepConvNet (Schirrmeister et al., 2017), and EEGNet (Lawhern et al., 2018). The EEGProjection encoder has the highest mean accuracy in both protocols in this experiment.

Table 19: Within-subject and LOSO Top-1 accuracy (%) for various EEG encoders

Brain encoder	Within Top-1	LOSO Top-1
EEGProjection	$\mathbf{95.4 \pm 0.3}$	$\mathbf{35.5 \pm 0.3}$
TSConv (TSception)	87.9 ± 0.6	30.4 ± 0.5
ShallowConvNet	79.0 ± 0.8	23.9 ± 0.6
EEGNet	77.2 ± 0.6	20.7 ± 0.8
DeepConvNet	64.7 ± 0.1	15.3 ± 0.4

F.2 EFFECT OF EEG TIME WINDOWS.

We next examine which portions of the EEG response support image retrieval. We train the model using equal-duration windows from the beginning $[0, d]$ or the end $[1000 - d, 1000]$ of the EEG signal. As shown in Figure 6, windows starting at stimulus onset yield higher Top-1 accuracy than equally long windows from the end of the signal in both tasks. Both early- and late-window sweeps show that retrieval benefits from retaining early EEG activity, while the substantial gap between LOSO and within-subject accuracy highlights room for improving cross-subject generalization in the future works.

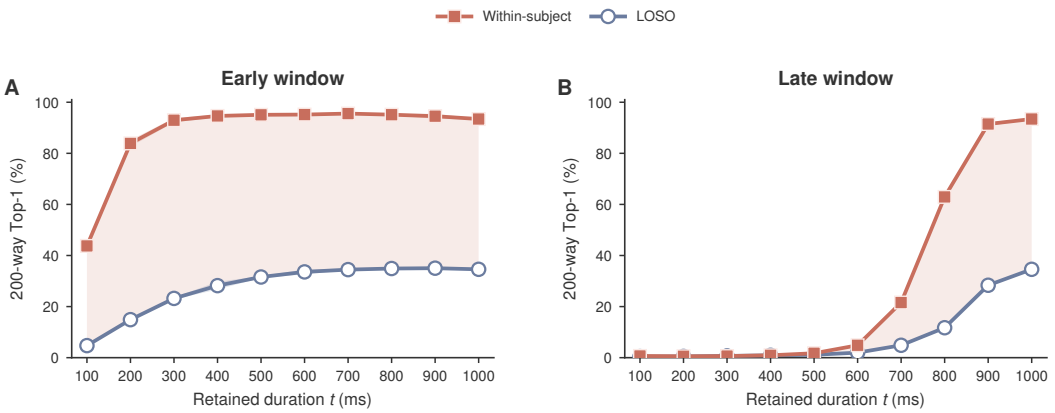

Figure 6: Top-1 accuracy for matched-duration EEG windows retained from the start, $[0, d]$, or end, $[1000 - d, 1000]$.

G Error-Focused Qualitative Retrievals

The within-subject model used for this montage achieves 96% Top-1 accuracy for Subject 1 over the complete 200-image test set. Figure 7 shows seven selected Top-1 errors and one Top-1 success. Among the seven errors, the correct image ranks second in six rows and sixth in one. For each trial, the model ranks all 200 candidate images by cosine similarity between the embedding of the 80-repetition-averaged EEG response and each candidate's visual embedding.

