CORTIVA: CANDIDATE-SCORE FUSION OF COMPLEMENTARY VISUAL TEACHERS FOR EEG- AND MEG-TO-IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding visual experience from non-invasive brain activity is central to neuroscience and brain–computer interfaces. Functional magnetic resonance imaging (fMRI) offers fine spatial detail, but its slow hemodynamics and burdensome acquisition limit temporally resolved decoding. Electroencephalography (EEG) and magnetoencephalography (MEG) provide millisecond resolution, making image retrieval compelling: identify the viewed image from neural responses and a fixed candidate bank. Contrastive alignment to pretrained visual representations enables zero-shot retrieval from EEG and MEG, but most systems collapse heterogeneous visual supervision into a single embedding before ranking. This early consolidation imposes one similarity geometry on every candidate order and removes encoder-specific disagreements from the final ranking. We propose CORTIVA, a candidate-score fusion method that retains this complementary evidence. Three decoding routes are aligned to heterogeneous visual targets, score the same indexed candidates independently, and combine only their temperature-scaled score vectors before ranking. On the 200-way THINGS-EEG2 benchmark, CORTIVA reaches **73.5% Top-1 and 95.3% Top-5** across ten participants, exceeding the strongest reported baseline by **10.3 and 5.4 percentage points**. With a modality-specific neural encoder, the same fusion principle reaches **42.4% Top-1 on THINGS-MEG**. Matched route-removal retraining and four weight controls show that the gain comes from combining complementary route scores and persists under uniform weighting, without depending on a specialized weighting rule. Analyses in an independent DINOv2 space recover the same local error-neighborhood enrichment and posterior neural–visual correspondence. These results support candidate-score fusion as a simple, testable alternative to embedding-level consolidation for neural image retrieval.

1 INTRODUCTION

Neural decoding infers perceptual, cognitive, or motor states from brain activity and supports brain–computer interfaces. Functional magnetic resonance imaging established natural-image decoding by resolving distributed cortical representations (Kay et al., 2008; Haxby et al., 2001; Shen et al., 2019), but its slow hemodynamics and burdensome acquisition limit interactive use. EEG and MEG preserve millisecond dynamics (Thorpe et al., 1996; Cichy et al., 2014; Grootswagers et al., 2017), motivating participant-level zero-shot image retrieval on public natural-image benchmarks (Gifford et al., 2022; Song et al., 2024; Zhang et al., 2026).

Recent EEG decoders align neural responses with visual or multimodal embeddings and then rank candidate images in one space (Song et al., 2024; Du et al., 2023; Song et al., 2025; Wei et al., 2024; Zhang et al., 2025; Li et al., 2024; 2025; Wu et al., 2025). These systems differ in encoders and supervision, yet a single similarity geometry ultimately orders all candidates. Disagreement among visual targets is therefore compressed before ranking.

CORTIVA instead aligns three routes to heterogeneous visual targets and fuses only their candidate scores. The routes score the same indexed candidate bank independently, so their disagreements remain available when the final ranking is formed. The main contributions are:

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.

- a retrieval architecture that preserves inter-target ranking differences by combining **score vectors, not embeddings**;
- **matched route-removal retraining and weight controls** that separate score integration from weight estimation; and
- **participant-level and cross-participant validation across EEG and MEG**, complemented by visual-geometry and neural-similarity analyses.

2 Related Work

Single-target neural image retrieval. THINGS-EEG and THINGS-EEG2 provide public, stimulus-controlled natural-image benchmarks (Grootswagers et al., 2022; Gifford et al., 2022). Early decoders classified stimuli within closed label sets using compact networks such as EEGNet (Lawhern et al., 2018) and deeper spatiotemporal convolutional architectures (Schirrmeister et al., 2017). Contrastive methods reframed the task as retrieval by mapping neural responses into a frozen visual embedding space. NICE aligns EEG with CLIP image embeddings (Song et al., 2024; Radford et al., 2021), and UBP models uncertainty in that alignment (Wu et al., 2025). ATM and Neural-MCRL introduce adaptive and spectral–temporal encoders (Li et al., 2024; 2025), while NeuroBridge combines self-supervised cognitive priors with bidirectional semantic alignment (Zhang et al., 2026). These systems differ substantially in neural encoding, but one similarity function over one embedding space determines the final candidate ordering.

Structured and multi-source visual targets. BraVL combines visual and linguistic features (Du et al., 2023); NICE++ adds language-guided supervision (Song et al., 2025); MB2C introduces bidirectional cycle consistency (Wei et al., 2024); and CognitionCapturer aligns EEG with image, text, and depth modalities (Zhang et al., 2025). Candidate targets can also be constructed from supervised backbones, self-supervised representations, or encoders optimized to predict cortical responses (He et al., 2016; Dosovitskiy et al., 2021; Caron et al., 2021; He et al., 2022; Yamins et al., 2014). Recent methods enrich representations before candidate scoring through structural representations, participant-adaptive granularity, and multilevel alignment (Tang et al., 2026; Jiang et al., 2026; Liu et al., 2026). In each case, the resulting composite representation still defines one ranking geometry. We discuss generative reconstruction separately because its perceptual and semantic metrics assess synthesized images rather than rankings within a fixed candidate set (Takagi & Nishimoto, 2023; Scotti et al., 2024; Benchetrit et al., 2024).

Score-level fusion and multi-teacher distillation. Classical classifier-fusion analysis shows that the sum rule can be insensitive to small route-specific posterior errors (Kittler et al., 1998), and late score fusion is effective in multimodal semantic recognition (Snoek et al., 2005). Gated mixtures of experts instead learn query-dependent weights (Jacobs et al., 1991). CORTIVA applies bidirectional contrastive learning, geometry preservation, and candidate-distribution distillation within each route (van den Oord et al., 2018; Radford et al., 2021; You et al., 2017; Hinton et al., 2015), but postpones integration until route-specific score vectors are available.

3 Method

Datasets. THINGS-EEG2 provides 63-channel, 1000-Hz EEG from ten participants, and THINGS-MEG provides 271-sensor, 1200-Hz recordings from four different participants; both use images from the 1,854-concept THINGS collection (Hebart et al., 2019; Gifford et al., 2022; Hebart et al., 2023). EEG trials are baseline-corrected with the -100 to 0 ms interval, cropped to 0–500 ms, and averaged over adjacent sample pairs to 500 Hz, yielding a 63×250 tensor. The primary EEG evaluation averages one test-session mean from each of four sessions. The final 200-image test set is held out from all training, checkpoint selection, and fusion-parameter selection.

Targets and routes. Throughout, a *target* denotes the candidate embedding geometry defined by frozen visual features and fixed transforms; a *route* denotes the neural mapping trained against one target; and a *score vector* contains that route's similarities to the identically indexed candidates. SAR applies a trainable projection to CLIP-RN50 features (Radford et al., 2021). CVR combines

OpenCLIP ViT-B/32, SynCLR, and SDXL-VAE features through fixed source-specific projectors and a fixed composite-target model (Ilharco et al., 2021; Cherti et al., 2023; Tian et al., 2024; Kingma & Welling, 2014; Podell et al., 2024). EPR averages three fixed CLIP-RN50 image-projection checkpoints. At inference, the three routes score the same 200-image bank. The complete data flow is summarized in Fig. 1.

For participant s, query i, route $b \in \{\mathrm{sar}, \mathrm{cvr}, \mathrm{epr}\}$, route representation $z_{s,i}^{(b)}$, and candidate representation $v_j^{(b)}$, each route produces

$$S_{sij}^{(b)} = \tau_b \left\langle \frac{z_{s,i}^{(b)}}{\|z_{s,i}^{(b)}\|_2}, \frac{v_j^{(b)}}{\|v_j^{(b)}\|_2} \right\rangle, \tag{1}$$

where $\tau_b = \min[\exp(\eta_b), 100]$ is a learned inverse temperature. A participant-specific router produces query-dependent route weights and may be blended with a score-margin confidence vector selected on validation data:

$$S_{sij}^{\mathrm{fused}} = \sum_b w_{s,i}^{(b)} S_{sij}^{(b)}. \tag{2}$$

Candidates are sorted by descending fused score. The objective combines route-specific bidirectional contrastive losses and distillation terms with cross-entropy on the fused score matrix and router supervision. Loss weights, route definitions, and validation grids were specified before test evaluation; Appendix B reports the architecture, objective, and selection procedure.

Neural encoders and selection. The shared EEG encoder combines a participant-specific spatial transform, graph attention over a dense directed channel graph, channel-token attention, coordinate-conditioned gating, and a temporal-convolutional projection. SAR uses this projection, CVR applies a residual head, and EPR uses an independent spectral–temporal route. Participant-specific single-target models are trained first; their pairs selected on validation data initialize CORTIVA, after which the target/reference networks are frozen and the route and fusion modules are jointly optimized. Checkpoints are selected by validation loss, and fusion parameters by validation mean rank on concept-disjoint data.

The objective combines route-specific contrastive and distillation terms with cross-entropy for the fused score matrix and router supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SAR}} + \mathcal{L}_{\mathrm{CVR}} + \mathcal{L}_{\mathrm{EPR}} + 0.05\mathcal{L}_{\mathrm{rce}} + 0.04\mathcal{L}_{\mathrm{trace}}. \tag{3}$$

The three route losses retain matched-pair contrast, feature direction, within-batch relational geometry, and candidate-score distributions. We optimize encoder-side parameters at 2×10^{-5} and the heads and router at 2×10^{-4} using AdamW with weight decay 0.01, gradient clipping at 1.0, and batches of 384. Training is limited to 60 epochs with patience 5, and the checkpoint with the lowest validation loss is restored before participant-specific fusion parameters are selected by validation mean rank over $\alpha \in \{0, 0.1, 0.2, 0.25, 0.3, 0.4\}$ and $T_c \in \{2.5, 5, 10\}$. The held-out test set is scored once after selection.

Computational complexity. The trainable Stage 2 set contains 13.8 million parameters per participant. For the Sub01 seed-2026 model on an RTX 4080 SUPER, one query takes 6.30 ms and 92.3 MiB of peak allocated GPU memory when candidate banks are kept on device, excluding offline visual feature extraction. Appendix Table S4 reports the parameter, latency, and memory measurements.

4 Experimental Protocol

The primary endpoint is participant-level 200-way retrieval on THINGS-EEG2. Each held-out image serves as one query and one candidate; candidates are ranked by descending fused score, with mid-ranks for ties. We report Top-1, Top-5, and mean rank after averaging seeds within participant. Three orthogonal negative controls test candidate identity, post-stimulus timing, and posterior spatial information by permuting candidate labels after scoring, replacing EEG with the duration-matched pre-stimulus interval, or retaining only frontal sensors. Repetition experiments

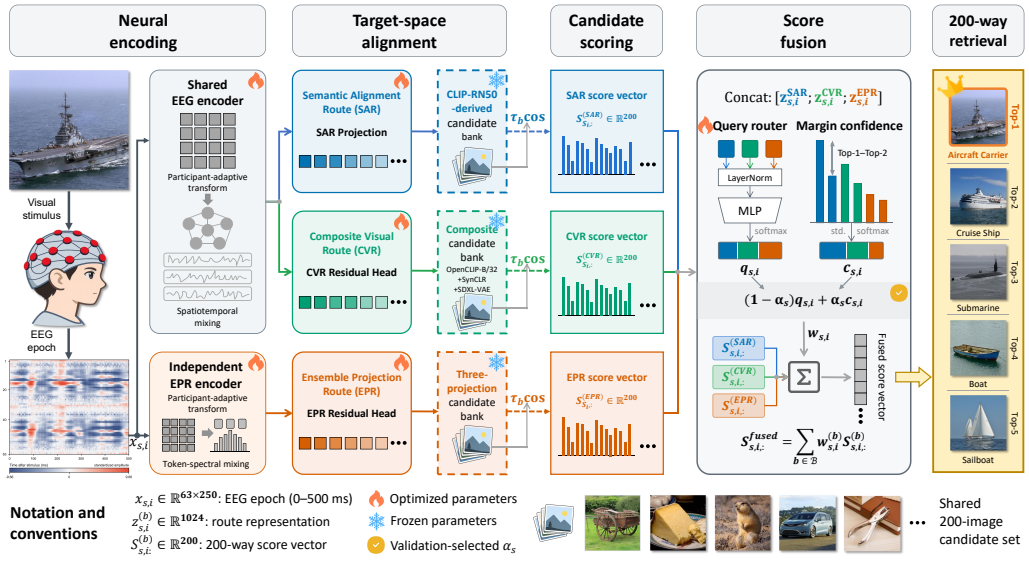

Figure 1: **CORTIVA architecture and candidate-score fusion.** The shared EEG encoder supports SAR and CVR, while EPR uses an independent encoder. Offline visual targets derive from CLIP-RN50, a composite OpenCLIP/SynCLR/SDXL-VAE representation, and a three-projection CLIP-RN50 ensemble. The three identically indexed score vectors are weighted and summed before ranking the common 200-image candidate bank; dashed outlines denote frozen candidate banks.

average every combination of one through four test-session means. Selective retrieval orders queries by the fused Top-1-minus-Top-2 margin and reports area under the risk–coverage curve.

