SHADOW BRAIN: BIOMIMETIC LEARNING FOR DECODING VISUAL NEURAL SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual neural decoding aims to infer the visual information perceived by an individual from brain activity and is key to connecting biological perception and artificial intelligence in brain–computer interface research. Visual information undergoes parallel processing, selective convergence, and hierarchical integration along biological visual pathways, giving rise to neural responses that vary across individuals. This variability, together with the low signal-to-noise ratio of electroencephalography (EEG), makes accurate alignment between visual information and EEG responses challenging. We propose Shadow Brain, a biomimetic visual–EEG learning framework built on the information-processing principles of the visual pathway from the retina through the lateral geniculate nucleus to the visual cortex. The visual branch draws on the functional sequence from retinal photoreception and signal processing to ganglion cell output. It constructs feature representations under five visual conditions and selectively integrates them using subject-specific learnable weights shared across stimuli within each subject. Inspired by population-level information integration in the visual cortex, the EEG branch employs NeuroWeft as a unified module to jointly model spatial relationships and temporal structure in multichannel EEG. Through contrastive learning, visual fusion, NeuroWeft, and cross-modal projectors are jointly optimized in a shared semantic space, adapting visual information integration to individual EEG signals. Within-subject 200-way image retrieval experiments using trial-averaged EEG from ten subjects in THINGS-EEG achieve mean Top-1 and Top-5 accuracies of 68.4% and 90.9%, respectively. Shadow Brain incorporates biological visual processing principles into visual–EEG correspondence learning, shifting from direct visual–EEG alignment to personalized visual–neural alignment constrained by biological visual processing principles.

1 INTRODUCTION

Visual neural decoding aims to recover information about perceived stimuli from brain activity (Kay et al., 2008). Cross-modal contrastive learning maps electroencephalography (EEG) signals and image representations into a shared space for zero-shot EEG-to-image retrieval (Song et al., 2024; Li et al., 2024; 2025). Many approaches use pretrained image features as fixed alignment targets for encoded EEG. However, biological vision transforms stimulus information through parallel processing and local integration before cortical population activity contributes to scalp EEG (Buzsáki et al., 2012). Subject variation also challenges generalization in EEG visual decoding (Huang et al., 2025); individual representational geometry is associated with perceived similarity (Charest et al., 2014). Recent work improves machine–human visual alignment (Muttenthaler et al., 2025) and primate cortical alignment (Yerxa et al., 2024), and introduces cognitive priors into EEG decoding (Zhang et al., 2026b). These findings motivate learning which visual information best aligns with individual neural responses, beyond matching EEG to a single predetermined image representation.

The biological visual system combines parallel pathways with successive integration stages (Figure 1). Retinal circuits distribute features across bipolar-cell pathways (Franke et al., 2017; Hsiang et al., 2024), with diverse ganglion-cell outputs (Baden et al., 2016). Center–surround organization and nonlinear pooling shape spatial integration (Dacey et al., 2000; Schreyer & Gollisch, 2021; Karamanlis et al., 2025); retinal sampling limits visual resolution (Rossi & Roorda, 2010), and cone noise affects output reliability (Ala-Laurila et al., 2011). Ganglion-cell signals reach visual cortex

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.

through the LGN (Nassi & Callaway, 2009). At the cortical level, neural-response-trained models generalize to new stimulus types (Wang et al., 2025a). These findings motivate complementary visual representations whose integration is learned from neural responses.

Inspired by parallel processing in the retina (Franke et al., 2017) and distributed cortical category representations (Huth et al., 2012), we propose Shadow Brain, a biomimetic framework for cross-modal visual–neural learning. The visual branch constructs five conditions with different informational properties for each stimulus and learns subject-specific fusion coefficients shared across stimuli under the EEG alignment objective, allowing individual neural responses to guide visual information selection and integration. The EEG branch uses the unified NeuroWeft encoder to jointly model signal details, channel interactions, and spatiotemporal structure. Visual fusion, NeuroWeft, and the cross-modal projectors are jointly optimized through contrastive learning to establish visual–EEG correspondences adapted to individual neural responses. Within-subject and cross-subject retrieval experiments and ablations on THINGS-EEG validate the full framework and its two core designs.

Our main contributions are as follows:

1. **A visual–neural decoding framework driven by biological visual principles.** We propose **Shadow Brain**, which unifies parallel visual processing, selective information integration, and distributed neural representation within a visual–EEG cross-modal learning framework, allowing the formation of visual representations to participate in learning visual–neural correspondence.

2. **Neural-response-driven, subject-specific visual integration.** We construct five parallel visual representations with different informational properties and propose a subject-specific learnable fusion mechanism. EEG responses help determine the relative contributions of different visual information, yielding visual representations adapted to individual neural responses.

3. **NeuroWeft for multichannel EEG representation.** NeuroWeft integrates signal detail processing, cross-channel interactions, and spatiotemporal encoding within a unified framework to learn neural information distributed across channels and time in visually evoked EEG.

4. **Experimental validation.** On within-subject 200-way image retrieval using THINGS-EEG, Shadow Brain achieves mean accuracies of **68.4% Top-1 and 90.9% Top-5** across ten subjects. Visual-side comparisons and NeuroWeft ablations further validate the effectiveness of the two core designs.

2 Related Work

Visual representations for neural decoding. BraVL aligns brain, visual, and linguistic features (Du et al., 2023); NICE and ATM learn EEG–image correspondences (Song et al., 2024; Li et al., 2024). DreamDiffusion combines masked EEG modeling with diffusion-based generation (Bai et al., 2024). CognitionCapturer uses multimodal experts (Zhang et al., 2025), extended with asymmetric alignment in CogCapPro (Zhang et al., 2026a). UBP introduces a blur prior (Wu et al., 2025); NeuroBridge averages augmented features (Zhang et al., 2026b); and ViEEG uses contour, object, and scene representations (Liu et al., 2026). NEED studies cross-subject and cross-task reconstruction (Huang et al., 2025), while Neuro-3D extends decoding to colored 3D objects (Guo et al., 2025). EEG2Video and EEGMirror address dynamic visual decoding (Liu et al., 2024; 2025); SeeEEG explores semantic retrieval-augmented generation (Kim et al., 2025). In fMRI, MindEye2 learns shared-subject representations (Scotti et al., 2024). Shadow Brain learns subject-specific weights over five visual conditions, shared across stimuli and jointly optimized with NeuroWeft.

EEG representation learning. LaBraM predicts masked neural tokens (Jiang et al., 2024), EEGPT learns general-purpose EEG representations through self-supervised pretraining (Wang et al., 2024), and EEG2Rep predicts masked latent representations (Mohammadi Foumani et al., 2024). CBraMod separately models spatial and temporal dependencies (Wang et al., 2025b). REVE accommodates heterogeneous electrode arrangements (El Ouahidi et al., 2025), while DeeperBrain incorporates volume-conduction and neural-dynamics priors (Wang et al., 2026). These pretrained models provide context for NeuroWeft's complementary focus: learning signal refinement, channel interaction, and spatiotemporal encoding from scratch within the visual–EEG alignment objective.

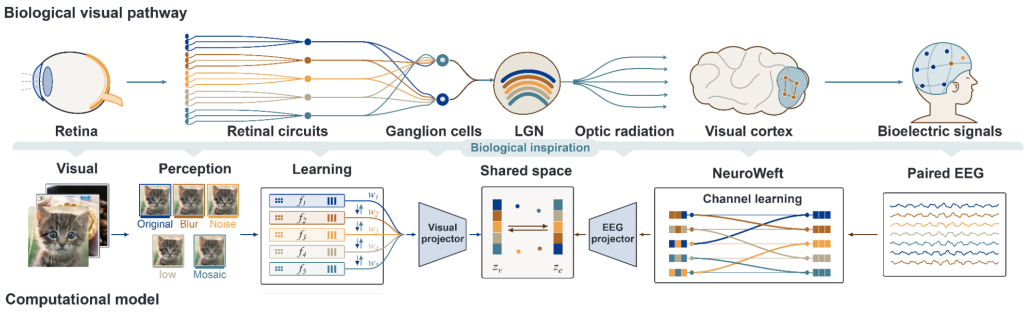

Figure 1: Shadow Brain framework. Biological processing motifs along the retina–LGN–visual cortex pathway. Five visual conditions are encoded into frozen features and integrated with subject-specific weights. NeuroWeft jointly encodes EEG spatial relationships and temporal structure. The two branches are optimized through cross-modal contrastive learning.