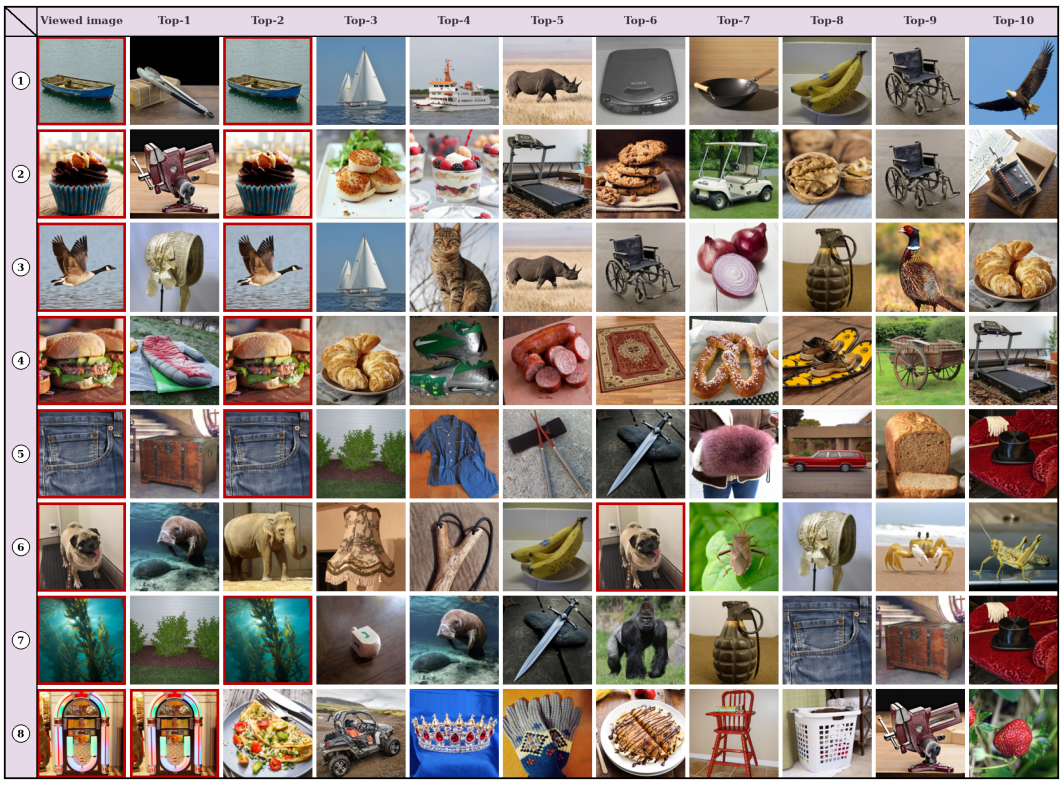

Figure 7: Error-focused Top-10 retrievals for Subject 1 on THINGS-EEG2: seven rows are selected Top-1 errors and one is a Top-1 success, while the complete test-set Top-1 accuracy is 96%; red outlines identify the viewed image and its position among the ranked candidates.

G.1 CROSS-MODAL SIMILARITY MATRIX

Figure 8 visualizes 200-way retrieval for the 10 within-subject models from the same training run. Cell (i, j) is the cosine similarity between the EEG embedding for test image i and the visual embedding for test image j; diagonal cells are the correct pairs. We average the participant-level similarity matrices because all participants share the same test-image ordering. Accuracy is computed for each participant before averaging, not from the averaged matrix.

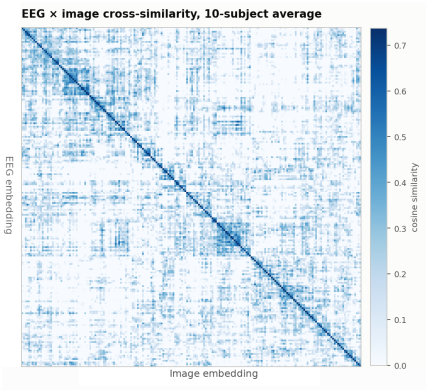

Figure 8: EEG–image cosine-similarity matrix averaged over 10 within-subject models. Rows are EEG embeddings, columns are visual embeddings, and the diagonal contains the correct pairs.

H MEG SENSOR PROJECTION

THINGS-MEG records $C = 271$ sensors. Flattening all sensors and time points creates a large input layer in the brain encoder. We instead learn a bias-free matrix $M \in \mathbb{R}^{K \times C}$ and compute

$$Y = MX, \tag{8}$$

where $X \in \mathbb{R}^{C \times T}$ is one MEG trial and $Y \in \mathbb{R}^{K \times T}$ is the resulting MEG signal. The same M maps the C sensor channels at every time point to K learned channels. We flatten Y and pass it to the otherwise unchanged brain encoder.