For each structural ablation, we retrain the model after removing SAR, CVR, or EPR while retaining the remaining optimization and selection procedure. Four weight controls retain the same 50 participant-by-seed route-score matrices but replace query-aligned weights with within-participant row permutations, participant means, uniform weights, or the fixed prior. Cross-participant experiments initialize each target participant from the best-validation source among the other nine participants and adapt on concept-disjoint target data. The THINGS-MEG configuration uses modality-specific EEGProject and TSConv neural front ends, 512-dimensional route spaces, the full 0–1000 ms epoch, and the same target families and candidate-score fusion rule.

Participant is the unit of inference. Seeds are averaged within participant; percentile-bootstrap intervals summarize endpoints and nonlinear statistics. Paired contrasts use exact sign-flip tests over all 2^{10} assignments and 50,000 participant-bootstrap resamples, with Holm adjustment across route-by-metric families. Representational similarity analysis correlates the upper triangles of participant-specific EEG and visual similarity matrices with Spearman ρ. DINOv2 ViT-B/14 defines a comparison geometry that is not used by any CORTIVA target. Appendices B and S1 report expanded definitions, hyperparameters, and all comparison sources.

5 Results

Within-participant EEG retrieval. After averaging seeds within participant, CORTIVA attains **73.5% Top-1** accuracy (SD 6.6 percentage points [pp]; 95% bootstrap CI 69.6–77.3%) and **95.3% Top-5** accuracy (SD 2.4 pp; 95% CI 93.8–96.6%) on 200-way THINGS-EEG2 retrieval (Table 1; Fig. 2). The strongest reported baseline in the same within-participant setting is NeuroBridge at 63.2% Top-1 and 89.9% Top-5 (Zhang et al., 2026); CORTIVA improves these by **10.3 and 5.4 pp**.

Participant Top-1 ranges from 63.9% (Sub03) to 81.5% (Sub08), Top-5 from 89.8% (Sub05) to 97.6% (Sub10), and mean rank from 1.49 to 2.75 (group mean 1.84). The median rank is 1 for every participant. Performance is stable across seeds at $73.53 \pm 0.43\%$ Top-1, whereas the between-participant SD is 6.6 pp. Validation loss falls by 7.1% on average when training extends beyond 18 epochs; the matched analyses therefore use the 60-epoch schedule.

Table 1: Participant-level 200-way retrieval on THINGS-EEG2.

Method	Metric	Sub01	Sub02	Sub03	Sub04	Sub05	Sub06	Sub07	Sub08	Sub09	Sub10	Mean
Within-participant: train and test on one participant												
BraVL (Du et al., 2023)	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-GA (Song et al., 2024)	Top-1	15.2	13.9	14.7	17.6	9.0	16.4	14.9	20.3	14.1	19.6	15.6
	Top-5	40.1	40.1	42.7	48.9	29.7	44.4	43.1	52.1	39.7	46.7	42.8
NICE++ w/ GA (Song et al., 2025)	Top-1	16.6	20.2	19.6	25.1	13.5	16.5	17.8	26.2	18.2	22.9	19.7
	Top-5	46.9	47.6	52.6	61.0	42.3	48.2	51.6	62.1	48.7	53.5	51.5
MB2C (Wei et al., 2024)	Top-1	23.7	22.7	26.3	34.8	21.3	31.0	25.0	39.0	27.5	33.2	28.5
	Top-5	56.3	50.5	60.2	67.0	53.0	62.3	54.8	69.3	59.3	70.8	60.4
Neural-MCRL (Li et al., 2025)	Top-1	27.5	28.5	37.0	35.0	22.5	31.5	31.5	42.0	30.5	37.5	32.3
	Top-5	64.0	61.5	69.0	66.0	51.5	61.0	62.5	74.5	59.5	71.0	64.2
CognitionCapturer (Zhang et al., 2025)	Top-1	27.2	28.7	37.2	37.7	21.8	31.6	32.8	47.6	33.4	35.1	33.3
	Top-5	59.5	57.0	66.1	63.2	47.8	58.1	59.6	73.5	57.6	63.6	60.6
UBP (Wu et al., 2025)	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
NeuroBridge (Zhang et al., 2026)	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
CORTIVA (Ours)	**Top-1**	**79.1**	**67.2**	**63.9**	**69.2**	**65.6**	**81.3**	**77.7**	**81.5**	**75.0**	**74.8**	**73.5**
	Top-5	**95.5**	**95.0**	**93.9**	**93.9**	**89.8**	**97.3**	**97.3**	**97.0**	**95.8**	**97.6**	**95.3**
Cross-participant: leave one participant out for testing												
BraVL (Du et al., 2023)	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 (Song et al., 2024)	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
ATM-S (Li et al., 2024)	Top-1	10.5	7.1	11.9	14.7	7.0	11.1	16.1	15.0	4.9	20.5	11.84
	Top-5	26.8	24.8	33.8	39.4	23.9	35.8	43.5	40.3	22.7	46.5	33.73
UBP (Wu et al., 2025)	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
Neural-MCRL (Li et al., 2025)	Top-1	13.0	12.0	14.5	12.5	11.5	13.5	14.0	18.5	13.5	17.0	14.0
	Top-5	31.5	30.5	35.5	35.5	29.0	35.5	36.0	38.5	32.5	39.0	34.3
NeuroBridge (Zhang et al., 2026)	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
CORTIVA (Ours)	**Top-1**	**27.0**	**26.0**	**28.2**	**25.6**	**27.9**	**29.0**	**19.5**	**32.7**	**19.2**	**29.5**	**26.5**
	Top-5	**57.5**	**57.3**	**58.2**	**64.1**	**59.5**	**67.7**	**51.6**	**65.5**	**48.5**	**64.7**	**59.5**

Values are participant means on the 200-way test. CORTIVA reports the five-seed average; analytical chance is 0.5%/2.5% for Top-1/Top-5. Supplementary Table S1 summarizes comparison sources and evaluation settings.

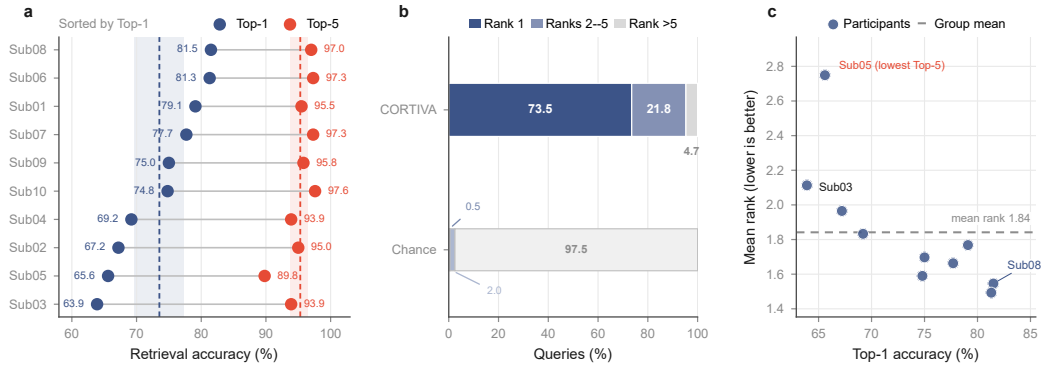

Figure 2: **Participant-level EEG retrieval and rank concentration.** (a) Five-seed mean Top-1 and Top-5 accuracy for Sub01–Sub10, ordered by Top-1 accuracy; dashed lines mark group means and shaded bands show their 95% participant-bootstrap intervals. (b) Distribution of the 10,000 participant–seed–query ranks partitioned into rank 1, ranks 2–5, and ranks above 5. (c) Participant-level relation between Top-1 accuracy and mean rank; lower mean rank indicates tighter concentration near the target.

Score integration, not a single route. CORTIVA exceeds each single-route output in all ten participants: participant means are 9.7 pp above SAR, 16.0 pp above CVR, and 10.9 pp above EPR. Removing CVR and EPR costs 5.47 and 5.40 pp Top-1, respectively, after matched retraining, even though their standalone Top-1 accuracies differ by 5.1 pp (Fig. 3). Across 10,000 participant–seed–query outcomes, fusion ranks the target first in 1,460 cases where EPR does not, while EPR does so in 370 cases where fusion does not. All four route-weight control intervals span zero around the validation-fixed query-aligned rule; uniform summation gives the highest point estimate, 74.2%

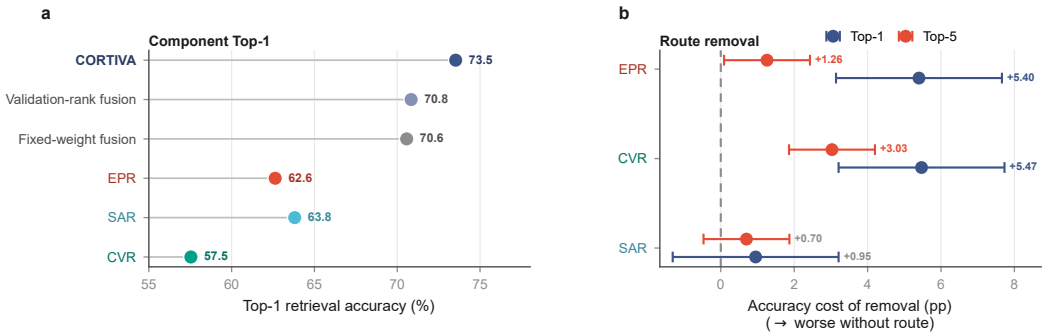

Figure 3: **Matched structural ablations.** (a) Participant-mean Top-1 retrieval for the three constituent route outputs, two fusion variants, and full CORTIVA. (b) Participant-paired Top-1 and Top-5 accuracy losses after matched retraining without each route; points show full-minus-ablated means, and horizontal bars show simultaneous 95% participant-bootstrap intervals within each metric. Positive values denote lower accuracy after removal.

versus 73.5% for the selected query-aligned rule (Appendix Fig. 7). These controls indicate that the gain comes from retaining and summing route scores, not from a finely tuned weighting rule.

Cross-participant and MEG transfer. Under target-adapted leave-one-participant-out evaluation, CORTIVA reaches **26.5% Top-1 and 59.5% Top-5** on THINGS-EEG2, compared with 19.0% and 45.9% for NeuroBridge. With a modality-specific MEG neural encoder and the same target construction and fusion principle, CORTIVA reaches **42.4% Top-1 and 73.6% Top-5** over four THINGS-MEG participants.

MEG participant Top-1 ranges from 20.6% to 65.6%, while the seed-wise SD of the four-participant mean is 1.5 pp. In a frozen-checkpoint input analysis, Top-1 rises from 29.2% with 0–500 ms retained, through 38.5% with 0–700 ms, to 42.4% with the full 0–1000 ms epoch. The cumulative profile indicates that useful MEG information is distributed across the full one-second response rather than confined to an early interval.

The complete participant-level comparison is given in Table 2.

Table 2: Within-participant 200-way retrieval on THINGS-MEG.

Method	Metric	Sub01	Sub02	Sub03	Sub04	Mean
Within-participant: train and test on one participant						
NICE (Song et al., 2024)	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 (Wu et al., 2025)	Top-1	15.0	46.0	27.3	18.5	26.7
	Top-5	38.0	80.5	59.0	43.5	55.2
NeuroBridge (Zhang et al., 2026)	Top-1	16.5	53.7	40.4	18.1	32.2
	Top-5	41.6	85.3	73.2	43.1	60.8
CORTIVA (Ours)	**Top-1**	**20.6**	**65.6**	**52.6**	**30.9**	**42.4**
	Top-5	**50.8**	**91.3**	**86.5**	**66.1**	**73.6**

Sub01–Sub04 denote the MEG cohort. CORTIVA averages seeds 2026–2031; Supplementary Table S1 summarizes the comparison settings.

Ablations, controls, and neural correspondence. Matched 60-epoch ablations quantify route-specific effects inside the fused system: removing CVR or EPR produces clear Top-1 losses, whereas the smaller SAR effect is less certain (Fig. 3). Negative controls return retrieval to analytical chance: candidate-label permutation yields 0.50% Top-1 and 2.50% Top-5, pre-stimulus input yields 0.60% and 2.41%, and frontal-only input yields 0.54% and 2.41%. In an independent DINOv2 ViT-B/14 space not used to construct any target, 21.9% of errors lie among the target's Top-10 neighbors versus

5.0% at random. EEG–visual representational-similarity analyses show stronger posterior than frontal correspondence, with the posterior-minus-frontal contrast positive in all ten participants.

Repeated observations and input sensitivity. Both retrieval accuracy and confidence ordering improve as additional observations are averaged. From one through four observations, Top-1 rises from 62.7% through 69.3% and 71.8% to 73.5%, while Top-5 increases from 90.1% to 95.3%; over the same range, AURC decreases from 0.166 to 0.098. At the selected checkpoint, retaining only O1, Oz, and O2 yields 29.5% Top-1. Gaussian noise at 0.05 of the training-set channel SD lowers Top-1 by 1.9 pp. Figure 4 shows the complete repetition-efficiency profile; Appendix Sections S10 and S16 report the sensor, temporal, perturbation, and repetition details.