Retinal processing as a design principle. Recordings reveal distributed feature coding across parallel bipolar-cell pathways (Hsiang et al., 2024) and nonlinear pooling in ganglion-cell responses to natural scenes (Karamanlis et al., 2025). These findings motivate parallel representations and selective integration. Biophysical adaptation models also improve retinal-response prediction under changing inputs (Idrees et al., 2024). Our five conditions operationalize variation in spatial detail, local integration, sampling, and reliability as computational design choices, without assuming a one-to-one mapping to retinal cell types.

3 Method

3.1 Framework And Biological Motivation

Biological vision combines parallel processing and distributed representation (Figure 1). Bipolar-cell pathways encode distinct stimulus features (Franke et al., 2017; Hsiang et al., 2024), and nonlinear pooling shapes retinal output (Karamanlis et al., 2025). Along the retina–LGN–cortical pathway (Nassi & Callaway, 2009), attention signals fed back from the visual cortex can modulate the response gain of the LGN (O'Connor et al., 2002), before distributed cortical category representations emerge (Huth et al., 2012).

For subject s, let $D_s = \{(x_i, e_i^s)\}_{i=1}^{N_s}$ denote paired visual stimuli and EEG responses. Shadow Brain translates the retinal–LGN–cortical processing sequence into a learnable framework. Parallel retinal processing motivates five complementary visual conditions; selective convergence motivates subject-specific feature coefficients shared across stimuli; and distributed neural representation motivates NeuroWeft's joint encoding of multichannel EEG structure and temporal information. Cross-modal projectors bring both branches into a shared space, where contrastive learning adapts visual integration and EEG encoding to individual responses.

3.2 Five Visual Conditions

We construct five parallel visual conditions for each stimulus: original, Gaussian blur, Gaussian noise, low resolution, and mosaic. The original retains image detail; blur and mosaic vary local spatial integration, low resolution restricts spatial sampling, and noise perturbs information reliability. These operations provide controlled visual variations for neural alignment. Following the feature-extraction pipeline used for the released augmented features, Gaussian blur uses a fixed 15×15 kernel, Gaussian noise has zero mean and pixel-space standard deviation 25, low resolution downsamples to 10% of the original width and height before bilinear upsampling, and mosaic uses level 5 nearest-neighbor block reconstruction. Let $\mathcal{V} = \{\text{orig}, \text{blur}, \text{noise}, \text{low}, \text{mosaic}\}$ index the five visual conditions and T_k denote the corresponding image operation. A frozen CLIP RN50 encoder F (Radford et al., 2021), with pretrained weights loaded through OpenCLIP (Cherti et al., 2023), produces precomputed features $f_{i,k} = F(T_k(x_i))$ with $d = 1024$. The precomputed CLIP RN50

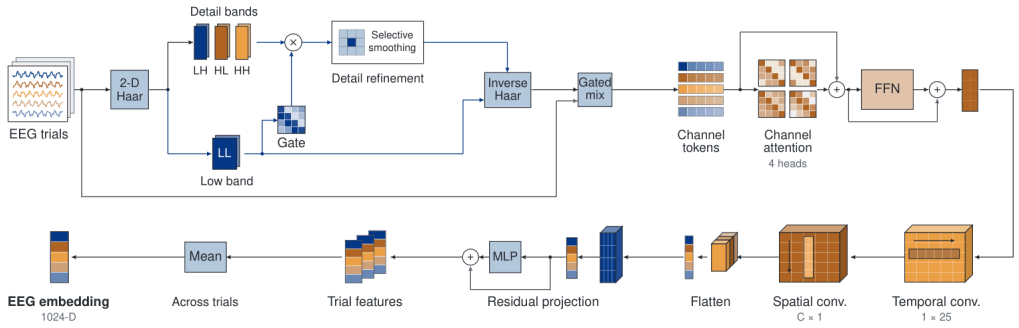

Figure 2: NeuroWeft as an EEG representation module. The NeuroWeft architecture combines low-band-guided detail refinement and gated reconstruction, channel interaction, and spatiotemporal encoding with residual projection and trial aggregation. Residual and low-band paths connect the stages within the encoder.

features are the inputs to selective fusion. The same five conditions are used to construct candidate image representations for retrieval.

3.3 Subject-Specific Selective Fusion

In within-subject learning, each subject model has five trainable logits $a_{s,k}$. Softmax converts the five logits into a normalized mixture:

$$w_{s,k} = \frac{\exp(a_{s,k})}{\sum_{\ell \in \mathcal{V}} \exp(a_{s,\ell})}, \qquad h_i^s = \sum_{k \in \mathcal{V}} w_{s,k} f_{i,k}. \tag{1}$$

The visual fusion coefficients are initialized equally and shared across all stimuli for subject s. The fusion coefficients remain fixed across candidates at inference. Subject-specific fusion introduces five logits per subject, corresponding to four independent mixture degrees of freedom. The result is a persistent subject-level fusion profile across stimuli.

Retinal pooling provides a functional motivation for combining parallel inputs (Karamanlis et al., 2025); our weighted mixture implements a computational abstraction of selective integration. Paired EEG supplies the learning signal that sets relative contributions. The visual fusion coefficients quantify each condition's contribution to the EEG-aligned visual representation.

For equal fusion, $w_{s,k} = 1/5$. For a single-condition model, the selected condition has coefficient one and all other coefficients are zero. Equal fusion and single-condition models use the same representation family and define the visual comparisons in Section 4.

For leave-one-subject-out evaluation, each fold learns one fusion vector jointly from the training participants. The jointly learned fusion vector is shared across the training participants' stimuli and reused for the held-out participant without fitting new participant-specific fusion parameters.

3.4 Cross-Modal Learning And Retrieval

Let G_s denote the NeuroWeft EEG encoder, and let P_v and P_e denote the separate visual and EEG projection mappings. The alignment configuration normalizes image features and retains EEG feature magnitudes:

$$z_i^{v,s} = \frac{P_v(h_i^s)}{\| P_v(h_i^s) \|_2}, \qquad z_i^{e,s} = P_e(G_s(e_i^s)). \tag{2}$$

NeuroWeft jointly models multichannel EEG as a unified representation module (Figure 2). Its signal-refinement stage combines low-band-guided detail processing with gated reconstruction us-

ing the Haar analysis and synthesis shown in Figure 2 (Haar, 1910). Channel interaction and spatiotemporal encoding then organize the refined signals, followed by residual projection and trial aggregation. For a paired mini-batch of size B, define $u_{ij} = \langle z_i^{e,s}, z_j^{v,s} \rangle / \tau$, where $\tau > 0$ controls logit scaling. Bidirectional contrastive alignment can be expressed as (Radford et al., 2021)

$$\mathcal{L} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(u_{ii})}{\sum_{j=1}^{B} \exp(u_{ij})} + \log \frac{\exp(u_{ii})}{\sum_{j=1}^{B} \exp(u_{ji})} \right]. \tag{3}$$

The alignment signal updates the fusion logits, EEG encoder, and projection parameters, while the pretrained visual features remain fixed. Thus the mixture is optimized for cross-modal matching. Appendix A.1 gives the resulting gradient with respect to a fusion logit.

At test time, a query EEG embedding is compared with all candidate-image embeddings from the corresponding subject or held-out-fold model. Candidates are ranked by cosine similarity, and Top-*K* accuracy records whether the matched image appears among the first *K* entries. For 200 candidates, uniform random rankings yield expected Top-1 and Top-5 accuracies of 0.5% and 2.5%, respectively.

The biologically inspired visual recognition framework gives Shadow Brain parallel visual representations, inspectable subject-specific integration, and spatiotemporal EEG encoding. Joint optimization allows individual neural responses to shape the visual mixture while learning an EEG representation for cross-modal retrieval.

4 Experiments

4.1 Experimental Setup

Data and preprocessing. THINGS-EEG (Gifford et al., 2022) contains ten participants. Training comprises 1,654 concepts, ten images per concept, and four repetitions per image; testing uses 200 concepts, one image per concept, and 80 repetitions. We follow NeuroBridge and its extended preprocessing details (Zhang et al., 2026b): 1,000-Hz recordings are filtered at 0.1–100 Hz, baseline-corrected using -200–0 ms, epoched to 0–1,000 ms after stimulus onset, downsampled to 250 Hz, and processed with multivariate noise normalization (MVNN) (Guggenmos et al., 2018). Averaging repetitions of each image yields 16,540 training and 200 test samples per participant. Within-subject inputs use 17 posterior electrodes and 250 time samples. Preprocessing choices can materially affect EEG decoding (Kessler et al., 2025); Appendix E.1 records our fixed input pipeline.