Table 20: Effect of the MEG sensor-projection output dimension. "No projection" uses all 271 sensors directly.

	no projection	projection output dimension K						
	271	32	48	64	96	112	128	160
within Top-1	51.3 ± 0.5	58.7 ± 1.1	59.2 ± 0.7	$\mathbf{65.2 \pm 1.0}$	64.0 ± 3.0	63.4 ± 2.1	64.2 ± 1.8	63.2 ± 1.5
LOSO Top-1	5.1 ± 0.7	5.8 ± 0.2	6.2 ± 0.5	$\mathbf{6.7 \pm 0.4}$	5.9 ± 0.4	6.4 ± 0.2	6.1 ± 0.2	6.0 ± 0.4

I EEG-TO-IMAGE RECONSTRUCTION

We evaluate EEG-to-image reconstruction with a separately trained EEG encoder, comparing RPA-only alignment with RPA + pooled-CLIP alignment.

I.1 IMAGE RECONSTRUCTION PIPELINE

The dual-stream reconstruction pipeline is adapted from ATM (Li et al., 2024; Podell et al., 2024; Sauer et al., 2024; Ye et al., 2023). It uses the projection brain-encoder architecture with a separate training configuration. Let $r_b^g \in \mathbb{R}^{1024}$ denote the EEG representation supplied to the priors, and let z_b^g denote the normalized EEG embedding used for contrastive training. The RPA used in this pipeline produces the normalized visual embedding z_{RPA}; the normalized pooled CLIP-H embedding is z_{pool}. The dual alignment objective is

$$\mathcal{L}_{\mathrm{trunk}} = \mathcal{L}_{\mathrm{NCE}}(z_b^g, z_{\mathrm{RPA}}) + 0.3\,\mathcal{L}_{\mathrm{NCE}}(z_b^g, z_{\mathrm{pool}}). \tag{9}$$

Two conditional priors map the EEG representation to the conditions used by the image generator. A semantic diffusion prior P_{sem} predicts a pooled CLIP-H embedding, followed by a per-dimension affine correction $\mathcal{A}$ that matches the per-coordinate mean and variance of pooled CLIP-H embeddings estimated from 4,000 training trials. In parallel, a structural diffusion prior P_{str} predicts an SDXL-VAE latent:

$$\widehat{c} = \mathcal{A}(P_{\text{sem}}(r_b^g)), \qquad \widehat{y}_{\text{VAE}} = P_{\text{str}}(r_b^g) \in \mathbb{R}^{4\times 64\times 64}. \tag{10}$$

The structural latent is decoded and blurred to form a coarse initialization, while $\widehat{c}$ supplies the IP-Adapter condition:

$$I_{\text{init}} = \text{Blur}\left(D_{\text{VAE}}(\widehat{y}_{\text{VAE}})\right), \qquad \widehat{I} = G_{\text{Turbo}}\left(I_{\text{init}}; \text{IP}(\widehat{c})\right). \tag{11}$$

The semantic and structural priors each use 50 sampling steps, with classifier-free-guidance scales 5.0 and 3.0, respectively; the img2img stage uses strength 0.5 and 10 denoising steps. We first train the EEG encoder with the alignment objective, then fit the semantic and structural priors to their corresponding image targets. At test time, the priors predict both conditions from an EEG response for image synthesis. RPA provides alignment supervision; the predicted CLIP-H and VAE features condition the generator.

I.2 Generation-objective comparison

To examine whether pooled-CLIP supervision complements RPA alignment for reconstruction, we compare RPA-only alignment with dual alignment using Equation 9. As shown in Table 21, dual alignment yields higher PixCorr, SSIM, AlexNet-2, AlexNet-5, Inception, and CLIP scores, whereas RPA-only alignment achieves a lower SwAV score. The differences are small but favor adding pooled-CLIP supervision on six of the seven reconstruction metrics.

Table 21: Reconstruction results.

Configuration	PixCorr ↑	SSIM ↑	AlexNet-2 ↑	AlexNet-5 ↑	Inception ↑	CLIP ↑	SwAV ↓
RPA-only	0.2245	0.4520	0.8599	0.9044	0.7702	0.8093	0.5212
RPA + pooled-CLIP	0.2251	0.4541	0.8645	0.9086	0.7832	0.8107	0.5256

I.3 Qualitative reconstructions across participants

Figures 9–13 show reconstructions of 40 selected held-out stimuli across five categories. Each grid shows the viewed images in the top row and one reconstruction per participant below.