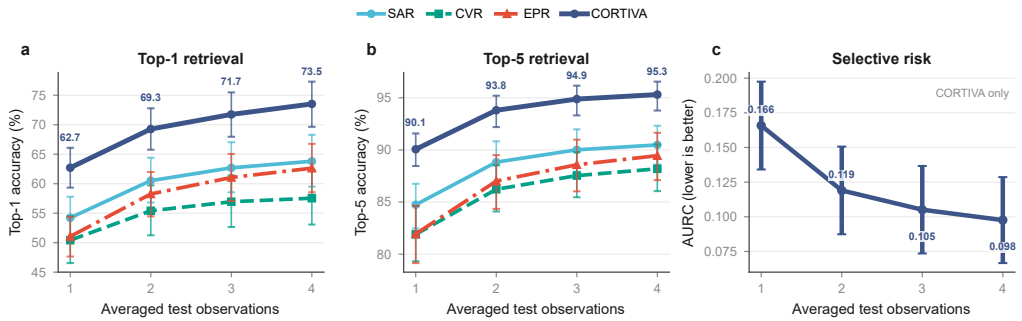

Figure 4: **Retrieval efficiency across available test observations. a,b**, Top-1 and Top-5 retrieval for CORTIVA and its three constituent route outputs after averaging every combination of one to four observations. Markers show participant means, whiskers show 95% participant-bootstrap intervals, and labels report CORTIVA means. **c**, CORTIVA AURC under the same observation counts; labels give the four AURC estimates.

External visual geometry. In DINOv2, 21.9% of incorrect Top-1 predictions lie among the target's Top-10 neighbors versus 5.0% at random (4.36×; 10/10 participants above random; $p = 0.00098$). In SAR's CLIP-RN50 geometry, the corresponding rate is 25.9% (5.15× enrichment). The corresponding sensor-wise temporal structure is shown in Fig. 6; the error-neighborhood structure is shown in Fig. 5. EEG–DINOv2 RSA shows the same posterior organization as the training-space analysis: mean $\rho = 0.0316$ over posterior sensors and a posterior-minus-frontal contrast of $\Delta\rho = 0.0108$ (10/10 positive, $p = 0.00098$).

6 Discussion

Standalone accuracy versus removal cost. CVR illustrates why route contribution must be assessed inside the fused system. Its standalone output reaches 57.5% Top-1, yet removing it costs 5.47 pp, essentially matching the 5.40-pp cost of removing the 62.6% EPR route. Fusion also recovers 166 outcomes missed at rank one by every constituent route. Full-score complementarity, including useful evidence below rank one, is therefore a more informative basis for route retention than standalone accuracy.

Adaptive weighting and score integration. Partly nonoverlapping route successes and near-uniform learned weights (mean normalized entropy 0.938) accord with sum-rule analysis (Kittler et al., 1998). Across the four Top-1 controls, intervals span zero and uniform summation gives the highest point estimate. The controls point to score integration as a main source of the gain and suggest uniform summation as a parameter-free alternative.

Modality and participant generalization. The score-space formulation extends across modalities and sensor layouts. For EEG, the 6.6-pp between-participant SD is approximately 15 times the 0.43-pp between-seed SD of the group mean. MEG Top-1 spans 45.0 pp, compared with a 17.6-pp EEG range; the seed-wise SD of the MEG group mean is 1.5 pp. Participant-to-participant variation is therefore much larger than optimization-seed variation in both cohorts.

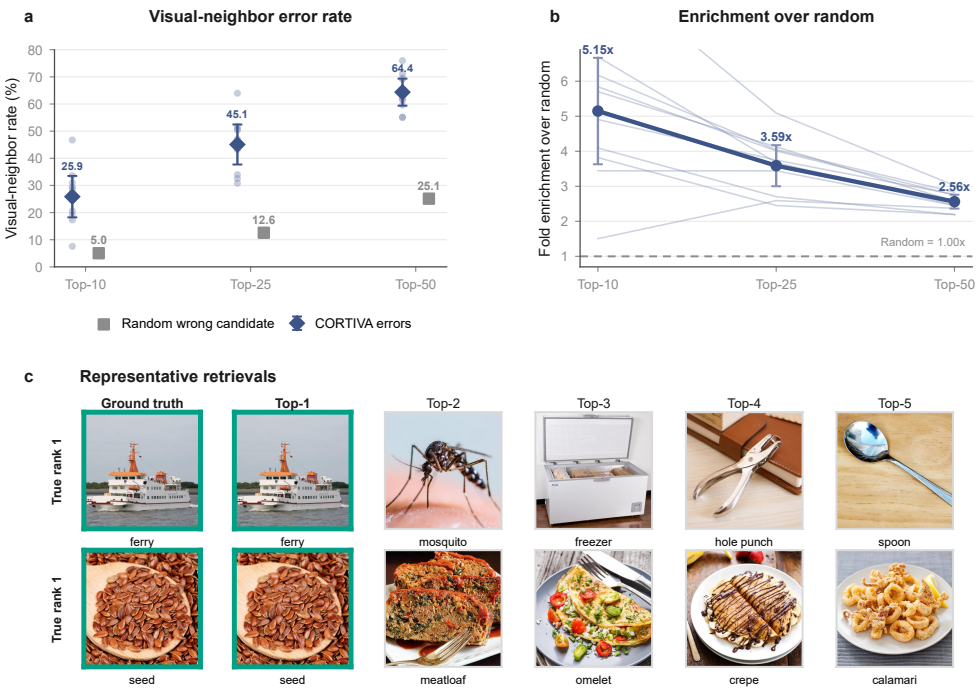

Figure 5: **CLIP-neighbor structure of retrieval errors and real-image examples.** (a) Participant-level rates at which incorrect CORTIVA predictions fall within the true image's Top-10, Top-25, or Top-50 CLIP neighborhood; squares mark random-wrong-candidate references. (b) Fold enrichment over the corresponding random references, computed within participant before group averaging. (c) Two Top-1 successes from Sub10 (seed 2026), sampled without reference to score or difficulty. Each row shows the ground truth followed by the five highest-ranked candidates; teal borders identify the target.

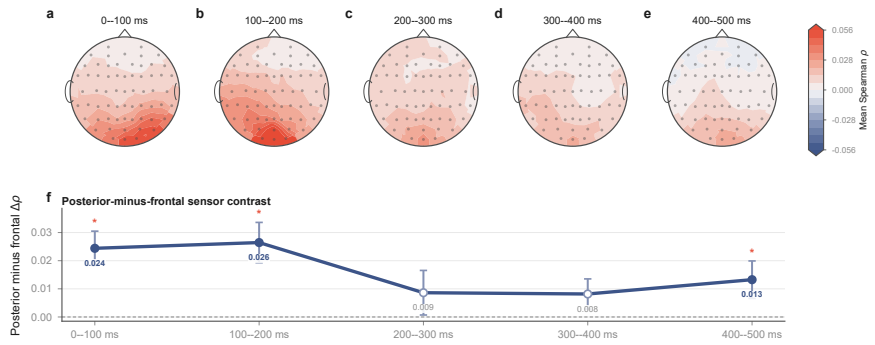

Figure 6: **Spatiotemporal EEG–CLIP representational similarity. a–e**, Mean single-sensor Spearman correlations for Sub01–Sub10 across five non-overlapping 100-ms windows on one shared asymmetric color scale. Values are interpolated for display only; statistical analyses use the recorded sensor locations. **f**, Posterior-minus-frontal contrasts with 95% t-intervals across participants; asterisks mark Holm-adjusted one-sided $p < 0.05$ over the five-window family. Regional trajectories appear in Supplementary Fig. S6.

Neural correspondence and limitations. Error-neighbor enrichment and RSA measure different properties: the former is local and target-centered, whereas RSA compares all image pairs globally. Local error-neighbor enrichment coexists with modest global EEG–visual correlations, and both patterns also appear in DINOv2, which is not used to construct any target. The posterior-minus-frontal contrast places the stronger association over posterior sensors in both visual spaces. Limitations include candidate-bank retrieval rather than open-ended generation, participant-specific training, and

evaluation on public THINGS-family datasets. Broader clinical and online BCI settings remain future work.

7 Conclusion

In this article, we introduced CORTIVA, a candidate-score fusion framework that preserves route-specific rankings from three heterogeneous visual teachers until the final decision. On 200-way THINGS-EEG2, CORTIVA reaches **73.5% Top-1 and 95.3% Top-5**, exceeding the strongest reported baseline by 10.3 and 5.4 points; the same principle transfers to THINGS-MEG and cross-participant adaptation. Matched route-removal and weight controls show that full-score complementarity, rather than standalone accuracy or specialized weighting, drives the gain and recovers cases unresolved by every route at rank one. DINOv2 and EEG representational analyses further reveal semantically structured errors and stronger posterior neural–visual correspondence beyond the training targets. CORTIVA therefore establishes candidate-score fusion as an effective and interpretable design principle for non-invasive neural image retrieval.

Data/code. THINGS-EEG2 and THINGS-MEG are public datasets (Gifford, 2021; Gifford et al., 2026; Hebart et al., 2023); anonymous code and source data accompany review. **AI use statement.** AI was used only for text translation and polishing; the authors verified the text. **Ethics.** Source studies report approval and consent (Gifford et al., 2022; Hebart et al., 2023); no identification or clinical deployment is attempted. **Reproducibility.** The main text gives the protocol; Appendices B and S1 give details; code is supplied anonymously.

References

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: Toward real-time reconstruction of visual perception. In *International Conference on Learning Representations*, pp. 7846–7858, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fb0a4de9c14f5557eeea886e22569cd-Abstract-Conference.html.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

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, Vancouver, BC, Canada, June 2023. doi: 10.1109/CVPR52729.2023.00276.

Radoslaw M. Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, 2014. doi: 10.1038/nn.3635.

Radoslaw M. Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:27755, 2016. doi: 10.1038/srep27755.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

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.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY, USA, 1993.

Michael D. Ernst. Permutation methods: A basis for exact inference. *Statistical Science*, 19(4): 676–685, 2004. doi: 10.1214/088342304000000396.

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.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. THINGS-EEG2: A large and rich EEG dataset for modeling human visual object recognition. Neuroelectromagnetic Data Archive and Tools Resource (NeMAR), dataset NM000232, 2026. URL https://doi.org/10.82901/nemar.nm000232. Version 1.1.0; accessed July 12, 2026.

Alessandro Thomas Gifford. A large and rich EEG dataset for modeling human visual object recognition. Open Science Framework (OSF) project, 2021. URL https://osf.io/3jk45/. Accessed July 12, 2026.

Krzysztof J. Gorgolewski, Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, Satrajit S. Ghosh, Tristan Glatard, Yaroslav O. Halchenko, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3:160044, 2016. doi: 10.1038/sdata.2016.44.

Tijl Grootswagers, Susan G. Wardle, and Thomas A. Carlson. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, 29(4):677–697, 2017. doi: 10.1162/jocn_a_01068.

Tijl Grootswagers, Ivy Zhou, Amanda K. Robinson, Martin N. Hebart, and Thomas A. Carlson. Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022. doi: 10.1038/s41597-021-01102-7.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.

James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001. doi: 10.1126/science.1063736.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, 2019. doi: 10.1371/journal.pone.0223792.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531. arXiv:1503.02531.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. Zenodo, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

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. arXiv:2604.17782.

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.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Computational Biology*, 10(11):e1003915, 2014. doi: 10.1371/journal.pcbi.1003915.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Banff, AB, Canada, April 2014. URL https://arxiv.org/abs/1312.6114.

Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998. doi: 10.1109/34.667881.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis: connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

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, pp. 102822–102864, Vancouver, BC, Canada, December 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 *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, Nantes, France, June 2025. doi: 10.1109/ICME59968.2025.11210130.

Jingtao Liu, Peiliang Gong, Chuhang Zheng, Yiheng Liu, and Qi Zhu. Multi-level bidirectional biomimetic learning for EEG-based visual decoding, 2026. URL https://arxiv.org/abs/2605.04680. arXiv:2605.04680.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, New Orleans, LA, USA, May 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Eric Maris and Robert Oostenveld. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1):177–190, 2007. doi: 10.1016/j.jneumeth.2007.03.024.

Thomas E. Nichols and Andrew P. Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25, 2002. doi: 10.1002/hbm.1058.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido 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. URL https://openreview.net/forum?id=a68SUt6zFt.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Cyril R. Pernet, Stefan Appelhoff, Krzysztof J. Gorgolewski, Guillaume Flandin, Christophe Phillips, Arnaud Delorme, and Robert Oostenveld. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Scientific Data*, 6:103, 2019. doi: 10.1038/s41597-019-0104-8.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, pp. 1862–1874, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/081b08068e4733ae3e7ad019fe8d172f-Abstract-Conference.html.

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. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294, 2022. doi: 10.52202/068431-1833.

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 of *Proceedings of Machine Learning Research*, pp. 44038–44059, Vienna, Austria, July 2024. PMLR. URL https://proceedings.mlr.press/v235/scotti24a.html.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):e1006633, 2019. doi: 10.1371/journal.pcbi.1006633.

Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th ACM International Conference on Multimedia*, pp. 399–402, Singapore, November 2005. doi: 10.1145/1101149.1101236.

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*, Vienna, Austria, May 2024. URL https://openreview.net/forum?id=dhLIno8FmH.

Yonghao Song, Yijun Wang, Huiguang He, and Xiaorong Gao. Recognizing natural images from EEG with language-guided contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):15896–15910, 2025. doi: 10.1109/TNNLS.2025.3562743.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, Vancouver, BC, Canada, June 2023. doi: 10.1109/CVPR52729.2023.01389.