Evaluation protocols. Within-subject evaluation trains a separate model for each participant. EEG leave-one-subject-out evaluation comprises ten folds, each training on nine participants and testing on the remaining participant. Each fold learns one shared fusion vector, as specified in Section 3.3. Each query is ranked against all 200 candidate images by cosine similarity, with Top-1 and Top-5 accuracy reported. These trial-averaged evaluations differ from the few-repetition setting studied by NEAR (Cui et al., 2026).

Implementation. Shadow Brain uses PyTorch with CUDA acceleration. Frozen CLIP RN50 features have 1024 dimensions per visual condition. NeuroWeft, five fusion logits, and two linear projectors into 512 dimensions are jointly optimized from scratch. Both EEG protocols train for 50 epochs with batch size 64 using AdamW (Loshchilov & Hutter, 2019); learning rate and weight decay are both 10^{-4}, and contrastive temperature is initialized to $\tau_0 = 0.07$. Training EEG is randomly smoothed with kernel size five and probability 0.3; test-time smoothing is disabled. Appendix E gives additional details.

Repeated runs. Except for Figure 3a, Shadow Brain retrieval accuracies are arithmetic means of five independent runs with different random seeds, reported per participant and averaged across participants. Figure 3 uses seed 2026 to examine visual fusion parameters. Appendix A.2 gives the fusion-parameter snapshots and trajectories; Appendix B gives MEG details.

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

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Average
Intra-Subject: train and test on one subject												
BraVL	Top-1	6.1	4.9	5.6	5.0	4.0	6.0	6.5	8.8	4.3	7.0	5.8
	Top-5	17.9	14.9	17.4	15.1	13.4	18.2	20.4	23.7	14.0	19.7	17.5
NICE	Top-1	13.2	13.5	14.5	20.6	10.1	16.5	17.0	22.9	15.4	17.4	16.1
	Top-5	39.5	40.3	42.7	52.7	31.5	44.0	42.1	56.1	41.6	45.8	43.6
UBP	Top-1	41.2	51.2	51.2	51.1	42.2	57.5	49.0	58.6	45.1	61.5	50.9
	Top-5	70.5	80.9	82.0	76.9	72.8	83.5	79.9	85.8	76.2	88.2	79.7
CogCapPro	Top-1	54.2	50.5	49.5	56.5	40.9	52.3	50.8	63.1	49.5	59.3	52.7
	Top-5	83.2	82.9	85.6	85.4	74.0	86.2	79.5	90.0	81.7	86.5	83.5
NeuroBridge	Top-1	50.0	63.2	61.6	61.4	54.8	69.7	62.7	71.2	64.0	73.6	63.2
	Top-5	77.6	90.6	91.1	**90.0**	85.0	92.9	**88.8**	**95.1**	**91.0**	**97.1**	89.9
Shadow Brain (ours)	Top-1	**65.3**	**64.7**	**63.2**	**65.8**	**68.3**	**73.2**	**63.7**	**76.3**	**67.7**	**75.8**	**68.4**
	Top-5	**87.3**	**91.7**	**92.2**	88.3	**86.8**	**94.2**	87.3	94.2	90.8	95.7	**90.9**
Inter-Subject: leave one subject out for test												
BraVL	Top-1	2.3	1.5	1.4	1.7	1.5	1.8	2.1	2.2	1.6	2.3	1.8
	Top-5	8.0	6.3	5.9	6.7	5.6	7.2	8.1	7.6	6.4	8.5	7.0
NICE	Top-1	7.6	5.9	6.0	6.3	4.4	5.6	5.6	6.3	5.7	8.4	6.2
	Top-5	22.8	20.5	22.3	20.7	18.3	22.2	19.7	22.0	17.6	28.3	21.4
UBP	Top-1	11.5	15.5	9.8	13.0	8.8	11.7	10.2	12.2	15.5	16.0	12.4
	Top-5	29.7	40.0	27.0	32.3	33.8	31.0	23.8	32.2	40.5	43.5	33.4
NeuroBridge	Top-1	23.2	21.2	**13.2**	17.0	14.5	**25.0**	15.3	20.1	13.7	27.2	19.0
	Top-5	**52.4**	49.3	36.5	**45.3**	37.7	**55.0**	45.1	44.9	36.5	**56.3**	45.9
Shadow Brain (ours)	Top-1	**24.3**	**27.7**	11.7	**19.3**	**15.7**	22.2	**17.3**	**20.8**	**17.4**	**27.8**	**20.4**
	Top-5	50.3	**50.2**	**37.2**	42.3	**38.7**	45.8	**45.8**	**47.0**	**50.5**	55.5	**46.3**

4.2 Zero-Shot EEG-to-Image Retrieval Performance

Table 1 compares Shadow Brain with published results for BraVL (Du et al., 2023), NICE (Song et al., 2024), UBP (Wu et al., 2025), CogCapPro (Zhang et al., 2026a), and NeuroBridge (Zhang et al., 2026b). Shadow Brain uses the settings in Section 4.1.

In within-subject retrieval, Shadow Brain achieves the highest mean Top-1 and Top-5 accuracies in Table 1, at 68.4% and 90.9%, exceeding NeuroBridge by 5.2 and 1.0 percentage points. Top-1 ranges from 63.2% to 76.3% and leads all listed methods for every subject.

In cross-subject retrieval, leave-one-subject-out Top-1 and Top-5 accuracies reach 20.4% and 46.3%, exceeding the next-highest averages by 1.4 and 0.4 percentage points. The fusion vector learned from training subjects is applied directly to the held-out subject, supporting retrieval without fitting subject-specific fusion parameters.

4.3 Visual Conditions And Learned Fusion

Figure 3 compares individual conditions, equal fusion, and learned fusion on sub-07, and shows fusion parameters and training trajectories across subjects.

Among the individual conditions, the strongest individual condition, Gaussian noise, achieves 46.0% Top-1; equal fusion reaches 54.5%, and learned fusion reaches 63.5% (Figure 3a). Learning the coefficients improves Top-1 by 9.0 percentage points over equal fusion and 22.5 points over original images. Learned fusion also achieves 87.0% Top-5, compared with 84.5% for equal fusion and 77.5% for Gaussian noise; blur and mosaic each reach 80.0%. Both combining conditions and learning their relative contributions improve retrieval.

Across subjects, mosaic receives the largest coefficient in nine subjects, while noise leads in sub-04 (Figure 3b,c). Coefficients are shared across stimuli within each subject.

The learned coefficients represent relative shares in the joint representation. Although noise has the highest individual-condition Top-1 on sub-07, mosaic receives a larger coefficient (23.91% versus 23.51%). Noise and mosaic shares increase from the 20% initialization (Figure 3d); epoch 41 marks the displayed sub-07 snapshot. Appendix A provides the gradient analysis and all-subject trajectories explaining how EEG alignment updates the mixture.