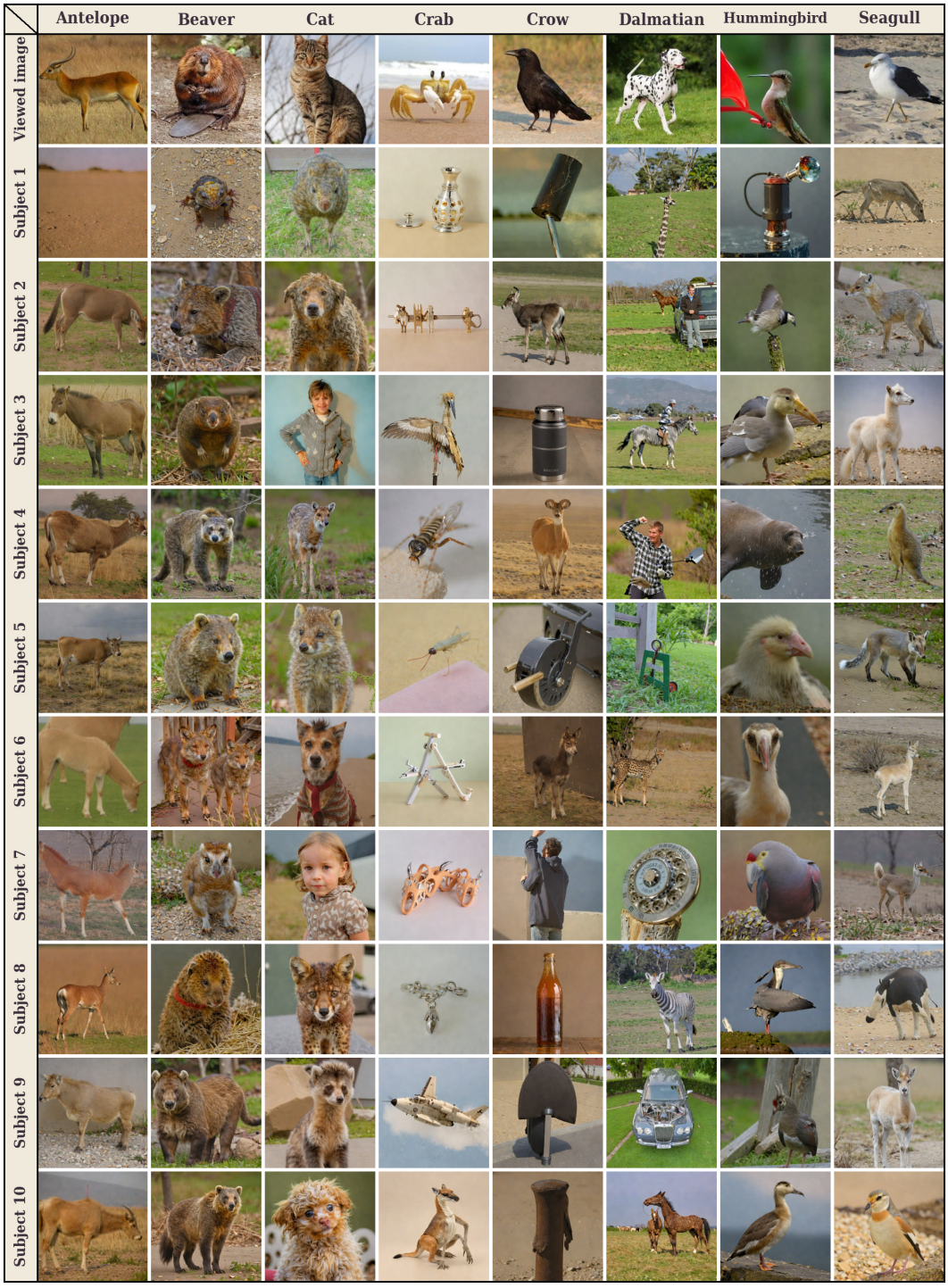

Figure 9: Viewed images (top) and participant-level reconstructions from 80-repetition-averaged EEG responses for eight animal stimuli.