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. arXiv:2603.07077.

Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996. doi: 10.1038/381520a0.

Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15887–15898, Seattle, WA, USA, June 2024. doi: 10.1109/CVPR52733.2024.01504.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748. arXiv:1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

Yayun Wei, Lei Cao, Hao Li, and Yilin Dong. MB2C: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8992–9000, Melbourne, VIC, Australia, October 2024. doi: 10.1145/3664647.3681292.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision–brain gap with an uncertainty-aware blur prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2246–2257, Nashville, TN, USA, June 2025. doi: 10.1109/CVPR52734.2025.00215.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1294, Halifax, NS, Canada, August 2017. doi: 10.1145/3097983.3098135.

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, Philadelphia, PA, USA, February 2025. doi: 10.1609/aaai.v39i13.33587. Issue 13.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18028–18036, 2026. doi: 10.1609/aaai.v40i21.38863. Issue 21.

A Expanded Background and Positioning

A.1 Introduction

Neural decoding infers perceptual, cognitive, or motor states from brain activity and supports brain–computer interfaces (BCIs). Early non-invasive systems focused on predefined commands or classes (Lawhern et al., 2018; Schirrmeister et al., 2017); recent work asks a harder question: which natural image produced a neural response? In retrieval, the decoder ranks a candidate bank to identify that image.

Functional magnetic resonance imaging (fMRI) established visual decoding by resolving distributed cortical representations and enabling natural-image identification and reconstruction (Kay et al., 2008; Haxby et al., 2001; Shen et al., 2019). Its seconds-scale hemodynamics and costly, immobile acquisition, however, limit interactive BCIs. Electroencephalography (EEG) and magnetoencephalography (MEG) preserve millisecond dynamics (Thorpe et al., 1996; Cichy et al., 2014; Grootswagers et al., 2017), motivating participant-level zero-shot retrieval on public benchmarks (Gifford et al., 2022; Song et al., 2024; Zhang et al., 2026).

EEG decoding has progressed from closed-set classification to contrastive retrieval of unseen concepts (Song et al., 2024; Radford et al., 2021). Recent systems add language and multimodal targets, uncertainty modeling, and spectral–temporal encoders (Du et al., 2023; Song et al., 2025; Wei et al., 2024; Zhang et al., 2025; Li et al., 2024; 2025; Wu et al., 2025). Yet they ultimately rank candidates in one embedding space. Disagreement among visual sources is therefore compressed before ranking.

CORTIVA instead aligns three routes to heterogeneous visual targets and fuses only their temperature-scaled candidate scores. The routes score the same indexed candidate bank independently, so their disagreements remain available when the final ranking is formed. Matched route-removal retraining and four weight controls test whether the gain comes from score integration rather than a specialized weighting rule. We evaluate the resulting retrieval behavior on THINGS-EEG2 and THINGS-MEG, including participant-level, cross-participant, and visual-geometry analyses (Gifford et al., 2022; Hebart et al., 2019; 2023).

The main contributions are:

- a retrieval architecture that scores a shared candidate bank under three heterogeneous visual targets and combines only the resulting score vectors, preserving inter-target ranking differences instead of absorbing them into a single embedding;
- matched route-removal retraining and four weight controls that quantify each route's removal effect, distinguish score integration from weight estimation, and test full-score complementarity as the basis for route retention; and
- participant-level and cross-participant validation across EEG and MEG, complemented by a DINOv2 geometry outside target construction that recovers the error-neighborhood and posterior neural-similarity patterns observed in the training spaces.

A.2 Related Work

A.2.1 Single-Target Neural Image Retrieval

THINGS-EEG and THINGS-EEG2 provide public, stimulus-controlled natural-image benchmarks (Grootswagers et al., 2022; Gifford et al., 2022). Early decoders classified stimuli within closed label sets using compact networks such as EEGNet (Lawhern et al., 2018) and deeper spatiotemporal convolutional architectures (Schirrmeister et al., 2017). Contrastive methods reframed the task as retrieval by mapping neural responses into a frozen visual embedding space. NICE aligns EEG with CLIP image embeddings (Song et al., 2024; Radford et al., 2021), and UBP models uncertainty in that alignment (Wu et al., 2025). ATM and Neural-MCRL introduce adaptive and spectral–temporal encoders (Li et al., 2024; 2025), while NeuroBridge combines self-supervised cognitive priors with bidirectional semantic alignment (Zhang et al., 2026). These systems differ substantially in neural encoding, but one similarity function over one embedding space determines the final candidate ordering.

A.2.2 Structured and Multi-Source Visual Targets

BraVL combines visual and linguistic features (Du et al., 2023); NICE++ adds language-guided supervision (Song et al., 2025); MB2C introduces bidirectional cycle consistency (Wei et al., 2024); and CognitionCapturer aligns EEG with image, text, and depth modalities (Zhang et al., 2025). More broadly, candidate targets can be constructed from supervised backbones (He et al., 2016; Dosovitskiy et al., 2021), self-supervised representations (Caron et al., 2021; He et al., 2022), or encoders optimized to predict cortical responses (Yamins et al., 2014). In these designs, enrichment occurs upstream of candidate scoring, and the resulting composite representation defines one ranking geometry. Recent methods likewise enrich the representation before candidate scoring through structural representations, participant-adaptive granularity, and multilevel alignment (Tang et al., 2026; Jiang et al., 2026; Liu et al., 2026). Table 3 summarizes participant-level 200-way retrieval results and their evaluation settings. We discuss generative reconstruction separately because its perceptual and semantic metrics assess synthesized images rather than rankings within a fixed candidate set (Takagi & Nishimoto, 2023; Scotti et al., 2024; Benchetrit et al., 2024).

A.2.3 Score-Level Fusion and Multi-Teacher Distillation

Classical classifier-fusion analysis shows that, when individual posterior estimates deviate only slightly from the true posteriors, the error of the sum rule is insensitive to those deviations to first order (Kittler et al., 1998). Late score fusion has also proved effective in multimodal semantic recognition (Snoek et al., 2005). A trained combiner adds parameters that must be estimated from finite data, whereas gated mixtures of experts learn query-dependent weights (Jacobs et al., 1991). Each CORTIVA route uses a bidirectional contrastive objective over matched and mismatched image–neural pairs (van den Oord et al., 2018; Radford et al., 2021); geometry- and candidate-distribution terms connect the objective to multi-teacher distillation (You et al., 2017; Hinton et al., 2015).

B Expanded Methods

B.1 Dataset and Participants

THINGS-EEG2 provides 63-channel, 1000-Hz EEG from ten participants (Sub01–Sub10), while THINGS-MEG provides 271-sensor, 1200-Hz recordings from four different participants; both datasets use images drawn from the 1,854-concept THINGS collection (Hebart et al., 2019; Gifford et al., 2022; Hebart et al., 2023). Models are trained separately for each participant and modality. EEG and MEG participant labels are modality-specific and do not identify the same individuals. Dataset identifiers and acquisition details appear in Supplementary Table S2.

We start from the released 1000-Hz all-channel arrays. Each trial is baseline-corrected with the -100 to 0 ms interval, cropped to 0–500 ms, and averaged over adjacent sample pairs to 500 Hz, yielding a 63×250 tensor. No additional filtering or artifact rejection is applied before training. For each training image, available session-level trial means are averaged into one input. The primary evaluation averages one trial mean from each of four sessions; a separate repetition analysis scores every combination of one through four test sessions. The 1,654 released concepts are partitioned into 1,454 training and 200 validation concepts (ten images each), producing 14,540 training and 2,000 validation inputs. The held-out set contains one image from each of 200 concepts. Every held-out image serves both as a query and as a candidate, and the entire set is reserved for final evaluation.

B.2 Visual Targets and Retrieval

Throughout, a *target* denotes the candidate embedding geometry defined by frozen visual features and their fixed transforms; a *route* denotes the neural mapping trained against one target; and a *score vector* contains that route's similarities to the identically indexed candidates. SAR applies a trainable 1,024-d image projection to frozen CLIP-RN50 features (Radford et al., 2021). For CVR, OpenCLIP ViT-B/32 (512-d), SynCLR (768-d), and SDXL-VAE (1,024-d) features pass through fixed source-specific projectors and a fixed composite-target model, whose normalized output has 1,024 dimensions (Ilharco et al., 2021; Cherti et al., 2023; Schuhmann et al., 2022; Tian et al., 2024; Kingma & Welling, 2014; Podell et al., 2024). EPR applies three fixed image-projection checkpoints to CLIP-RN50 features, normalizes their outputs, and averages them. The CVR and EPR targets and

all source feature banks are precomputed; SAR candidate representations are updated through its image projector. At inference, the three routes score fixed candidate banks under the same 200-image index. Supplementary Tables S3, S8, and S11 specify the models and checkpoints.

Stage 1 performs participant-specific, single-target EEG–image alignment on the same 1,454 training and 200 validation concepts used by the final model. It produces paired EEG and image-projection checkpoints, including the source-specific CVR projectors and the three EPR projection pairs. Stage 2 initializes CORTIVA from the pairs selected on validation data, freezes the target/reference networks, and jointly optimizes the trainable route and fusion modules described below. The held-out 200-image test set is excluded from both stages.

Let $b \in \mathcal{B} = \{\mathrm{sar}, \mathrm{cvr}, \mathrm{epr}\}$ index the three routes, $z_{s,i}^{(b)}$ denote the route representation for participant s and held-out query i, and $v_j^{(b)}$ denote candidate j in the corresponding route space under the selected model. Each route produces a temperature-scaled cosine score

$$S_{sij}^{(b)} = \tau_b \left\langle \frac{z_{s,i}^{(b)}}{\|z_{s,i}^{(b)}\|_2}, \frac{v_j^{(b)}}{\|v_j^{(b)}\|_2} \right\rangle. \tag{4}$$

Here $\tau_b = \min[\exp(\eta_b), 100]$ is the learned inverse temperature (logit scale, as in CLIP (Radford et al., 2021)) for route b. The common candidate index makes score-vector positions directly comparable. Fusion preserves each learned score scale; consequently, route influence depends jointly on τ_b, the score distribution, and the fusion weight rather than on weight magnitude alone.

B.3 EEG Input and Controls

The model input is a 63×250 tensor (channels $\times$ time samples) spanning 0–500 ms post-stimulus at 500 Hz; the default four-session test average follows Section B.1.

Three orthogonal negative controls test candidate indexing, post-stimulus timing, and posterior spatial information: candidate identities are permuted after scoring; the EEG input is replaced by the duration-matched pre-stimulus interval (-500 to 0 ms, all 63 channels); or only frontal sensors are retained over 0–500 ms. Supplementary Fig. S1 and Table S13 summarize these controls, and Supplementary Table S17 reports the sensor and temporal input scan. For selective retrieval, queries are ordered by the fused Top-1-minus-Top-2 score margin; risk is the Top-1 error rate among retained queries, and lower area under the risk–coverage curve (AURC) indicates better confidence ordering. Sensor-space figures display participant-mean statistics at the recorded sensors rather than model weights.

B.4 CORTIVA Model Family

The shared graph-token EEG encoder in Fig. 1 combines graph-based and token-based channel mixing. A participant-specific $63 \to 63$ spatial transform first remaps the channel axis. The resulting tokens pass through graph attention over a dense directed channel graph without self-loops, followed by channel-token attention. A learned channel gate and an additive bias from a 40-d electrode-coordinate embedding then modulate the output. A temporal-convolutional head projects the representation to 1,024 dimensions. EPR uses a separate participant-specific spatial transform and channel-token Transformer (Vaswani et al., 2017). Its spectral module computes a channel-wise FFT, reweights the coefficients with learned per-channel and per-band gains over five masks (0.5–4, 4–8, 8–13, 13–30, and 30–45 Hz), applies an inverse FFT, and blends the result residually with the input before a separate 1,024-d temporal-convolutional projection. SAR uses the graph-token projection, CVR applies a residual head to that vector, and EPR applies an independent residual head. Supplementary Table S3 specifies the trainable modules and spectral masks.