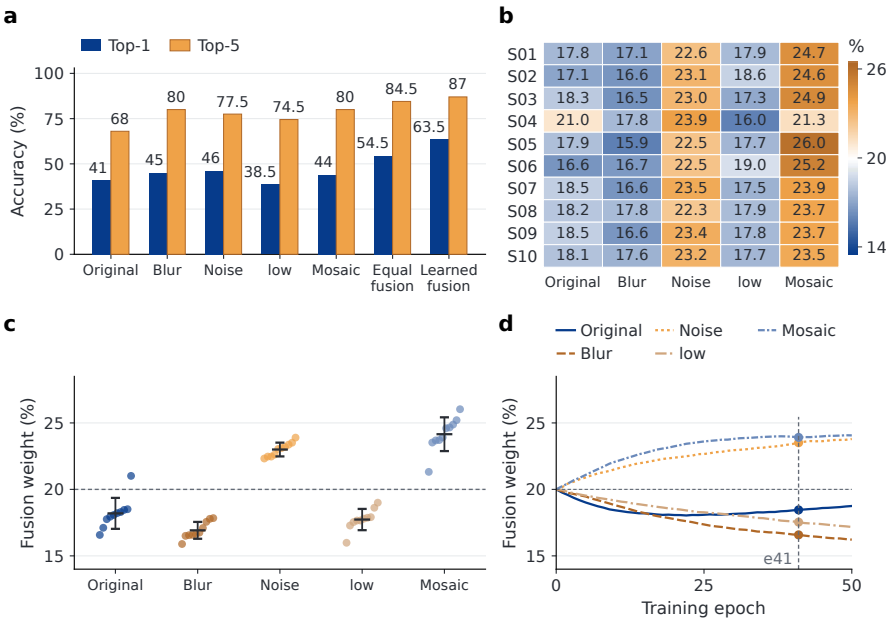

Figure 3: Visual-condition retrieval and learned mixtures for seed 2026. **a** Sub-07 accuracy for five individual conditions, equal fusion, and learned fusion. **b** Five-condition weights for ten subjects. **c** Across-subject mean and sample SD (n = 10), with individual weights. **d** Sub-07 trajectories; epoch 41 marks the snapshot in **b,c**. Dashed lines indicate equal fusion (20%). Weights are shared across stimuli within each subject.

4.4 Category Structure In Image And EEG Representations

We visualize category structure with t-SNE (van der Maaten & Hinton, 2008) visualizations of sub-07 contain 117 points per modality from animal, food, and tool categories (Figure 4). Both modalities show local category-related clustering with within-category variation; animals occupy a relatively distinct region, while food and tools partly overlap.

Across subjects, pooled EEG features show more dispersed groups and category mixing across 1,170 records per modality (Appendix C). Category organization complements image-level retrieval in Table 1 and Section 4.6.

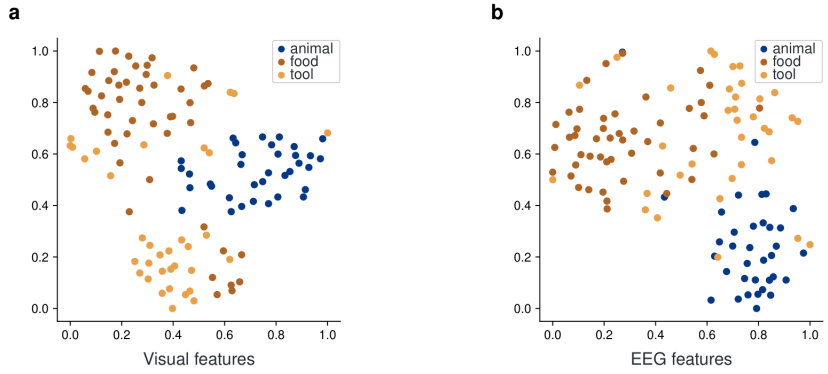

Figure 4: Category structure for sub-07 visualized with t-SNE. **a** Visual features and **b** EEG features, colored by animal, food, and tool. Each panel contains 117 points. The panels show category organization within each modality.

4.5 NeuroWeft Ablation And Signal Transformations

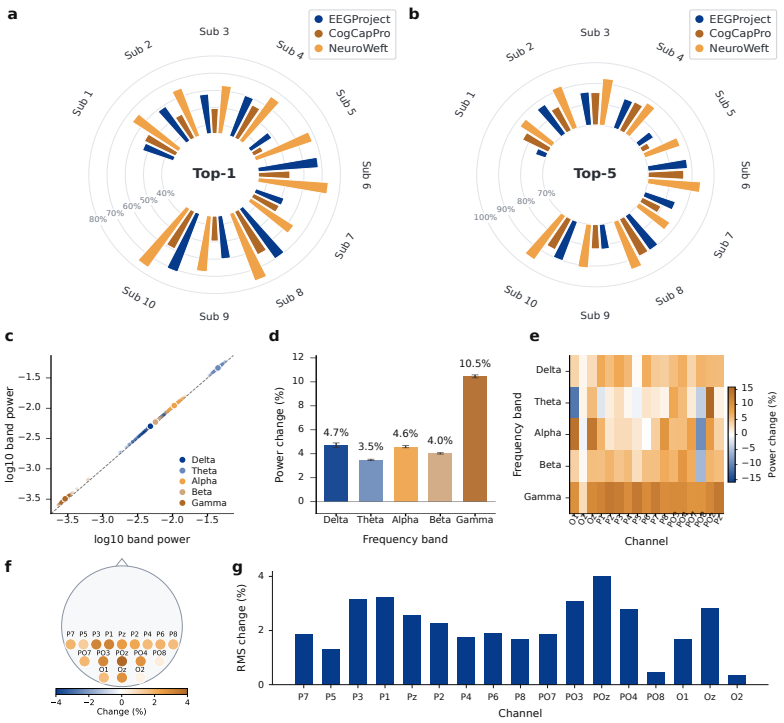

Figure 5: EEG encoder comparison and channel-level signal changes. **a,b** Top-1 and Top-5 retrieval accuracies across ten subjects for EEGProject, CogCapPro, and NeuroWeft, averaged over five independent runs. **c** Paired $\log_{10}$ band power on the horizontal axis (before processing) and vertical axis (after processing). **d** Mean relative band-power changes with query-bootstrap 95% confidence intervals. **e** Band-by-channel power changes. **f,g** Relative RMS changes with schematic electrode positions. Panels c–g use 200 paired queries from sub-07 and 17 spatially whitened model-input channels; band power in c–e is estimated from 1–45 Hz Welch spectra (Welch, 1967).

Table 2 reports sub-07 retrieval averaged over five independent runs with visual fusion, NeuroWeft, both, or neither. Here, "visual fusion" means the five-condition trainable mixture in Section 3.3; when it is disabled, the visual input is the original-condition feature. "NeuroWeft" denotes the full EEG encoder in Figure 2; when it is disabled, EEGProject is used as the baseline EEG encoder. The full model reaches 63.7% Top-1 and 87.3% Top-5, exceeding the configuration without either component by 27.0 and 28.4 percentage points. Adding NeuroWeft to visual fusion improves Top-1 by 12.2 points, confirming its contribution beyond visual fusion.

Table 2: Sub-07 visual fusion and NeuroWeft ablation.

Visual fusion	NeuroWeft	Top-1 (%)	Top-5 (%)
×	×	36.7	58.9
×	✓	41.6	68.3
✓	×	51.5	84.7
✓	✓	**63.7**	**87.3**

Figure 5a,b compares EEGProject, CogCapPro (Zhang et al., 2026a), and NeuroWeft across ten subjects using Top-1 and Top-5 accuracies averaged over five independent runs. For sub-07, the EEGProject bars correspond to the fusion-enabled baseline-encoder configuration in Table 2.

For the signal analysis, we use 200 paired queries and 17 posterior input channels from sub-07. Oz waveforms nearly overlap, with small structured residuals and preserved spectral peaks (Appendix D.1). Band-power changes vary by frequency, with the largest mean relative increase, 10.5%, in gamma (Figure 5c,d).

Across channels, POz shows the largest relative root-mean-square (RMS) change, with smaller changes at PO8 and O2 (Figure 5e–g). Mean relative power spectral density (PSD) change over 1–60 Hz is also greatest at POz but negative at PO8 (Appendix D.2). PSD and RMS describe spectral distribution and overall amplitude, respectively, capturing complementary aspects of the nonuniform signal transformations.

4.6 Cross-Modal Similarities And Retrieval Examples

Figure 6a,b shows cosine similarities between 200 EEG queries and 200 candidate images from sub-07, with the clustered Top-5 view highlighting each query's highest-ranked candidates.