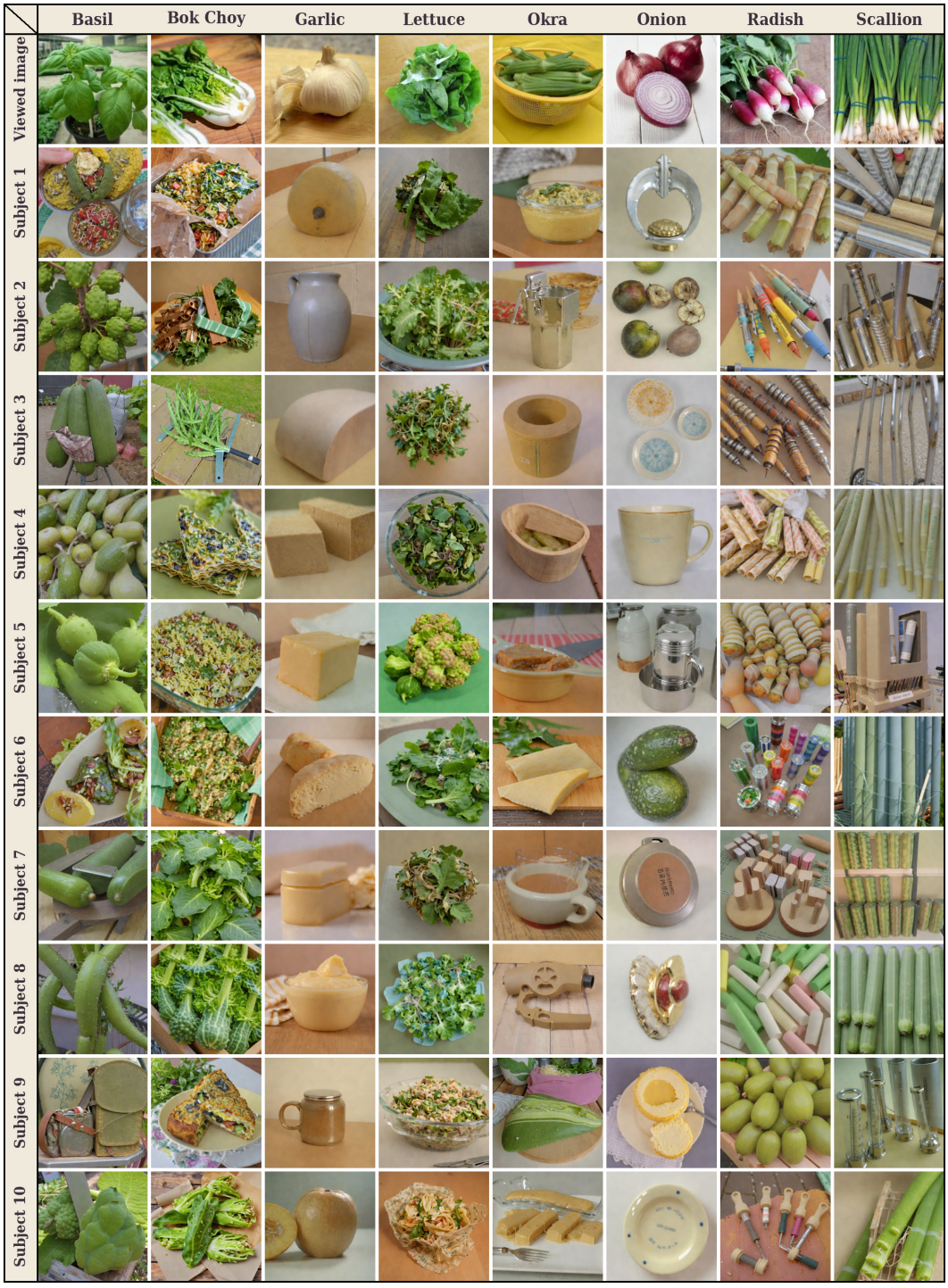

Figure 10: Viewed images (top) and participant-level reconstructions from 80-repetition-averaged EEG responses for eight vegetable and herb stimuli.