Each CVR and EPR residual head contains two 1,024-d linear layers with GELU activation and dropout ($p = 0.1$) between them. We initialize the final layer at zero, making the residual mapping the identity at initialization. For participant s, router $g_s(\cdot)$ applies layer normalization to concatenated route vectors followed by a $3072 \to 512 \to 3$ multilayer perceptron with GELU and 0.05 dropout. Its output weights start at zero, and its bias is initialized to the element-wise logarithm of the prespecified prior $(0.55, 0.25, 0.20)$, chosen before test evaluation and held fixed across all participants, seeds,

and ablations. The router outputs a softmax over the three routes:

$$q_{s,i} = \mathrm{softmax}\Big(g_s\Big[z_{s,i}^{(\mathrm{sar})}; z_{s,i}^{(\mathrm{cvr})}; z_{s,i}^{(\mathrm{epr})}\Big]\Big). \tag{5}$$

The router is optionally blended with a score-margin confidence vector $c_{s,i}$, obtained by standardizing each route's Top-1-minus-Top-2 margin within a query (Supplementary Eq. (S1)). The participant-specific blend α_s and confidence temperature $T_{c,s}$ are chosen on validation data. The final weights and fused score are

$$\begin{aligned} w_{s,i} &= (1-\alpha_s)q_{s,i} + \alpha_s c_{s,i}, \\ S_{sij}^{\mathrm{fused}} &= \sum_{b\in\mathcal{B}} w_{s,i}^{(b)} S_{sij}^{(b)}. \end{aligned} \tag{6}$$

Equation equation 6 implements candidate-score fusion: route scores are combined over aligned candidate indices without collapsing the targets into a common embedding. Candidates are sorted by descending fused score, with ties assigned mid-ranks. For $N = 200$ final queries and candidates, the correct-image rank and participant-level Top-k accuracy are

$$\begin{aligned} r_{s,i} = 1 &+ \sum_{j\neq y_i} \mathbb{1}\big[S_{sij}^{\mathrm{fused}} > S_{siy_i}^{\mathrm{fused}}\big] \\ &+ \tfrac{1}{2}\sum_{j\neq y_i} \mathbb{1}\big[S_{sij}^{\mathrm{fused}} = S_{siy_i}^{\mathrm{fused}}\big], \\ \text{Top-}\,k_s &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[r_{s,i}\le k]. \end{aligned} \tag{7}$$

The objective combines route-specific contrastive and distillation terms with cross-entropy for the fused score matrix and router supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SAR}} + \mathcal{L}_{\mathrm{CVR}} + \mathcal{L}_{\mathrm{EPR}} + 0.05\,\mathcal{L}_{\mathrm{rce}} + 0.04\,\mathcal{L}_{\mathrm{trace}}. \tag{8}$$

Each route loss, $\mathcal{L}_{\mathrm{SAR}}$, $\mathcal{L}_{\mathrm{CVR}}$, and $\mathcal{L}_{\mathrm{EPR}}$, is a weighted sum of a bidirectional contrastive term (van den Oord et al., 2018; Radford et al., 2021) and distillation terms that retain feature direction, within-batch relational geometry, and candidate-score distributions (Hinton et al., 2015; You et al., 2017). All 15 weights were specified before test evaluation and are shared across participants, seeds, and matched ablations; Supplementary Table S8 lists the objective and its weights. Contrastive weights are 0.18–0.20 and auxiliary weights are 0.02–0.08. For a training batch of size B, let $\widetilde{S}^{\mathrm{fused}} \in \mathbb{R}^{B\times B}$ denote the within-batch fused score matrix. Routed supervision is symmetric across EEG-to-image and image-to-EEG directions,

$$\mathcal{L}_{\mathrm{rce}} = \tfrac{1}{2}\Big[\mathrm{CE}(\widetilde{S}^{\mathrm{fused}}, I_B) + \mathrm{CE}((\widetilde{S}^{\mathrm{fused}})^{\mathsf{T}}, I_B)\Big], \tag{9}$$

where $I_B = (1,\ldots,B)$ denotes diagonal pairing. Router supervision is derived only from route-specific reference-score margins on training rows. If $U^{(b)}$ is the reference score matrix for route b, then

$$\begin{aligned} d_i^{(b)} &= U_{ii}^{(b)} - \max_{j\neq i} U_{ij}^{(b)}, \\ p_i^{(b)} &= \frac{\exp\Big(d_i^{(b)}/T_{\mathrm{trace}}\Big)}{\sum_{b'\in\mathcal{B}}\exp\Big(d_i^{(b')}/T_{\mathrm{trace}}\Big)}, \\ \mathcal{L}_{\mathrm{trace}} &= -\frac{1}{B}\sum_{i=1}^{B}\sum_{b\in\mathcal{B}} p_i^{(b)}\log q_i^{(b)}, \end{aligned} \tag{10}$$

with $T_{\mathrm{trace}} = 5$.

B.5 Training and Model Selection

Stage 2 updates the SAR and EPR projection and feature-normalization blocks, the SAR image projector, both residual heads, the router, and the three logit scales. Earlier backbone blocks, the CVR target model, and all reference networks remain frozen. AdamW (Loshchilov & Hutter, 2019)

Algorithm 1 CORTIVA training and model selection. Test evaluation follows checkpoint and fusion selection on concept-disjoint validation data.

Require: $\mathcal{D}_s^{\mathrm{fit}}, \mathcal{D}_s^{\mathrm{val}}, \mathcal{D}_s^{\mathrm{test}}$ for S participants; frozen $\{V^{(b)}\}_{b\in\mathcal{B}}$; grids $\mathcal{A}, \mathcal{T}$
Output: Participant-level retrieval results
for all participants $s \in \{1, \ldots, S\}$ **do**
  Initialize the participant-specific encoders, residual heads, and router
  for epoch $e = 1, \ldots, 60$ **do**
    Update trainable parameters on $\mathcal{D}_s^{\mathrm{fit}}$ using equation 8
    Compute $\mathcal{L}_s^{\mathrm{val}}(e)$; retain the minimum-loss checkpoint
    Stop after five epochs without improvement
  end for
  Restore the minimum-validation-loss checkpoint e_s^*
  for all $(\alpha, T_c) \in \mathcal{A} \times \mathcal{T}$ **do**
    Form $\overline{S}_s^{\mathrm{val}}$ by equation 11; compute ranks by equation 7
  end for
  Select $(\alpha_s^*, T_{c,s}^*)$ by equation 12 and the fixed tie rule
  Fix $(e_s^*, \alpha_s^*, T_{c,s}^*)$ and score $\mathcal{D}_s^{\mathrm{test}}$ using equation 4–equation 6
end for
return participant-level Top-1, Top-5, and ranks

uses $\beta = (0.9, 0.999)$, weight decay 0.01, gradient clipping at 1.0, batches of 384, and learning rates of 2×10^{-5} for encoder-side parameters and 2×10^{-4} for heads and router. Training runs for at most 60 epochs with patience 5, and validation loss selects the checkpoint within that budget.

Model parameters are trained on 1,454 concepts; checkpoints and fusion hyperparameters are selected on 200 disjoint validation concepts. Validation averages the ten images within each concept, whereas testing ranks individual held-out images (Supplementary Table S9). Because ranks are computed on concept-averaged rows, validation Top-1 is near ceiling and provides little discrimination among checkpoints. Checkpoint and fusion selection therefore use validation loss and mean rank. For $n_c = 10$ validation images per concept, the fused matrix is

$$\overline{S}_{s,cj}^{\mathrm{val}}(e, \alpha, T_c) = \frac{1}{n_c} \sum_{h=1}^{n_c} S_{s,(c,h),j}^{\mathrm{val}}(e, \alpha, T_c), \tag{11}$$

where (c, h) indexes the hth EEG row for concept c. Equation equation 11 produces the 200×200 matrix used for checkpoint and fusion selection; participant-specific parameters are listed in Supplementary Table S9.

For participant s, checkpoint epoch and fusion hyperparameters are selected sequentially from the validation data:

$$\begin{aligned} e_s^* &= \arg\min_e \mathcal{L}_s^{\mathrm{val}}(e), \\ (\alpha_s^*, T_{c,s}^*) &= \arg\min_{\alpha\in\mathcal{A},\, T_c\in\mathcal{T}} \frac{1}{200} \sum_{i=1}^{200} r_{s,i}^{\mathrm{val}}(e_s^*, \alpha, T_c), \end{aligned} \tag{12}$$

The blend-weight grid is $\mathcal{A} = \{0, 0.1, 0.2, 0.25, 0.3, 0.4\}$, and the confidence-temperature grid is $\mathcal{T} = \{2.5, 5.0, 10.0\}$. Mean-rank ties are resolved by Top-1, Top-5, lower α, and lower T_c, in that order. Algorithm 1 summarizes the selection procedure.

B.6 Computational Complexity

The trainable Stage 2 set contains 13.8 M parameters per participant. On an RTX 4080 SUPER, the Sub01 seed-2026 model processes one query in 6.30 ms (median over 1,000 timed iterations) with 92.3 MiB of peak allocated GPU memory when all three candidate banks are on device; offline visual feature extraction is excluded. Supplementary Table S4 reports parameter, latency, and memory details.

B.7 Structural Component Ablations

Three participant-paired variants remove SAR, CVR, or EPR from the fused score. For a removed route, its loss terms are dropped, its fusion weight is set to zero, its router-input slice is zeroed, and its route-exclusive parameters receive no gradient and remain at initialization. All other modules retain the full model's optimization settings. Because CVR depends on the shared EEG encoder, the no-SAR variant retains that encoder and isolates the SAR score vector; the no-CVR and no-EPR variants additionally remove route-exclusive parameters. The three interventions therefore differ in architectural scope while preserving matched optimization and selection.

Each leave-one-route-out model is independently retrained and selected under the same participant-wise evaluation design as full CORTIVA across Sub01–Sub10 and seeds 2026–2030. Participant-paired full-minus-ablated contrasts quantify route-specific effects; the CVR and EPR losses are clear, whereas the smaller SAR effect has a simultaneous interval that spans zero.

Training sufficiency is evaluated by extending otherwise matched runs from 18 to 60 epochs and comparing concept-disjoint validation loss (Supplementary Table S19).

B.8 THINGS-MEG Evaluation

THINGS-MEG contains 1,654 training concepts with 12 images each and 200 test images with 12 repetitions each (Hebart et al., 2023). For MEG, data are downsampled to 200 Hz, all 12 test repetitions are averaged, route spaces are 512-d, and the selected schedule is refitted on all 1,654 training concepts. The EEG backbones are not reused for MEG; instead, the neural encoder uses the EEGProject and TSConv components introduced by Zhang et al. (2026), while retaining CORTIVA's three target constructions and candidate-score fusion rule. Supplementary Table S5 lists the optimizer, warm-up, and augmentation settings.

For each participant and seed, 1,454 concepts form the training split and 200 form the disjoint selection split. A 15-epoch SAR warm-up precedes joint training for at most 60 epochs with patience 12. Warm-up duration, joint-training duration, and score-margin confidence parameters are chosen on validation data before the model is refitted on all 1,654 concepts using the selected schedule. Results average seeds 2026–2031; Supplementary Tables S5 and S6 list the corresponding settings.

B.9 Cross-Participant Adaptation

For each target participant, the other nine participants provide source models, and the source checkpoint with the lowest validation loss initializes adaptation. Target-participant observations are averaged by concept and divided into concept-disjoint adaptation and validation sets; the held-out retrieval concepts remain disjoint. Final encoder blocks, route heads, router, and target-participant parameters are updated for at most 200 epochs with patience 15; batch-normalization running statistics are then re-estimated on the adaptation split. Participant results average seeds 2026–2030; Supplementary Table S1 summarizes the evaluation setting.

B.10 Error and Representational Analyses

For each incorrect Top-1 prediction, we record whether the predicted image lies within the correct image's Top-k neighborhood in SAR's CLIP-RN50 space and compare the resulting rate with the analytical chance rate $k/199$ for a uniformly drawn wrong candidate. The external analysis repeats this calculation in DINOv2 ViT-B/14, a geometry outside CORTIVA target construction (Oquab et al., 2024). Representational similarity analysis (RSA) (Kriegeskorte et al., 2008; Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014) correlates the upper triangles of the 200×200 EEG and visual similarity matrices with Spearman ρ, separately for Sub01–Sub10. Posterior sensors have O, PO, P, or CP prefixes; frontal sensors have Fp, AF, or F prefixes; the occipital set is O1/Oz/O2. This analysis uses fixed EEG tensors and repeated measurements, keeping the external-geometry validation separate from model fitting and score selection.

To isolate the value of weight estimation from that of score integration, the controls retain the same 50 participant-by-seed route-score matrices while replacing query-aligned weights with 2,000 within-participant row permutations, participant means, uniform weights, or the $(0.55, 0.25, 0.20)$ prior. For

query i, normalized route-weight entropy is $-\sum_b w_i^{(b)} \log w_i^{(b)} / \log 3$, ranging from 0 for a single active route to 1 for uniform weights. Max-weight-route accuracy and an any-route Top-1 oracle characterize the range of route-wise performance.

B.11 STATISTICS

Seeds are averaged within participant before group inference. Query-level figures pool seed-level outcomes; inferential tests use participant summaries. Percentile-bootstrap intervals summarize participant variability and nonlinear statistics, while figure captions identify t-intervals. Paired contrasts use exact sign-flip tests over all $2^{10} = 1{,}024$ assignments (Ernst, 2004; Nichols & Holmes, 2002; Maris & Oostenveld, 2007) and 50,000 participant-bootstrap resamples (Efron & Tibshirani, 1993); endpoint intervals use 100,000 resamples. Posterior-versus-frontal hypotheses use one-sided exact sign tests.

Route-removal effects use simultaneous 95% participant-bootstrap intervals and Holm adjustment across nine route-by-metric tests (Holm, 1979); standardized paired effects are Cohen's d_z. Weight controls use paired percentile-bootstrap intervals with within-metric Holm adjustment. The full-window RSA contrast is primary, five temporal contrasts form a Holm-adjusted family, and 56 time–frequency contrasts use Benjamini–Hochberg correction (Benjamini & Hochberg, 1995).

B.12 ETHICS

This study is a secondary analysis of the public THINGS-EEG2 and THINGS-MEG releases; ethics approval and consent for the original acquisitions are reported by Gifford et al. (2022); Hebart et al. (2023).