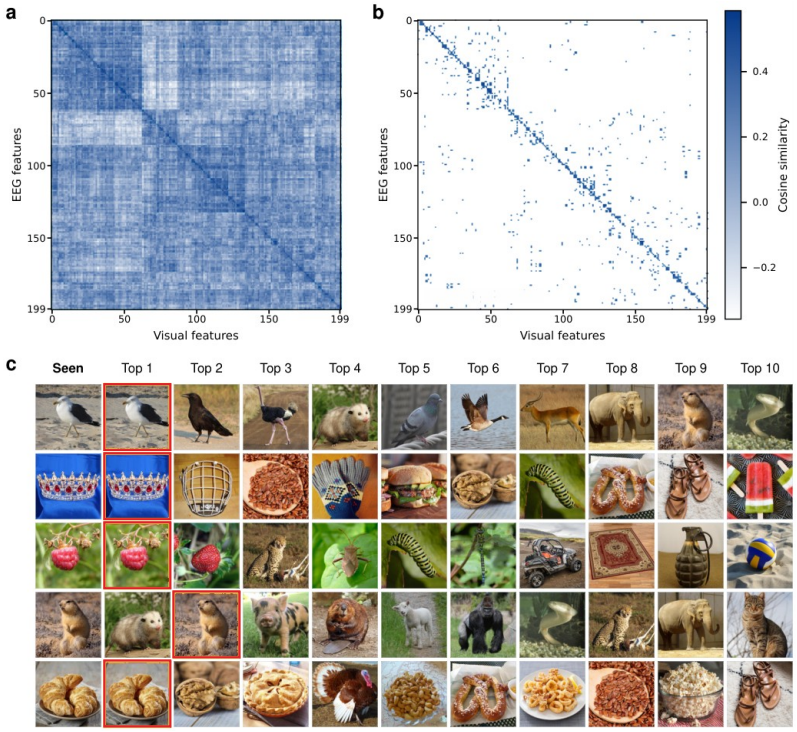

Figure 6: Sub-07 retrieval evidence. **a** Dense cosine similarities. **b** Clustered Top-5 similarities, with masked entries shown in white. Both matrices use Visual features as columns, EEG features as rows, and a shared color scale. **c** Five stimuli labeled "Seen" with Top-10 candidates; red borders mark matching images.

Among the five queries shown with Top-10 candidates, four matching images rank first and one ranks second (Figure 6c).

5 Conclusion

We present Shadow Brain, a biomimetic framework for visual–EEG alignment that combines subject-specific visual fusion with NeuroWeft. Joint optimization integrates parallel visual information according to individual EEG responses and learns spatiotemporal representations of multichannel signals. On THINGS-EEG, Shadow Brain achieves competitive performance in within-subject and cross-subject 200-way image retrieval. Ablation experiments support the contributions of visual fusion and NeuroWeft. By translating parallel processing, selective integration, and distributed neural encoding into learnable designs, Shadow Brain advances visual neural decoding and computational modeling inspired by biological mechanisms (LeCun et al., 2015). These results motivate personalized brain–computer interfaces guided by visual neural processing principles.

AI USE STATEMENT

In this work, Large Language Models (LLMs) were used solely to aid in the writing and polishing of the manuscript, including language editing, translation, sentence rephrasing, grammar checking, and improving readability and clarity.

LLMs were not used for research ideation, methodological or experimental design, data generation, model training or evaluation, result selection, or data analysis. All scientific content, experimental results, and conclusions were developed and verified by the authors.

All AI-assisted text was reviewed and revised by the authors. The authors take full responsibility for the final content of the manuscript.

REFERENCES

Petri Ala-Laurila, Martin Greschner, E. J. Chichilnisky, and Fred Rieke. Cone photoreceptor contributions to noise and correlations in the retinal output. *Nature Neuroscience*, 14(10):1309–1316, 2011. doi: 10.1038/nn.2927.

Tom Baden, Philipp Berens, Katrin Franke, Miroslav Román Rosón, Matthias Bethge, and Thomas Euler. The functional diversity of retinal ganglion cells in the mouse. *Nature*, 529(7586):345–350, 2016. doi: 10.1038/nature16468.

Yunpeng Bai, Xintao Wang, Yan-Pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. DreamDiffusion: High-Quality EEG-to-Image Generation with Temporal Masked Signal Modeling and CLIP Alignment. In *Computer Vision – ECCV 2024*, volume 15089 of *Lecture Notes in Computer Science*, pp. 472–488, 2024. doi: 10.1007/978-3-031-72751-1_27.

György Buzsáki, Costas A. Anastassiou, and Christof Koch. The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, 13(6):407–420, 2012. doi: 10.1038/nrn3241.

Ian Charest, Rogier A. Kievit, Taylor W. Schmitz, Diana Deca, and Nikolaus Kriegeskorte. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, 111(40):14565–14570, 2014. doi: 10.1073/pnas.1402594111.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023. doi: 10.1109/cvpr52729.2023.00276.

Zhenyao Cui, Siyuan Kan, Dingkun Liu, and Dongrui Wu. Beyond Trial Averaging: Anchoring Neural and Visual Representations for Few-Repetition Brain-to-Image Retrieval. arXiv preprint arXiv:2608.19128, 2026.

Dennis Dacey, Orin S. Packer, Lisa Diller, David Brainard, Beth Peterson, and Barry Lee. Center surround receptive field structure of cone bipolar cells in primate retina. *Vision Research*, 40(14):1801–1811, 2000. doi: 10.1016/S0042-6989(00)00039-0.

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.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A Foundation Model for EEG – Adapting to Any Setup with Large-Scale Pretraining on 25,000 Subjects. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0760.

Katrin Franke, Philipp Berens, Timm Schubert, Matthias Bethge, Thomas Euler, and Tom Baden. Inhibition decorrelates visual feature representations in the inner retina. *Nature*, 542(7642):439–444, 2017. doi: 10.1038/nature21394.

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. doi: 10.1016/j.neuroimage.2022.119754.

Matthias Guggenmos, Philipp Sterzer, and Radoslaw Martin Cichy. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, 173:434–447, 2018. doi: 10.1016/j.neuroimage.2018.02.044.

Zhanqiang Guo, Jiamin Wu, Yonghao Song, Jiahui Bu, Weijian Mai, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. Neuro-3D: Towards 3D Visual Decoding from EEG Signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23870–23880, 2025.

Alfred Haar. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69(3): 331–371, 1910.

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. doi: 10.7554/elife.82580.

Jen-Chun Hsiang, Ning Shen, Florentina Soto, and Daniel Kerschensteiner. Distributed feature representations of natural stimuli across parallel retinal pathways. *Nature Communications*, 15: 1920, 2024. doi: 10.1038/s41467-024-46348-y.

Shuai Huang, Huan Luo, Haodong Jing, Qixian Zhang, Litao Chang, Yating Feng, Xiao Lin, Chendong Qin, Han Chen, Shuwen Jia, Siyi Sun, and Yongxiong Wang. NEED: Cross-Subject and Cross-Task Generalization for Video and Image Reconstruction from EEG Signals. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5761.

Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, 76(6):1210–1224, 2012. doi: 10.1016/j.neuron.2012.10.014.

Saad Idrees, Michael B. Manookin, Fred Rieke, Greg D. Field, and Joel Zylberberg. Biophysical neural adaptation mechanisms enable artificial neural networks to capture dynamic retinal computation. *Nature Communications*, 15:5957, 2024. doi: 10.1038/s41467-024-50114-5.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *International Conference on Learning Representations*, 2024.

Dimokratis Karamanlis, Mohammad H. Khani, Helene M. Schreyer, Sören J. Zapp, Matthias Mietsch, and Tim Gollisch. Nonlinear receptive fields evoke redundant retinal coding of natural scenes. *Nature*, 637:394–401, 2025. doi: 10.1038/s41586-024-08212-3.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. doi: 10.1038/nature06713.

Roman Kessler, Alexander Enge, and Michael A. Skeide. How EEG preprocessing shapes decoding performance. *Communications Biology*, 8:1039, 2025. doi: 10.1038/s42003-025-08464-3.

Jun-Mo Kim, Woohyeok Choi, Sang-Jun Park, Keun-Soo Heo, Young-Han Son, Ji-Hye Oh, Dong-Hee Shin, and Tae-Eui Kam. SeeEEG: Semantic-aware EEG-based Multi-Modal Retrieval-Augmented Generation for High-Fidelity Visual Brain Decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4883–4892, 2025.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

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, pp. 102822–102864, 2024. doi: 10.52202/079017-3266.

Yueyang Li, Zijian Kang, Shengyu Gong, Wenhao Dong, Weiming Zeng, Hongjie Yan, Wai Ting Siok, and Nizhuan Wang. Neural-MCRL: Neural Multimodal Contrastive Representation Learning for EEG-based Visual Decoding. In *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/icme59968.2025.11210130.

Minxu Liu, Donghai Guan, Chuhang Zheng, Chunwei Tian, Jie Wen, and Qi Zhu. ViEEG: Hierarchical Visual Neural Representation for EEG Brain Decoding. In *International Conference on Machine Learning*, 2026.

Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. EEG2Video: Towards Decoding Dynamic Visual Perception from EEG Signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2306.

Xuan-Hao Liu, Bao-Liang Lu, and Wei-Long Zheng. EEGMirror: Leveraging EEG Data in the Wild via Montage-Agnostic Self-Supervision for EEG to Video Decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18273–18283, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: Enhancing Self-supervised EEG Representation Through Informative Masked Inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5544–5555, 2024. doi: 10.1145/3637528.3671600.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C. Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K. Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, 2025. doi: 10.1038/s41586-025-09631-6.

Jonathan J. Nassi and Edward M. Callaway. Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience*, 10(5):360–372, 2009. doi: 10.1038/nrn2619.

Daniel H. O'Connor, Miki M. Fukui, Mark A. Pinsk, and Sabine Kastner. Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, 5(11):1203–1209, 2002. doi: 10.1038/nn957.

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 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Ethan A. Rossi and Austin Roorda. The relationship between visual resolution and cone spacing in the human fovea. *Nature Neuroscience*, 13(2):156–157, 2010. doi: 10.1038/nn.2465.

Helene Marianne Schreyer and Tim Gollisch. Nonlinear spatial integration in retinal bipolar cells shapes the encoding of artificial and natural stimuli. *Neuron*, 109(10):1692–1706.e8, 2021. doi: 10.1016/j.neuron.2021.03.015.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 44038–44059, 2024.

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*, pp. 47648–47665, 2024.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

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

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1239.

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*, 2025b.

Jiquan Wang, Sha Zhao, Yangxuan Zhou, Yiming Kang, Shijian Li, and Gang Pan. DeeperBrain: A Neuro-Grounded EEG Foundation Model Towards Universal BCI. arXiv preprint arXiv:2601.06134, 2026.

Peter D. Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967. doi: 10.1109/TAU.1967.1161901.

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.

Thomas Yerxa, Jenelle Feather, Eero P. Simoncelli, and SueYeon Chung. Contrastive-Equivariant Self-Supervised Learning Improves Alignment with Primate Visual Area IT. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3043.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. CognitionCapturer: Decoding Visual Stimuli from Human EEG Signal with Multimodal Information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14486–14493, 2025. doi: 10.1609/aaai.v39i13.33587.

Kaifan Zhang, Lihuo He, Junjie Ke, Yuqi Ji, Lukun Wu, Lizi Wang, and Xinbo Gao. CognitionCapturerPro: Towards High-Fidelity Visual Decoding from EEG/MEG via Multi-Modal Information and Asymmetric Alignment. arXiv preprint arXiv:2603.12722, 2026a.

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, 2026b. doi: 10.1609/aaai.v40i21.38863. Extended version: arXiv:2511.06836.

A Additional Visual-Side Analysis

A.1 How Cross-Modal Alignment Updates Fusion Parameters

This section explains how the contributions of visual conditions change during joint optimization. For the fused representation $h_i^s = \sum_k w_{s,k} f_{i,k}$ defined in Section 3.3, normalized weights are obtained by applying softmax to five trainable parameters. The derivative of the fused representation with respect to the parameter $a_{s,k}$ is

$$\frac{\partial h_i^s}{\partial a_{s,k}} = w_{s,k} \left(f_{i,k} - h_i^s \right). \tag{4}$$

By the chain rule, the gradient of the training loss with respect to this parameter is

$$\frac{\partial \mathcal{L}}{\partial a_{s,k}} = \sum_i w_{s,k} \left\langle \frac{\partial \mathcal{L}}{\partial h_i^s}, f_{i,k} - h_i^s \right\rangle. \tag{5}$$

Equation 5 relates the difference between a condition's features and the current fused representation to the downstream cross-modal alignment gradient. Each parameter update depends on the condition's features, the current fused representation, and the alignment-loss gradient, and therefore varies with the shares of the other conditions in the mixture. This gradient relationship explains the difference between individual-condition accuracy rankings and fusion-coefficient rankings in Section 4.3: fusion coefficients are optimized by the joint alignment objective and describe relative shares in the current mixture.

A.2 Subject-Level Fusion Parameter Snapshots And Learning Trajectories

The fusion parameters in Figure 3b,c and the trajectories in Figure 7 come from the seed-2026 training record for each subject. The five coefficients are normalized to sum to one and are shared across stimuli within each subject. The trajectories show the values recorded at each epoch without additional temporal smoothing.

The snapshot epochs for sub-01 through sub-10 are 33, 47, 45, 37, 38, 34, 41, 29, 48, and 38, respectively. The epoch-41 marker in Figure 3d corresponds to sub-07; the other subjects use their respective recorded snapshots. The dispersion in Figure 3c is the sample standard deviation of fusion coefficients across ten subjects.

The full trajectories show that noise and mosaic coefficients generally rise above the 20% equal initialization, while the magnitude and evolution of each condition's coefficient vary across subjects. The coefficient ordering for sub-04 differs from most subjects, consistent with the parameter allocation in Figure 3b. The epoch-resolved trajectories show how the common allocation pattern and subject differences emerge during training.

B Additional Evaluation On THINGS-MEG

B.1 Inputs And Model Adaptation

To examine the full framework in another neural recording modality, we evaluate within-subject and leave-one-subject-out image retrieval on the four participants in THINGS-MEG (Hebart et al., 2023). We read preprocessed MEG recordings and average repeated trials, using all stored sensors when no channel subset is specified. NeuroWeft is instantiated with the sensor and time dimensions of the MEG arrays rather than the EEG input dimensions.

The visual branch uses precomputed 1024-dimensional CLIP RN50 features for original, Gaussian blur, Gaussian noise, low resolution, and mosaic conditions, combined using five trainable fusion parameters. Separate linear projectors map visual and MEG representations into a 512-dimensional shared space. The neural encoder, fusion parameters, and projectors are trained jointly from scratch, while the visual backbone remains frozen.