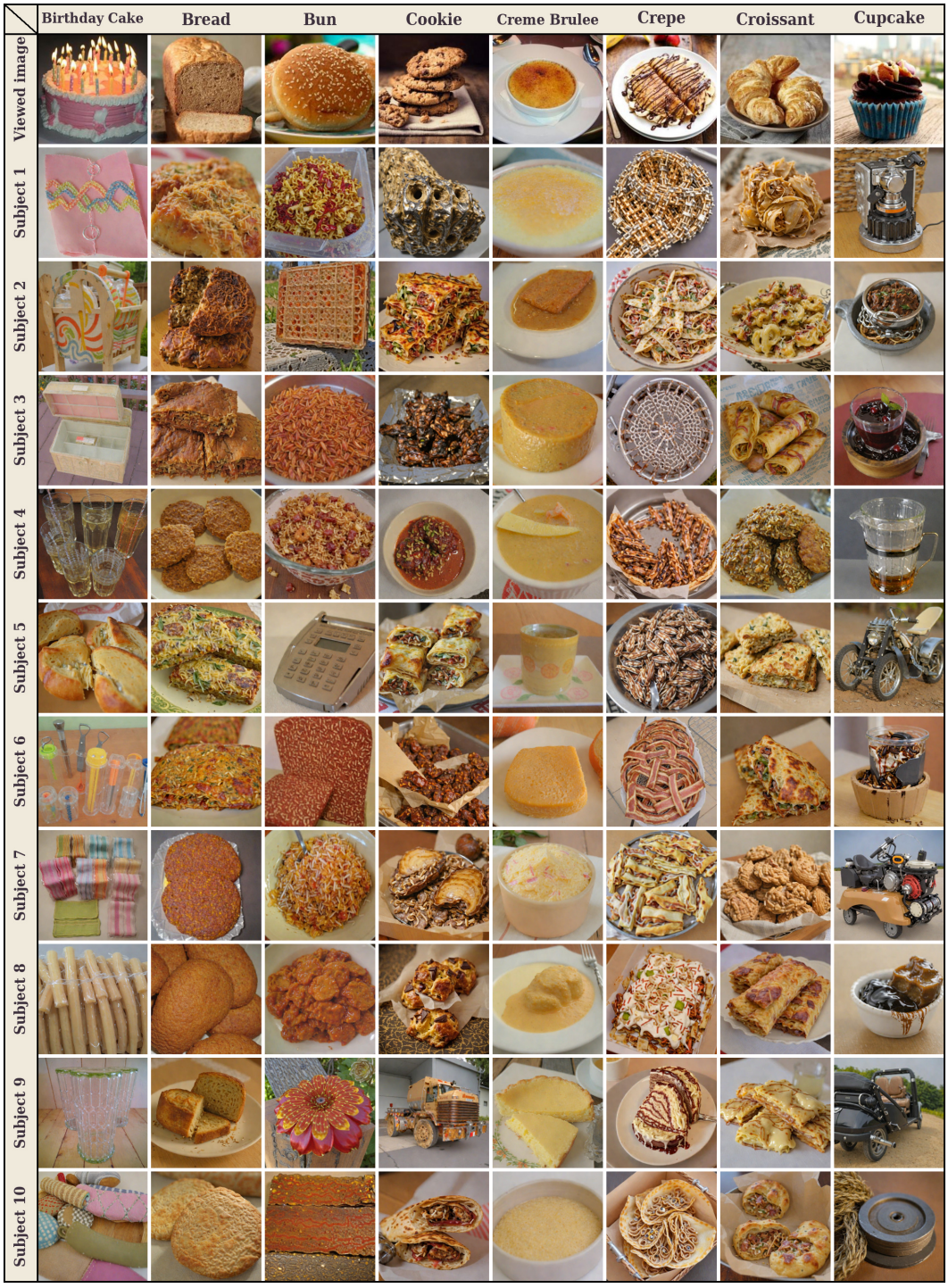

Figure 11: Viewed images (top) and participant-level reconstructions from 80-repetition-averaged EEG responses for eight baked-good and dessert stimuli.