C EXPANDED RESULTS

C.1 WITHIN-PARTICIPANT RETRIEVAL

After averaging seeds within participant, CORTIVA attains 73.5% Top-1 accuracy (SD 6.6 percentage points [pp]; 95% bootstrap CI 69.6–77.3%) and 95.3% Top-5 accuracy (SD 2.4 pp; 95% CI 93.8–96.6%) (main-text Fig. 2; Table 3). Top-1 ranges from 63.9% (Sub03) to 81.5% (Sub08), Top-5 from 89.8% (Sub05) to 97.6% (Sub10), and mean rank from 1.49 to 2.75 (group mean 1.84). The median rank is 1 for every participant. Because each participant contributes 200 queries, Top-1 differences are quantized at 0.5 pp.

Table 3: Participant-level 200-way retrieval on THINGS-EEG2: within-participant.

Method	Metric	Sub01	Sub02	Sub03	Sub04	Sub05	Sub06	Sub07	Sub08	Sub09	Sub10	Mean
Within-participant: train and test on one participant												
BraVL (Du et al., 2023)	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-GA (Song et al., 2024)	Top-1	15.2	13.9	14.7	17.6	9.0	16.4	14.9	20.3	14.1	19.6	15.6
	Top-5	40.1	40.1	42.7	48.9	29.7	44.4	43.1	52.1	39.7	46.7	42.8
NICE++ w/ GA (Song et al., 2025)	Top-1	16.6	20.2	19.6	25.1	13.5	16.5	17.8	26.2	18.2	22.9	19.7
	Top-5	46.9	47.6	52.6	61.0	42.3	48.2	51.6	62.1	48.7	53.5	51.5
MB2C (Wei et al., 2024)	Top-1	23.7	22.7	26.3	34.8	21.3	31.0	25.0	39.0	27.5	33.2	28.5
	Top-5	56.3	50.5	60.2	67.0	53.0	62.3	54.8	69.3	59.3	70.8	60.4
Neural-MCRL (Li et al., 2025)	Top-1	27.5	28.5	37.0	35.0	22.5	31.5	31.5	42.0	30.5	37.5	32.3
	Top-5	64.0	61.5	69.0	66.0	51.5	61.0	62.5	74.5	59.5	71.0	64.2
CognitionCapturer (Zhang et al., 2025)	Top-1	27.2	28.7	37.2	37.7	21.8	31.6	32.8	47.6	33.4	35.1	33.3
	Top-5	59.5	57.0	66.1	63.2	47.8	58.1	59.6	73.5	57.6	63.6	60.6
UBP (Wu et al., 2025)	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
NeuroBridge (Zhang et al., 2026)	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
CORTIVA (Ours)	**Top-1**	**79.1**	**67.2**	**63.9**	**69.2**	**65.6**	**81.3**	**77.7**	**81.5**	**75.0**	**74.8**	**73.5**
	Top-5	**95.5**	**95.0**	**93.9**	**93.9**	**89.8**	**97.3**	**97.3**	**97.0**	**95.8**	**97.6**	**95.3**

Table 4: Participant-level 200-way retrieval on THINGS-EEG2: cross-participant.

Method	Metric	Sub01	Sub02	Sub03	Sub04	Sub05	Sub06	Sub07	Sub08	Sub09	Sub10	Mean
BraVL (Du et al., 2023)	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 (Song et al., 2024)	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
ATM-S (Li et al., 2024)	Top-1	10.5	7.1	11.9	14.7	7.0	11.1	16.1	15.0	4.9	20.5	11.84
	Top-5	26.8	24.8	33.8	39.4	23.9	35.8	43.5	40.3	22.7	46.5	33.73
UBP (Wu et al., 2025)	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
Neural-MCRL (Li et al., 2025)	Top-1	13.0	12.0	14.5	12.5	11.5	13.5	14.0	18.5	13.5	17.0	14.0
	Top-5	31.5	30.5	35.5	35.5	29.0	35.5	36.0	38.5	32.5	39.0	34.3
NeuroBridge (Zhang et al., 2026)	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
CORTIVA (Ours)	**Top-1**	**27.0**	**26.0**	**28.2**	**25.6**	**27.9**	**29.0**	**19.5**	**32.7**	**19.2**	**29.5**	**26.5**
	Top-5	**57.5**	**57.3**	**58.2**	**64.1**	**59.5**	**67.7**	**51.6**	**65.5**	**48.5**	**64.7**	**59.5**

Values are participant means on the 200-way test. CORTIVA reports the five-seed average; analytical chance is 0.5%/2.5% for Top-1/Top-5. Supplementary Table S1 summarizes comparison sources and evaluation settings.

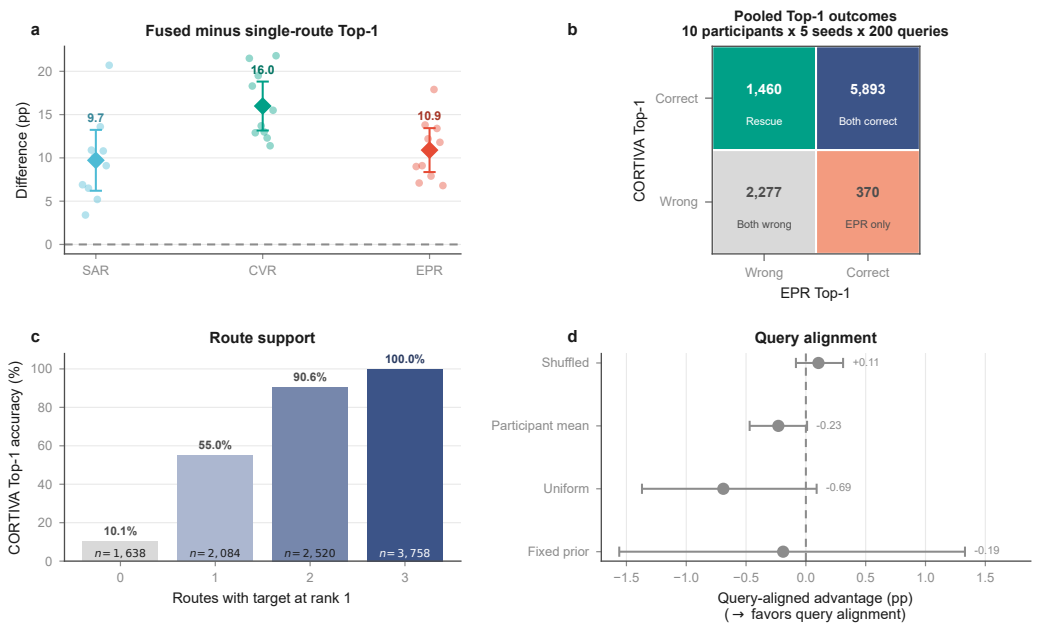

Figure 7: **Score integration and route-weight controls. a**, Participant-paired fused-minus-single-route Top-1 differences; diamonds and whiskers show means and 95% t-intervals. **b**, Top-1 outcome transitions between CORTIVA and EPR across all evaluated participant–seed–query combinations. **c**, CORTIVA Top-1 accuracy conditioned on how many routes individually rank the correct image first; bar labels give outcome counts. The zero-route column contains images recovered at rank 1 by score-weighted summation from sub-Top-1 route evidence. **d**, Participant-paired Top-1 advantages of query-aligned weights over four controls after averaging five seeds within participant. Whiskers are 95% participant-bootstrap intervals; positive values favor query alignment.

C.2 Stability Across Seeds and Training Duration

Participant-mean Top-1 is stable across seeds at $73.53 \pm 0.43\%$, whereas between-participant variability is larger (SD 6.6 pp; Supplementary Table S20). Extending training beyond 18 epochs reduces participant-wise validation loss by 7.1% on average, supporting the 60-epoch schedule used throughout the matched analysis (Supplementary Table S19).

C.3 Structural Component Ablations

The frozen EPR projection reference reaches 65.0% Top-1 before final CORTIVA training (Supplementary Table S14), and full CORTIVA exceeds it by 8.5 pp. Removing CVR and EPR produces Top-1 losses of 5.47 and 5.40 pp, respectively, each positive in all ten participants (main-text Fig. 3; Supplementary Table S15). Their single-route Top-1 accuracies differ by 5.1 pp (57.5% for CVR and 62.6% for EPR), so standalone accuracy and removal cost do not order the two routes in the same way.

CORTIVA exceeds each single-route output in all ten participants (Fig. 7a): participant means are 9.7 pp above SAR, 16.0 pp above CVR, and 10.9 pp above EPR. Together with the matched route-removal effects, these gains show that score integration combines route-specific evidence rather than inheriting the strongest route. Across 10,000 participant $\times$ seed $\times$ query outcomes, CORTIVA ranks the target first in 1,460 cases where EPR does not, while EPR ranks it first in 370 cases where CORTIVA does not (Fig. 7b). Conditioning on the number of individually correct routes, CORTIVA's Top-1 accuracy rises from 10.1% (zero routes) to 55.0% (one), 90.6% (two), and 100.0% (three). Thus, agreement among all routes is preserved, while fusion recovers 166 of the 1,638 outcomes in which no route ranks the target first.

Table 5: Within-participant 200-way retrieval on THINGS-MEG.

Method	Metric	Sub01	Sub02	Sub03	Sub04	Mean
Within-participant: train and test on one participant						
NICE (Song et al., 2024)	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 (Wu et al., 2025)	Top-1	15.0	46.0	27.3	18.5	26.7
	Top-5	38.0	80.5	59.0	43.5	55.2
NeuroBridge (Zhang et al., 2026)	Top-1	16.5	53.7	40.4	18.1	32.2
	Top-5	41.6	85.3	73.2	43.1	60.8
CORTIVA (Ours)	**Top-1**	**20.6**	**65.6**	**52.6**	**30.9**	**42.4**
	Top-5	**50.8**	**91.3**	**86.5**	**66.1**	**73.6**

Sub01–Sub04 denote the MEG cohort. CORTIVA averages seeds 2026–2031; Supplementary Table S1 summarizes the comparison settings.

All four Top-1 control intervals span zero around the validation-fixed query-aligned rule (Supplementary Table S16). Uniform summation gives the highest point estimate, 74.2% versus 73.5% for query-aligned weighting (query-aligned minus uniform: -0.69 pp, 95% CI $[-1.37, 0.09]$; Fig. 7d). The validation-fixed rule defines the primary result, and uniform summation provides a parameter-free deployment configuration.

Concept-averaged validation Top-1 ranges from 98.5% to 100.0%, and eight of ten participants select $\alpha_s = 0$ for every seed (Supplementary Table S9). Mean rank supplies the finer selection signal under this ceiling-level Top-1 accuracy.

C.4 Cross-Participant Adaptation

Under target-adapted leave-one-participant-out evaluation, CORTIVA achieves 26.5% Top-1 and 59.5% Top-5 across five seeds. Participant means range from 19.2% to 32.7% and from 48.5% to 67.7%, respectively. Table 4 presents the cross-participant results; Supplementary Table S1 summarizes the comparison settings.

C.5 Internal Fusion Variants

Fixed-weight and validation-rank fusion reach 70.6% and 70.8% Top-1, respectively, compared with 73.5% for CORTIVA (Supplementary Table S14).

C.6 Retrieval on THINGS-MEG

The MEG instantiation reuses the EEGProject and TSConv neural-front-end components introduced by NeuroBridge while retaining CORTIVA's target construction and score fusion. Across six seeds, CORTIVA attains 42.4% Top-1 and 73.6% Top-5. Participant Top-1 ranges from 20.6% (Sub01) to 65.6% (Sub02), while the seed-wise SD of the four-participant mean is 1.5 pp (Supplementary Table S6). In the frozen-checkpoint input analysis, Top-1 rises monotonically as the retained interval expands: 29.2% for 0–500 ms, 38.5% for 0–700 ms, and 42.4% for 0–1000 ms. The cumulative profile indicates useful information across the full 1-s epoch (Supplementary Fig. S8).

C.7 Repeated-Observation Efficiency and Input Sensitivity

Both retrieval accuracy and confidence ordering improve as additional observations are averaged: from one to four repetitions, Top-1 rises from 62.7% through 69.3% and 71.8% to 73.5%, while Top-5 increases from 90.1% to 95.3% (main-text Fig. 4). Over the same range, AURC decreases from 0.166 to 0.098.

All three negative controls return retrieval to analytical chance (Supplementary Fig. S1; Table S13): candidate-label permutation yields 0.50% Top-1 and 2.50% Top-5, pre-stimulus input yields 0.60% and 2.41%, and frontal-only input yields 0.54% and 2.41%, respectively. At the selected checkpoint,

retaining only O1, Oz, and O2 yields 29.5% Top-1. Additive Gaussian noise at 0.05 of the training-set channel SD lowers Top-1 from 73.5% to 71.6% (-1.9 pp); Supplementary Figs. S3 and S4 show the complete sensor, temporal, and perturbation profiles.

C.8 Visual Geometry of Errors and EEG Responses

The local error structure extends to a visual space not used to construct any target. In DINOv2 ViT-B/14, 21.9% of errors lie among the target's Top-10 neighbors versus 5.0% at random (4.36×; 10/10 participants above random; one-sided exact sign test, $p = 0.00098$; Supplementary Table S21). In SAR's CLIP-RN50 training geometry, incorrect Top-1 predictions fall within the correct image's Top-10 neighborhood in 25.9% of cases, compared with 5.0% for a random wrong candidate (5.15× enrichment). The corresponding Top-25 and Top-50 rates are 45.1% versus 12.6% and 64.4% versus 25.1% (2.56× at Top-50; main-text Fig. 5).