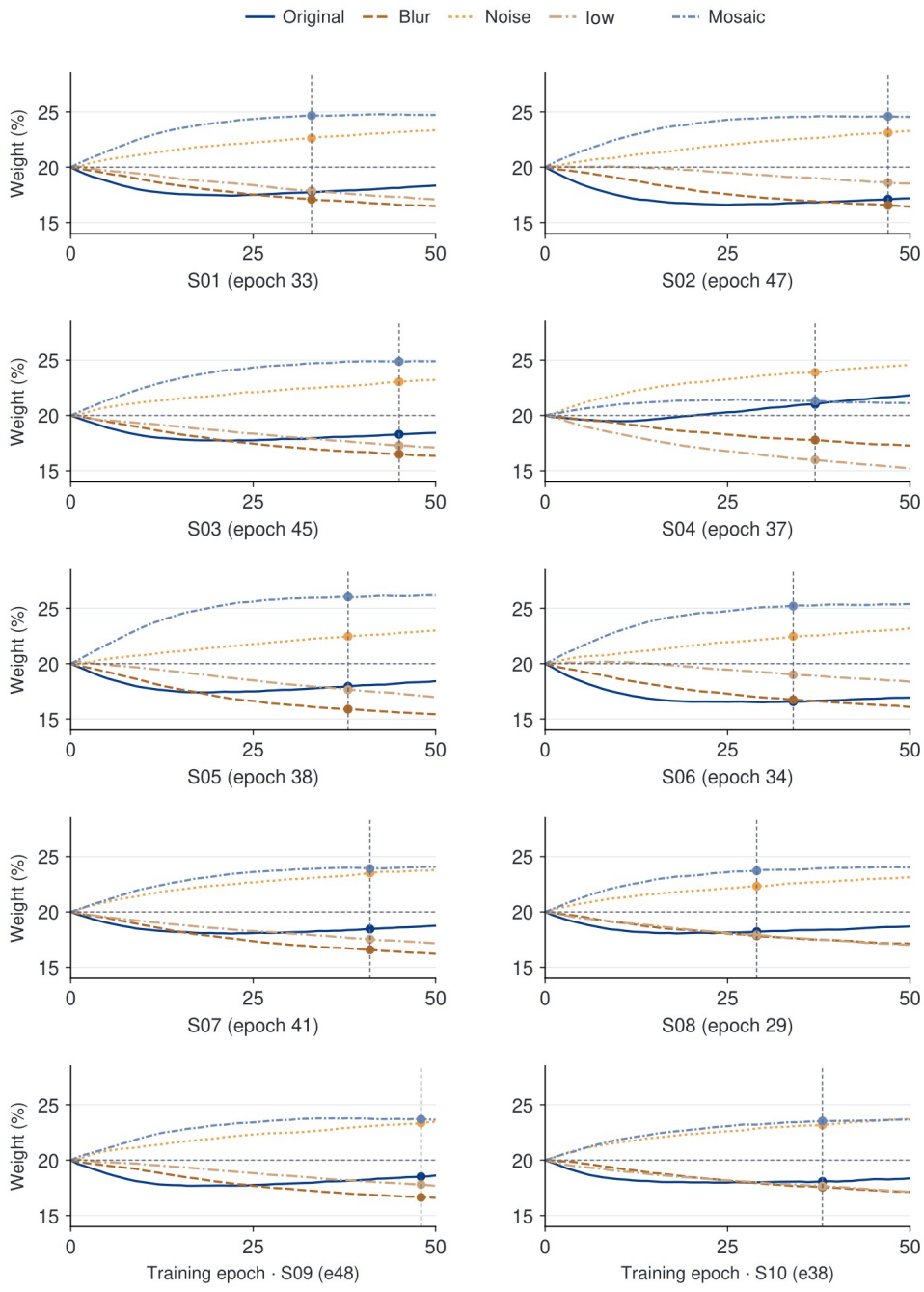

Figure 7: Five-condition fusion trajectories from the seed-2026 training run per subject, covering all ten subjects. Colors and line styles are identical across panels. Horizontal lines indicate equal weighting, and markers identify the reported snapshot for each subject. The shared trend toward larger noise and mosaic coefficients coexists with variation in the magnitude and trajectory of each condition's contribution.

Table 3: MEG image retrieval accuracy (%) for four participants under within-subject and leave-one-subject-out evaluation. Reference results follow the extended version of NeuroBridge (Zhang et al., 2026b); CogCapPro uses the multimodal-fusion results of Zhang et al. (2026a). Averages are displayed to one decimal place.

Method		Sub 1	Sub 2	Sub 3	Sub 4	Average
Intra-Subject: train and test on one subject						
NICE	Top-1	9.6	18.5	14.2	9.0	12.8
	Top-5	27.8	47.8	41.6	26.6	36.0
UBP	Top-1	15.0	46.0	27.3	18.5	26.7
	Top-5	38.0	80.5	59.0	43.5	55.2
CogCapPro	Top-1	15.0	54.5	36.3	21.5	31.8
	Top-5	**45.0**	89.5	72.2	**51.7**	64.6
NeuroBridge	Top-1	16.5	53.7	40.4	18.1	32.2
	Top-5	41.6	85.3	73.2	43.1	60.8
Shadow Brain (ours)	Top-1	**21.3**	**69.4**	**52.2**	**26.7**	**42.4**
	Top-5	44.7	**94.2**	**83.8**	49.3	**68.0**
Inter-Subject: leave one subject out for test						
UBP	Top-1	2.0	1.5	2.7	2.5	2.2
	Top-5	5.7	**17.2**	10.5	8.0	10.4
NeuroBridge	Top-1	**4.3**	3.6	3.0	2.5	3.4
	Top-5	**13.1**	15.6	11.2	11.3	12.8
Shadow Brain (ours)	Top-1	3.5	**7.3**	**6.2**	**5.6**	**5.7**
	Top-5	9.5	15.1	**14.6**	**13.2**	**13.1**

B.2 Training And Evaluation Settings

Within-subject evaluation trains an independent model for each participant. Cross-subject evaluation comprises four folds, each training on three participants and testing on the held-out participant. Each fold learns one shared fusion vector from the training participants and applies it directly to the held-out participant at test time.

Both protocols train for 50 epochs with training batch size 64 and test batch size four. We use AdamW with learning rate and weight decay 10^{-4} and initial contrastive temperature 0.07. Unlike the EEG configuration, the MEG experiments do not use random signal smoothing. Each query is ranked against the complete set of 200 candidate images by cosine similarity, and Top-1 and Top-5 accuracies are averaged over five independent training runs.

B.3 Retrieval Results

In within-subject THINGS-MEG retrieval, Shadow Brain achieves mean Top-1 and Top-5 accuracies of 42.4% and 68.0%, respectively, the highest among the methods in Table 3. Mean Top-1 exceeds NeuroBridge's 32.2% by 10.2 percentage points, and mean Top-5 exceeds CogCapPro's 64.6% by 3.4 points. Shadow Brain achieves the highest Top-1 for all four participants and the highest Top-5 for participants 2 and 3.

In leave-one-subject-out THINGS-MEG retrieval, Shadow Brain achieves mean Top-1 and Top-5 accuracies of 5.7% and 13.1%, both above the averages of the other methods in Table 3. Retraining the full framework on MEG supports within-subject matching and image retrieval for held-out participants, extending the main evaluation to another neural recording modality.

C Representation Visualization Across All Subjects

The category visualization across all ten subjects contains 1,170 records per modality: 350 animals, 470 foods, and 350 tools (Figure 8). Image features exhibit category-related organization, with some points tightly overlapping. EEG features form multiple local groups, some dominated by a single category and others containing mixed categories.

Compared with the sub-07 view, the pooled visualization retains category-related neighborhoods but has more dispersed local groups and greater overlap. We use t-SNE to display local organization of high-dimensional representations in a low-dimensional space (van der Maaten & Hinton, 2008). Distributed cortical semantic organization provides a research context for category-representation

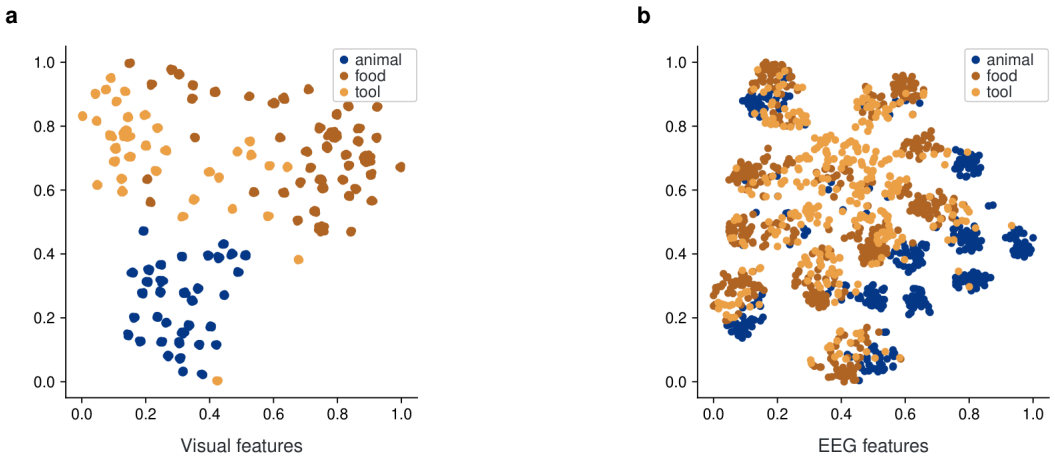

Figure 8: All-subject image and EEG feature visualizations for three categories. Colors are consistent with Figure 4. Each modality contains 1,170 plotted records aggregated across subjects.

analysis (Huth et al., 2012), while retrieval accuracies and candidate rankings describe image-level matching performance.

D Signal Analysis Within NeuroWeft

D.1 Paired Waveform, Spectrum, And Time–Frequency Structure

We analyze signals before and after processing in sub-07 through waveform, spectral, and time–frequency views (Figure 9). The Oz waveforms nearly overlap, while the residual, displayed on a separate vertical scale, shows small structured changes. The spectrum aggregated over 200 queries retains its main peaks. Mean power increases in all five bands, with the largest relative increase in gamma.

Within the displayed frequency range, the time–frequency power ratio increases more at higher frequencies, consistent with the spectral and band-power analyses. Intermediate processing in NeuroWeft preserves the main shape of the example waveform while adjusting power in a frequency-dependent manner. Figures 5 and 9 use signals from the same intermediate processing step, with statistics computed in spatially whitened model-input coordinates.

D.2 Frequency- And Channel-Resolved Changes

Mean relative PSD change is positive across the displayed frequency range and generally larger at higher frequencies (Figure 10a). Resolving the changes by channel and band shows positive gamma-band changes in most channels, while lower bands show increases or decreases depending on the channel.