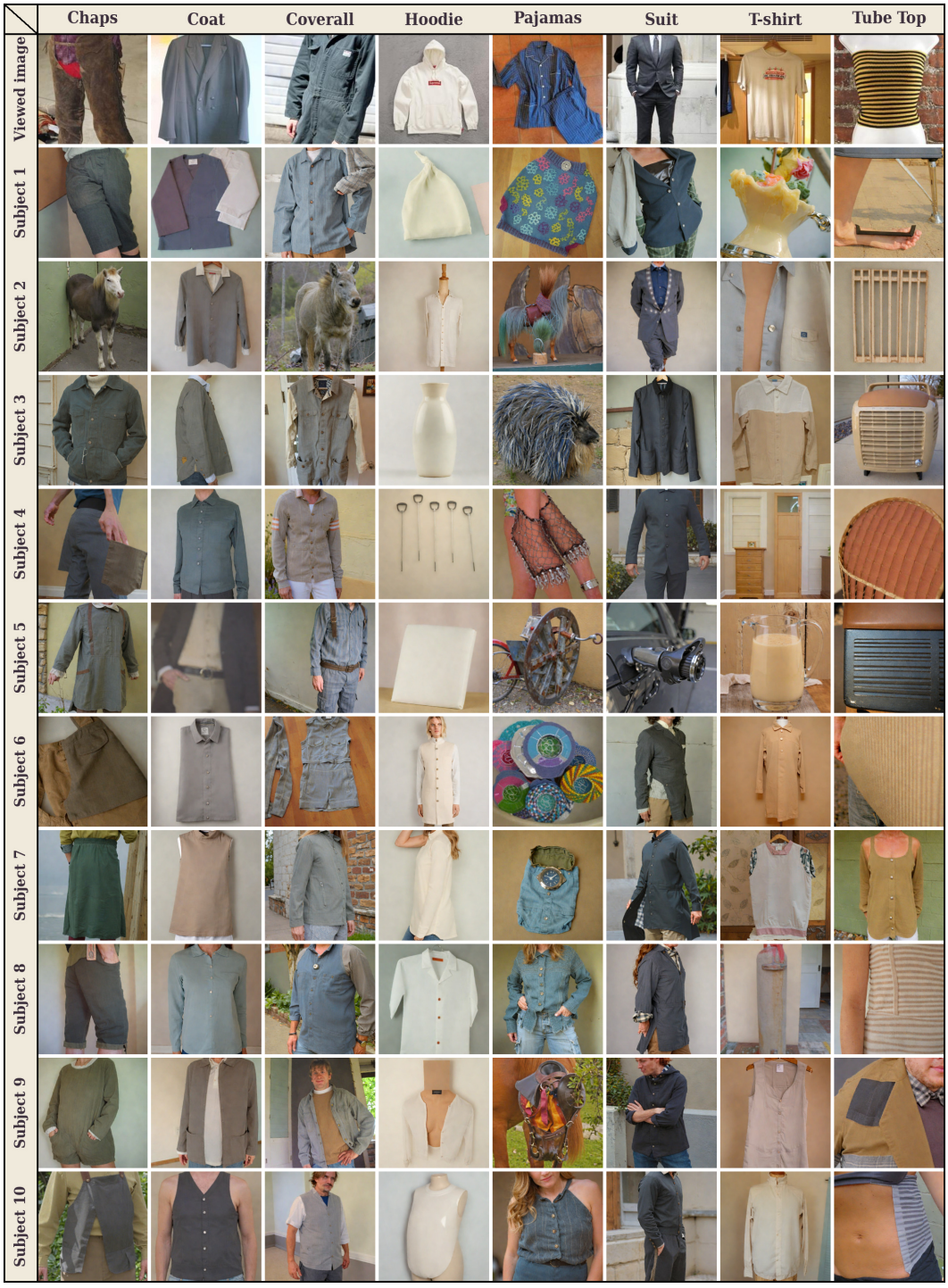

Figure 12: Viewed images (top) and participant-level reconstructions from 80-repetition-averaged EEG responses for eight clothing stimuli.