Correlations between participant-level EEG and CLIP similarity matrices are consistently positive: the mean Spearman ρ over posterior sensors is 0.047, positive in all ten participants, and the occipital group (O1, Oz, and O2) has a full-window mean of $\rho = 0.064$. The full-window posterior-minus-frontal contrast is $\Delta\rho = 0.027$ (10/10 positive; one-sided exact sign test, $p = 0.00098$). Among the five 100-ms windows, the largest descriptive contrast occurs at 100–200 ms ($\Delta\rho = 0.026$); contrasts remain positive after Holm correction at 0–100, 100–200, and 400–500 ms. EEG–DINOv2 RSA shows the same organization, with mean $\rho = 0.0316$ over posterior sensors and a posterior-minus-frontal contrast of $\Delta\rho = 0.0108$ (10/10 positive; $p = 0.00098$; Supplementary Table S21). Supplementary Figs. S6 and S7 detail the CLIP-based temporal and spectral structure.

D Expanded Discussion

D.1 Standalone Accuracy versus Removal Cost

CVR illustrates why route contribution must be assessed inside the fused system. Its standalone output reaches 57.5% Top-1, yet removing it costs 5.47 pp, essentially matching the 5.40-pp cost of removing the 62.6% EPR route. Fusion also recovers 166 outcomes missed at rank one by every constituent route. Full-score complementarity, including useful evidence below rank one, is therefore a more informative basis for route retention than standalone accuracy.

D.2 Adaptive Weighting and Score Integration

Partly nonoverlapping route successes and near-uniform learned weights (mean normalized entropy 0.938) accord with the sum-rule analysis of Kittler et al. (1998). Across the four Top-1 controls, intervals span zero and uniform summation gives the highest point estimate. The controls point to score integration as a main source of the gain and suggest uniform summation as a parameter-free alternative.

D.3 Modality and Participant Generalization

The score-space formulation extends across modalities and sensor layouts: it reaches 42.4% Top-1 with 271-sensor MEG and 26.5% under target-adapted cross-participant evaluation. For EEG, the 6.6-pp between-participant SD is approximately 15 times the 0.43-pp between-seed SD of the group mean. MEG Top-1 spans 45.0 pp (20.6–65.6%), compared with a 17.6-pp EEG range (63.9–81.5%); the seed-wise SD of the MEG group mean is 1.5 pp. Participant-to-participant variation is therefore much larger than optimization-seed variation in both cohorts.

D.4 Neural Correspondence

Error-neighbor enrichment and RSA measure different properties. The former is local and target-centered, whereas RSA compares all image pairs globally. Local error-neighbor enrichment coexists with modest global EEG–visual correlations, and both patterns also appear in DINOv2, which is not used to construct any target. The posterior-minus-frontal contrast places the stronger association over posterior sensors in both visual spaces.

D.5 Design Implications

The results favor retaining target-specific candidate scores through the final stage of neural retrieval. This design adds one candidate-score vector per target, supports direct route-removal tests, and does not require query-dependent weighting. Evaluation in DINOv2 separates the observed error organization from the visual spaces used to train the routes.

E Expanded Conclusion

CORTIVA retains route-specific rankings from three heterogeneous visual teachers until final candidate scoring. On 200-way THINGS-EEG2, it reaches 73.5% Top-1 and 95.3% Top-5, exceeding the strongest reported baseline by 10.3 and 5.4 points; the same principle transfers to THINGS-MEG and cross-participant adaptation. Matched route-removal and weight controls show that full-score complementarity, not standalone accuracy or specialized weighting, drives the gain and recovers cases unresolved by every route at rank one. DINOv2 and EEG representational analyses show semantically structured errors and stronger posterior neural–visual correspondence beyond the training targets. Together, these findings support candidate-score fusion as an effective and interpretable design for non-invasive neural image retrieval.

S1 SUPPLEMENTARY OVERVIEW

All group-level retrieval, ablation, routing, and representational-similarity analyses use the participant as the unit of inference. Query-level counts summarize 10,000 participant $\times$ seed $\times$ query outcomes, while inferential tests use participant-level summaries. Percentile-bootstrap intervals summarize endpoints, paired effects, and nonlinear statistics; figures explicitly marked as descriptive means use t-intervals. THINGS-MEG results aggregate four participants over six optimization seeds. Figures use a color-vision-deficiency-safe palette with redundant marker and line-style encodings where applicable.

S2 COMPARISON SOURCES AND SETTINGS

Table S1 summarizes the source and evaluation setting of each comparison. Participant entries and aggregate means follow the cited reports. In the NICE family, GA denotes graph attention; CognitionCapturer refers to its visual-output setting.

Table S1: Sources and evaluation settings for 200-way results.

Method	Publication	Result source	Evaluation setting
A. THINGS-EEG2 within-participant retrieval			
BraVL (Du et al., 2023)	IEEE TPAMI 2023	Table VIII	Within participant
NICE-GA (Song et al., 2024)	ICLR 2024	Table 2	Within participant
NICE++ w/ GA (Song et al., 2025)	IEEE TNNLS 2025	Table III	Within participant
MB2C (Wei et al., 2024)	ACM MM 2024	Table 1	Within participant
Neural-MCRL (Li et al., 2025)	IEEE ICME 2025	Table I	Within participant
CognitionCapturer (Zhang et al., 2025)	AAAI 2025	Table 2, image branch	Within participant
UBP (Wu et al., 2025)	CVPR 2025	Table 1	Within participant
NeuroBridge (Zhang et al., 2026)	AAAI 2026	Table 1	Within participant
CORTIVA	Current study	Main Table 1	Within participant
B. THINGS-EEG2 cross-participant retrieval			
BraVL (Du et al., 2023)	IEEE TPAMI 2023	Table VIII, 200-way V&T	Cross participant
NICE (Song et al., 2024)	ICLR 2024	Table 2	Cross participant
ATM-S (Li et al., 2024)†	NeurIPS 2024	UBP Table 1; NeuroBridge Table 1	Cross participant
UBP (Wu et al., 2025)	CVPR 2025	Table 1	Cross participant
Neural-MCRL (Li et al., 2025)	IEEE ICME 2025	Table I	Cross participant
NeuroBridge (Zhang et al., 2026)	AAAI 2026	Table 1	Cross participant
CORTIVA	Current study	Main Table 1	Cross participant
C. THINGS-MEG within-participant retrieval			
NICE (Song et al., 2024)	ICLR 2024	Appendix Table 8, retrieval	Within participant
UBP (Wu et al., 2025)	CVPR 2025	Table 2, intra-subject	Within participant

Table S1: Sources and evaluation settings for 200-way results (Continued).

Method	Publication	Result source	Evaluation setting
NeuroBridge (Zhang et al., 2026)	AAAI 2026	Table 6, intra-subject	Within participant
CORTIVA	Current study	Main Table 3; Table S6	Within participant

†ATM-S participant entries use the ten-participant results reported by UBP and NeuroBridge; the Mean column uses the ATM-S aggregate. CORTIVA uses 160 target-participant adaptation concepts and 40 validation concepts; test concepts are disjoint.

S3 Dataset and acquisition details

The EEG dataset is described in the THINGS-EEG2 publication (Gifford et al., 2022) and publicly distributed via OSF and NeMAR (Gifford, 2021; Gifford et al., 2026). THINGS-data describes the MEG collection (Hebart et al., 2023). Table S2 lists identifiers and acquisition metadata (Gorgolewski et al., 2016; Pernet et al., 2019).

Table S2: Datasets and acquisition parameters used in the CORTIVA analysis.

Property	Value	Source
Dataset	THINGS-EEG2	Dataset paper and NeMAR
NeMAR identifier	`NM000232`	NeMAR
Original OSF DOI	`10.17605/OSF.IO/3JK45`	OSF
NeMAR DOI	`10.82901/nemar.nm000232`	NeMAR
EEG participants	10	Dataset paper and NeMAR
EEG channels	63	Dataset paper and NeMAR
Acquisition sampling rate	1000 Hz	Dataset paper and NeMAR
EEG analysis grid	0–500 ms; 500 Hz; 250 samples	Current preprocessing pipeline
EEG preprocessing	Baseline correction, -100 to 0 ms; adjacent-sample averaging; no additional filtering or artifact rejection	Current preprocessing pipeline
BIDS version	1.9.0	NeMAR
EEG release license	CC BY 4.0	OSF and NeMAR
Primary EEG retrieval test	200 images; four session-level test means per image are averaged	Current evaluation protocol
Dataset	THINGS-MEG	THINGS-data publication
OpenNeuro identifier	`ds004212`	OpenNeuro / THINGS-data
MEG participants	4 (distinct from the EEG cohort)	THINGS-data publication
MEG sensors	271	THINGS-data publication
MEG acquisition sampling rate	1200 Hz	THINGS-data publication
MEG analysis grid	0–1000 ms; 200 Hz	Current evaluation protocol
Training set	1,654 concepts, 12 images per concept	THINGS-data publication

Table S2: Datasets and acquisition parameters used in the CORTIVA analysis (Continued).

Property	Value	Source
Retrieval test	200 images, 12 repetitions per image	THINGS-data publication

S4 Ethics and data use

CORTIVA is a secondary analysis of THINGS-EEG2 and THINGS-MEG. The dataset publications report ethics approval and consent for the original collections (Gifford et al., 2022; Hebart et al., 2023); OSF and NeMAR record the public EEG-release license (Gifford, 2021; Gifford et al., 2026).

S5 Retrieval configuration and model selection

The data split and principal hyperparameters for CORTIVA's shared SAR/CVR EEG encoder and independent EPR encoder are summarized in Table S3. Checkpoint and fusion parameters are chosen on concept-disjoint validation data before the 200-image test set is scored.

Table S3: CORTIVA retrieval configuration and model-selection summary.

Item	Value	Interpretation
Participants	10	Separate within-participant models for EEG Sub01–Sub10.
Training split	1,454 concepts / 14,540 EEG inputs per participant	Ten images per concept are used for parameter estimation.
Concept-disjoint validation set	200 concepts / 2,000 EEG trials per participant	Ten images per concept are aggregated for checkpoint and fusion selection.
Held-out retrieval test	200 image queries and candidates; four session-level test means averaged per query	Fixed THINGS-EEG2 query and candidate bank used for final Top-1 and Top-5 scoring.
Training schedule	Maximum 60 epochs, patience 5	Validation selects a checkpoint within the 60-epoch cap.
Optimization	AdamW (Loshchilov & Hutter, 2019); $\beta = (0.9, 0.999)$; batch 384; eval batch 200; base LR 2×10^{-5}; head/router LR 2×10^{-4}; weight decay 0.01; gradient norm 1.0	Optimizer hyperparameters.
Trainable modules	SAR/EPR projection and feature-normalization blocks; SAR image projector; CVR/EPR residual heads; router; route logit scales	13.79 M updated parameters per participant; earlier backbone blocks and all target/reference networks remain frozen.
EEG encoders	Shared EEG encoder for SAR/CVR and independent ensemble encoder for EPR	Backbone architectures.

Table S3: CORTIVA retrieval configuration and model-selection summary (Continued).

Item	Value	Interpretation
EPR spectral module	Five masks with thresholds 0.5–4, 4–8, 8–13, 13–30, and 30–45 on the module's fixed 250-Hz internal grid	Learns channel and band weights, applies an inverse FFT, and blends the result residually with the input.
Residual heads	CVR and EPR: $1024 \to 1024 \to 1024$, GELU, dropout 0.1, zero-initialized final layer	Residual adaptation initialized as the identity for both auxiliary routes.
Router	Concatenate three 1,024-d route vectors; LayerNorm; $3072 \to 512 \to 3$; GELU; dropout 0.05; prior $(0.55, 0.25, 0.20)$	Query-conditioned route weights before score-margin confidence weighting.
EEG input	63 channels; 0–500 ms; 500 Hz; 250 samples	Training session means are averaged by image; four test-session means are averaged for the primary evaluation.
Checkpoint selection	Concept-disjoint validation-set total loss	Selection criterion before scoring the 200-image test set.
Score-margin confidence selection	α grid: 0, 0.1, 0.2, 0.25, 0.3, 0.4; temperature grid: 2.5, 5.0, 10.0	Selected by validation mean rank.
Candidate-label permutation	10,000 candidate-column permutations per participant and seed	Re-scoring of each five-seed score matrix after candidate identities are permuted.

S6 COMPUTATIONAL FOOTPRINT

Table S4 distinguishes parameter adaptation from online routing and candidate scoring. Static counts are derived from tensor dimensions and the evaluation configuration. Measured rows use the frozen Sub01 seed-2026 model on an NVIDIA RTX 4080 SUPER with candidate banks resident; offline visual feature extraction is excluded. The multiply-accumulate subtotal covers the router's dense layers and the three candidate-bank dot products.

Table S4: Computational and resource footprint of within-participant EEG retrieval.