In the channel summaries over 1–60 Hz, POz has the largest mean relative PSD increase and PO8 has a negative change (Figure 10d). Frequency-averaged relative PSD changes and the overall RMS changes in Figure 5 characterize spectral and amplitude changes, respectively. Resolving the statistics by frequency and channel therefore provides complementary descriptions of the signal transformation.

E Implementation And Statistical Details

E.1 EEG Inputs And Data Loading

Within-subject EEG inputs select 17 posterior electrodes in the following order: P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, and O2. Each sample contains 250 time points.

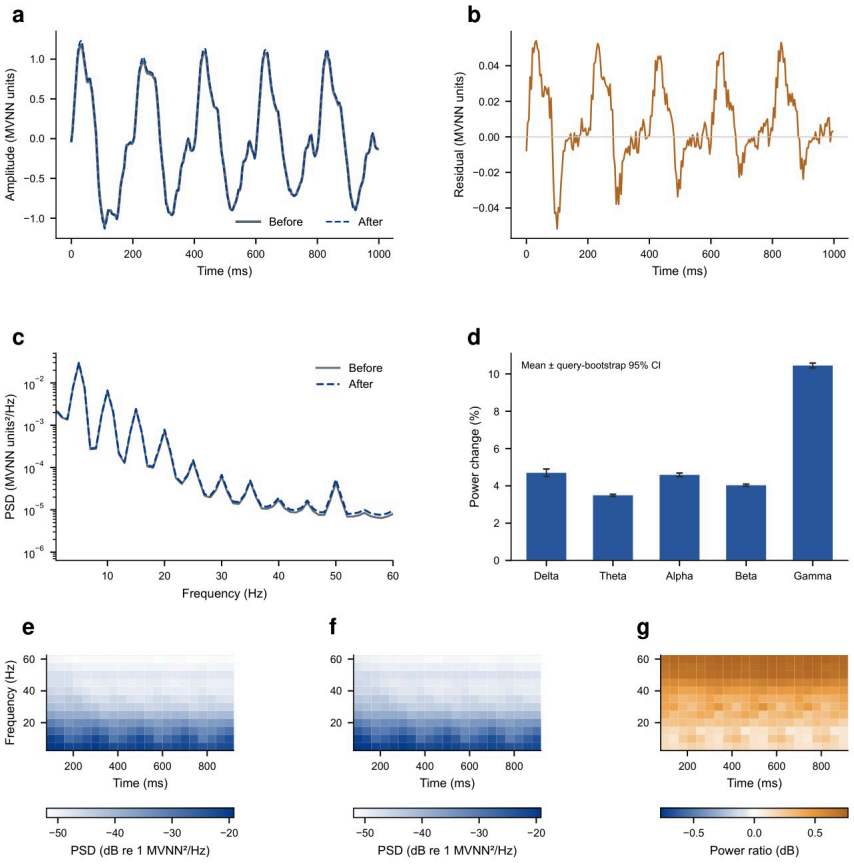

Figure 9: Paired signal and spectral diagnostics for sub-07. **a,b** An Oz waveform example and its residual at separate vertical scales. **c** Spectrum across 200 queries. **d** Paired band-power changes with query-bootstrap 95% confidence intervals. **e–g** Before, after, and relative time–frequency views. Signals are measured in MVNN units at the same intermediate step as Figure 5. Confidence intervals are obtained by resampling the 200 paired queries.

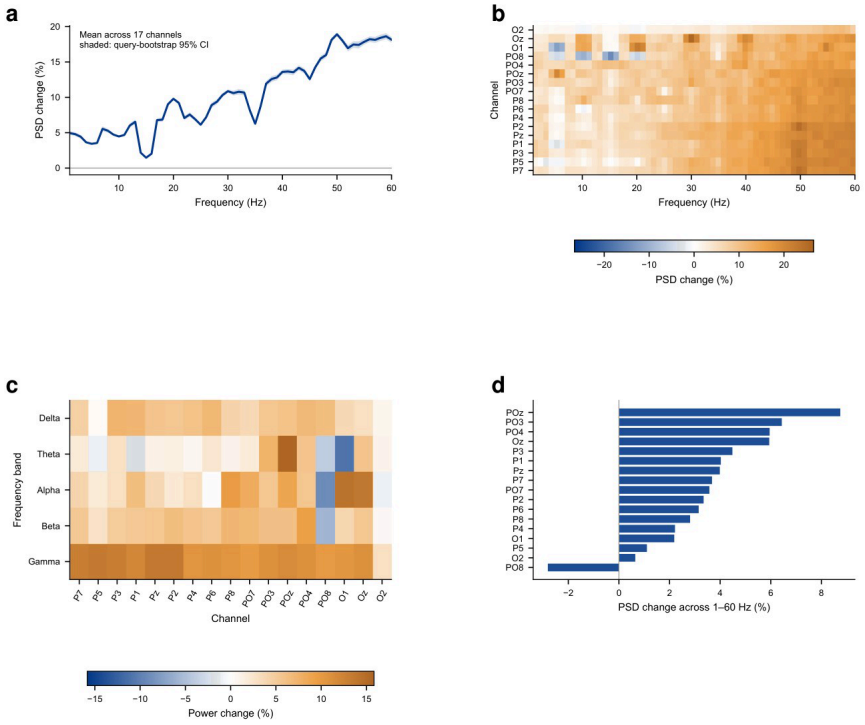

Figure 10: Frequency- and channel-resolved modulation for sub-07. **a** Mean relative PSD change over 17 channels with a query-bootstrap 95% interval. **b** Frequency-by-channel changes. **c** Band-by-channel changes. **d** Channel means across the displayed 1–60 Hz range for 200 paired queries. The frequency- and channel-resolved summaries characterize changes in model-input signal coordinates; their frequency averaging differs from the RMS summary in Figure 5.

Before channel selection and repetition averaging, training arrays have dimensions $1654 \times 10 \times 4 \times 63 \times 250$, and test arrays have dimensions $200 \times 1 \times 80 \times 63 \times 250$, ordered by concept, image, repetition, electrode, and time.

Training uses repetition averaging and applies random smoothing with kernel size five and probability 0.3; the test loader does not apply random smoothing. Training and test batch sizes are 64 and four, respectively, and both loaders use four workers. Training samples are shuffled, and incomplete final batches are excluded from batch updates. Each query is always compared with all 200 candidate images, so a test batch size of four does not change the number of retrieval candidates.

E.2 Features, Projection, And Optimization

NeuroWeft outputs and features from each visual condition have 1024 dimensions. Two independent linear projectors map EEG and image features to 512 dimensions. AdamW jointly updates the neural encoder, projectors, and five visual fusion parameters with learning rate and weight decay 10^{-4}; the pretrained visual backbone remains fixed. The loss configuration uses initial contrastive temperature 0.07, enables image-feature ℓ_2 normalization, and disables EEG-feature ℓ_2 normalization.

E.3 Cross-Subject EEG Training

Each of the ten leave-one-subject-out folds initializes a new NeuroWeft encoder, two projectors, visual fusion module, and AdamW optimizer. Training data come from the other nine participants, and test data come only from the held-out participant. Both splits use the same ordered channel selection. Each fold trains for 50 epochs with training batch size 64, test batch size four, learning rate and weight decay 10^{-4}, and initial contrastive temperature 0.07. Random smoothing is applied

only to training inputs. All training participants in a fold share one fusion vector, which is used directly for retrieval on the held-out participant.

E.4 Independent Runs And Statistical Summaries

Shadow Brain retrieval experiments use five independent training runs with different random seeds. Except for Figure 3a, reported retrieval accuracies are first averaged arithmetically over the five runs within each subject or held-out fold, and then averaged across subjects. Table 1 presents subject-level accuracies and the overall average, while Table 3 presents the corresponding MEG results.

The visual-condition comparison in Figure 3a comes from a single training analysis with seed 2026. Fusion parameters and epoch-resolved trajectories in Figure 3b–d and Figure 7 also come from the corresponding seed-2026 training record for each subject; snapshot epochs are listed in Appendix A.2. The mean and sample standard deviation in Figure 3c describe fusion-parameter differences across ten subjects. Intervals in the signal analyses of Figure 5c–g and Figures 9 and 10 are obtained by bootstrap resampling of 200 paired queries and describe query-level statistical uncertainty. Retrieval results from five independent runs are aggregated as specified above.