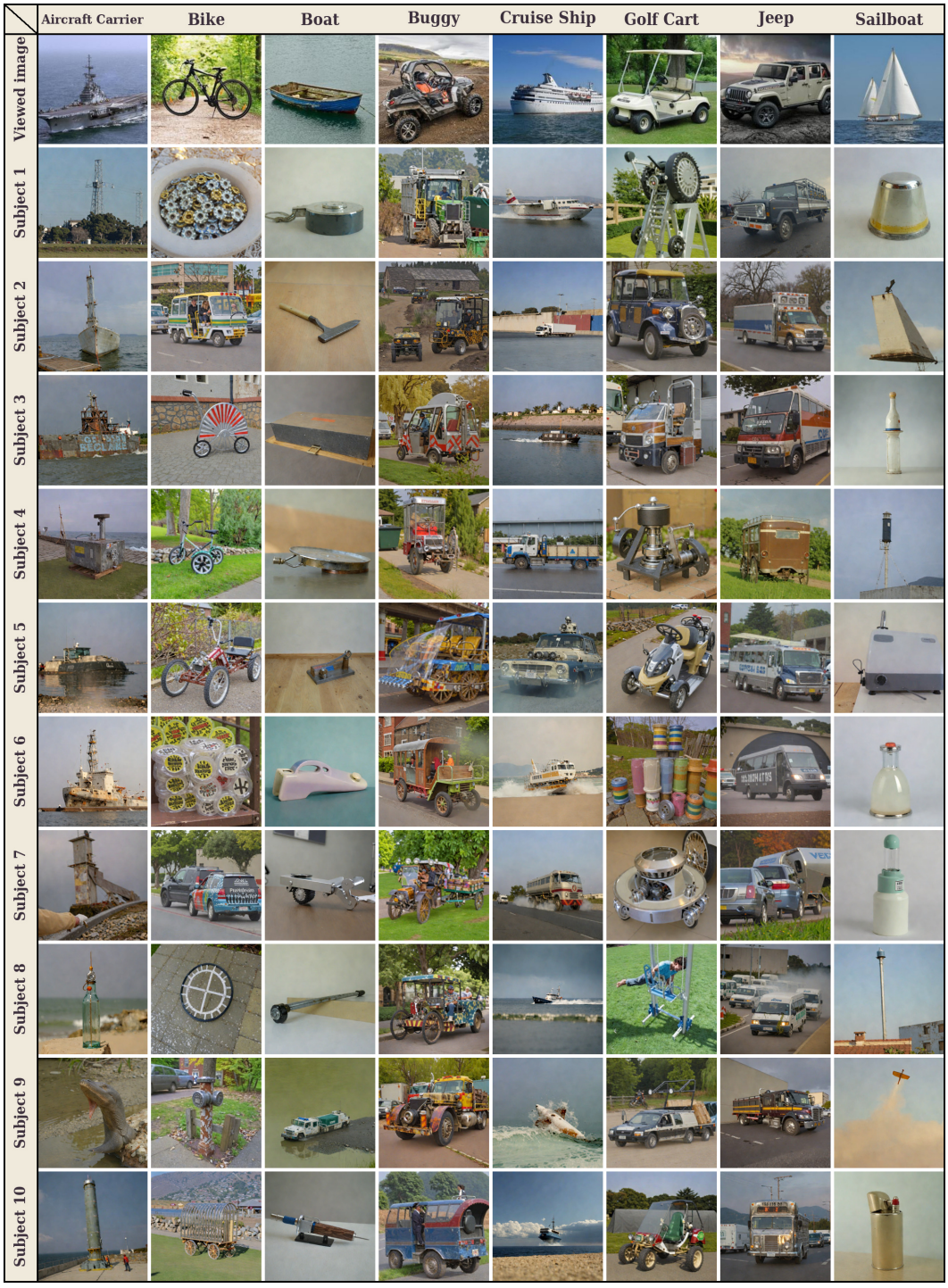

Figure 13: Viewed images (top) and participant-level reconstructions from 80-repetition-averaged EEG responses for eight vehicle stimuli.

J Limitations and Future Directions

RPA improves the visual representation used for brain–image alignment, but decoding performance remains dependent on repeated-response averaging and is substantially lower under LOSO evaluation. Learning effectively from individual responses and transferring across participants and record-

ing conditions therefore remain important directions. Our evaluation is also restricted to THINGS-EEG2 and THINGS-MEG with fixed 200-image galleries. In particular, the THINGS-EEG2 test gallery contains one image per unseen concept and does not directly evaluate discrimination among different images of the same concept. Larger galleries with visually similar candidates would help establish whether learned patch-token adaptation supports fine-grained visual discrimination. Although we examine multiple visual encoders and depths, the detailed adapter comparisons and token interventions concentrate on one selected intermediate representation. Extending token adaptation across layers and investigating time-resolved alignment could establish whether complementary visual representations improve decoding beyond the current single-layer design. Finally, the auxiliary feature probes measure descriptor predictability independently of RPA, while attribution maps characterize model behavior rather than establish the contribution of individual visual attributes. Controlled stimulus manipulations and direct analyses of the learned visual embedding could clarify how semantic and lower-level information contributes to brain–image alignment.