Resource or operation	Quantity	Derivation and scope	Stage
Trainable parameters	13.79 M (13,791,777) per participant	Final encoder blocks, prediction heads, and router; frozen visual targets remain fixed.	Training
Query router	1,581,059 parameters; 1,574,400 dense multiply-accumulates per query	LayerNorm over 3,072 features, $3072 \to 512 \to 3$ dense layers.	Training and online
Inference candidate banks	$3 \times 200 \times 1024 = 614{,}400$ float32 values (2.344 MiB)	One final 1,024-d candidate matrix per route after model selection.	Online

Table S4: Computational and resource footprint of within-participant EEG retrieval (Continued).

Resource or operation	Quantity	Derivation and scope	Stage
Visual feature sources	Excluded from the 13.8-million trainable count	Source feature extraction is offline; the selected SAR projector generates its final bank, while CVR and EPR load fixed targets.	Offline / finalization
Candidate dot products	614,400 multiply-accumulates per query	Three routes, 200 candidates, and 1,024 dimensions; excludes normalization and the final weighted sum.	Online
Measured inference, batch 1	6.30 ms median (IQR 0.22 ms); 92.26 MiB peak allocated (7.83 MiB above the loaded-model baseline)	Device-resident end-to-end inference over 1,000 timed iterations after 100 warm-up iterations.	Online
Measured inference, batch 200	14.22 ms median (IQR 0.12 ms); 663.83 MiB peak allocated (564.65 MiB incremental)	Device-resident end-to-end inference over 300 timed iterations after 100 warm-up iterations.	Online
Training input	$14{,}540 \times 63 \times 250$ float32 values (873.59 MiB)	Uncompressed participant-specific training tensor.	Training
Validation input	$2{,}000 \times 63 \times 250$ float32 values (120.16 MiB)	Concept-disjoint tensor used for checkpoint and fusion selection.	Validation
Optimization budget	38 batches per epoch; at most 2,280 updates per participant	$\lceil 14{,}540/384 \rceil \times 60$, with early-stopping patience 5.	Training
Confidence search	$6 \times 3 = 18$ configurations per participant	Six blend weights and three confidence temperatures evaluated on validation data.	Validation

S7 THINGS-MEG CONFIGURATION AND SEED STABILITY

The MEG evaluation uses the SAR, CVR, and EPR target families with candidate-score fusion. Its model uses 512-dimensional route spaces, a 271-sensor stem, and a modality-specific neural encoder. Table S5 gives the training and selection configuration.

Table S5: THINGS-MEG CORTIVA configuration.

Item	Value	Role
Participants	Sub01–Sub04 (MEG cohort)	Separate within-participant models; labels are distinct from the EEG cohort.
MEG input	271 sensors; 0–1 s; 200 Hz; participant-wise standardized	Input to the modality-specific sensor and temporal stem.
Training set	1,654 concepts; 12 images per concept	Participant-specific parameter estimation and full-data refit.
Selection split	1,454 training concepts / 200 concept-disjoint validation concepts	Selects the checkpoint and score-margin confidence parameters.

Table S5: THINGS-MEG CORTIVA configuration (Continued).

Item	Value	Role
Test retrieval	200 image queries and candidates; 12 MEG repetitions averaged per query	Final Top-1 and Top-5 scoring.
Routes	SAR, CVR, and EPR; 512-d projected score spaces	Three target-specific candidate-score vectors.
Neural encoders	EEGProject for SAR/CVR; TSConv for EPR (Zhang et al., 2026)	Modality-specific feature extraction.
SAR warm-up	Maximum 15 epochs	Selects the SAR initialization on validation mean rank.
Joint optimization	Maximum 60 epochs; patience 12; batch 1,024	Trains auxiliary routes and query-conditioned router.
Optimizer	AdamW; learning rate 10^{-4}; weight decay 10^{-4}	Shared optimization setting across all six seeds.
Training augmentation	Five-sample temporal smoothing independently applied to each channel with probability 0.30	Applied stochastically during training.
Confidence selection	α: 0, 0.1, 0.2, 0.25, 0.3, 0.4; T_c: 2.5, 5, 10	Validation-selected score-margin confidence weighting.
Full-data refit	Selected warm-up and joint durations on all 1,654 training concepts	Produces the final checkpoint before test scoring.
Optimization seeds	2026–2031	Six independent training runs.

Table S6: Six-seed THINGS-MEG retrieval and participant-level stability. Sub01–Sub04 denote the MEG cohort and are distinct from the identically numbered EEG participants.

A. Seed-level four-participant means

Seed	Top-1 (%)	Top-5 (%)
2026	43.000	73.750
2027	42.125	73.000
2028	43.375	75.000
2029	40.875	71.875
2030	40.625	73.875
2031	44.500	74.375
Mean $\pm$ SD	42.417 ± 1.501	73.646 ± 1.094

B. Participant-level summaries across six seeds

Participant	Top-1 (%)	Top-5 (%)
Sub01	20.583 ± 0.801	50.750 ± 2.019
Sub02	65.583 ± 4.421	91.250 ± 1.891
Sub03	52.583 ± 2.990	86.500 ± 1.871
Sub04	30.917 ± 1.201	66.083 ± 1.625

S8 ROUTE AND VISUAL-TARGET SPECIFICATIONS

Each EEG route is trained against a target candidate geometry derived from one or more frozen visual encoders. CLIP-RN50 supplies both SAR and EPR features (Radford et al., 2021); OpenCLIP ViT-B/32, SynCLR, and SDXL-VAE form the composite visual sources for CVR (Ilharco et al., 2021; Cherti et al., 2023; Schuhmann et al., 2022; Tian et al., 2024; Kingma & Welling, 2014; Podell et al., 2024). SAR and EPR share a feature source but use different target transforms: SAR updates its image projector during training, whereas EPR averages three frozen projections.

Table S7: CORTIVA route-to-target mapping.

Route	**Neural encoder**	**Visual source and target construction**	**Online role**
Semantic Alignment Route (SAR)	Shared graph-token EEG encoder; 1,024-d output	CLIP-RN50 features followed by the trainable SAR image projector	Produces $S^{(\mathrm{sar})}$ over the 200 candidates.
Composite Visual Route (CVR)	Residual CVR head applied to the SAR EEG vector	OpenCLIP-B/32, SynCLR, and SDXL-VAE features transformed by fixed source-specific projectors and fused by a preceding composite-target model into a normalized 1,024-d target	Produces $S^{(\mathrm{cvr})}$.
Ensemble Projection Route (EPR)	Independent spectral–temporal encoder plus residual EPR head; 1,024-d output	Normalized mean of three frozen image-projection checkpoints applied to CLIP-RN50 features, followed by final normalization	Produces $S^{(\mathrm{epr})}$.
Final CORTIVA score	Query-conditioned router blended with validation-selected score-margin weights	Weighted sum of the three route-specific 200-way score vectors	Produces $S^{(\mathrm{fused})}$.

Stage 1 performs participant-specific, single-target EEG–image alignment on the training and validation concepts. It produces the paired EEG and image-projection checkpoints used to initialize Stage 2, including the CVR source projectors and composite-target model and the three EPR projection pairs. Stage 2 freezes those target/reference networks and trains CORTIVA for 60 epochs. Participant-specific initialization pairs are selected by Stage 1 validation loss.

Candidate scoring for route b uses the temperature-scaled cosine similarity in Eq. (1) of the main text. The router receives concatenated SAR, CVR, and EPR EEG vectors. For each route, let $S^{(b)}_{si(1)} \geq S^{(b)}_{si(2)}$ be the two largest candidate scores for query i. The score-margin confidence vector is

$$
\begin{aligned}
m^{(b)}_{s,i} &= S^{(b)}_{si(1)} - S^{(b)}_{si(2)}, \\
\mu_{s,i} &= |\mathcal{B}|^{-1} \sum_{b \in \mathcal{B}} m^{(b)}_{s,i}, \\
\sigma_{s,i} &= \left[|\mathcal{B}|^{-1} \sum_{b \in \mathcal{B}} \left(m^{(b)}_{s,i} - \mu_{s,i} \right)^2 \right]^{1/2}, \\
\widehat{m}^{(b)}_{s,i} &= \frac{m^{(b)}_{s,i} - \mu_{s,i}}{\sigma_{s,i} + 10^{-6}}, \\
c^{(b)}_{s,i} &= \frac{\exp\left(\widehat{m}^{(b)}_{s,i} / T_{c,s} \right)}{\sum_{b' \in \mathcal{B}} \exp\left(\widehat{m}^{(b')}_{s,i} / T_{c,s} \right)}.
\end{aligned}
\tag{S1}
$$

Participant-specific α_s and $T_{c,s}$ blend $c_{s,i}$ with the learned router weights before ranking. Candidate-score fusion preserves the learned inverse-temperature scale of each route vector.

S9 Training objective and validation-selected fusion parameters

The training objective contains 15 terms whose weights are fixed globally across participants, seeds, and matched ablations. Let $\bar{x} = x/\|x\|_2$, let $I_B = (1, \ldots, B)$ denote diagonal labels for a batch of size B, and let softmax operate row-wise. The symmetric contrastive, cosine-preservation, and relational losses are

$$\mathcal{C}_{\tau_b}(A, V) = \tfrac{1}{2}\left[\mathrm{CE}(\tau_b \bar{A}\bar{V}^\mathsf{T}, I_B) + \mathrm{CE}(\tau_b \bar{V}\bar{A}^\mathsf{T}, I_B)\right], \tag{S2}$$

$$\mathcal{D}_{\cos}(A, T) = 1 - \frac{1}{B}\sum_{i=1}^{B} \bar{a}_i^\mathsf{T} \bar{t}_i, \tag{S3}$$

$$P_\tau(A) = \mathrm{softmax}(\bar{A}\bar{A}^\mathsf{T}/\tau), \tag{S4}$$

$$\mathcal{D}_{\mathrm{rel}}(A, T) = \tau^2\, \mathrm{KL}(P_\tau(T) \,\|\, P_\tau(A)). \tag{S5}$$

where $\tau_b = \min[\exp(\eta_b), 100]$ is the learned inverse temperature of route b, matching the main text. For candidate-distribution distillation, define

$$P_\tau(A; V) = \mathrm{softmax}(\bar{A}\bar{V}^\mathsf{T}/\tau), \tag{S6}$$

$$\mathcal{D}_{\mathrm{logit}}(A, T; V) = \tau^2\, \mathrm{KL}(P_\tau(T; V) \,\|\, P_\tau(A; V)). \tag{S7}$$

For the M-member EPR projection ensemble, the target distribution is the arithmetic mean of member probabilities:

$$P_A^{\mathrm{ens}} = \frac{1}{M}\sum_{m=1}^{M} P_\tau(A; V_m), \tag{S8}$$

$$P_T^{\mathrm{ens}} = \frac{1}{M}\sum_{m=1}^{M} P_\tau(T_m; V_m), \tag{S9}$$

$$\mathcal{D}_{\mathrm{ens}}(A, \{T_m, V_m\}) = \tau^2\, \mathrm{KL}(P_T^{\mathrm{ens}} \,\|\, P_A^{\mathrm{ens}}). \tag{S10}$$

Every KL divergence is averaged over batch rows.

Let Z_{s} and V_{s} be the SAR EEG and trainable projected-image representations, X_{t} the frozen text representation, and $\widetilde{Z}_{\mathrm{s}}, \widetilde{V}_{\mathrm{s}}$ the normalized SAR reference consensus. Let $Z_{\mathrm{c}}, V_{\mathrm{c}}$ denote the CVR prediction and frozen composite target. Let Z_{e} be the EPR prediction, $\widetilde{Z}_{\mathrm{e}}$ and V_{e} the normalized means of the frozen EPR EEG and image reference projections, and $\{\widetilde{Z}_{\mathrm{e},r}, V_{\mathrm{e},r}\}_{r=1}^{M}$ their individual pairs. The objective is

$$\begin{aligned}
\mathcal{L} = {} & 0.20\,\mathcal{C}_{\tau_{\mathrm{s}}}(Z_{\mathrm{s}}, V_{\mathrm{s}}) + 0.03\,\mathcal{C}_{\tau_{\mathrm{s}}}(V_{\mathrm{s}}, X_{\mathrm{t}}) + 0.08\,\mathcal{D}_{\cos}(Z_{\mathrm{s}}, \widetilde{Z}_{\mathrm{s}}) \\
& + 0.03\,\mathcal{D}_{\mathrm{rel}}(Z_{\mathrm{s}}, \widetilde{Z}_{\mathrm{s}}) + 0.03\,\mathcal{D}_{\mathrm{logit}}(Z_{\mathrm{s}}, \widetilde{Z}_{\mathrm{s}}; \widetilde{V}_{\mathrm{s}}) + 0.02\,\mathcal{D}_{\cos}(V_{\mathrm{s}}, \widetilde{V}_{\mathrm{s}}) \\
& + 0.18\,\mathcal{C}_{\tau_{\mathrm{c}}}(Z_{\mathrm{c}}, V_{\mathrm{c}}) + 0.05\,\mathcal{D}_{\cos}(Z_{\mathrm{c}}, V_{\mathrm{c}}) + 0.04\,\mathcal{D}_{\mathrm{rel}}(Z_{\mathrm{c}}, V_{\mathrm{c}}) \\
& + 0.18\,\mathcal{C}_{\tau_{\mathrm{e}}}(Z_{\mathrm{e}}, V_{\mathrm{e}}) + 0.05\,\mathcal{D}_{\cos}(Z_{\mathrm{e}}, \widetilde{Z}_{\mathrm{e}}) + 0.04\,\mathcal{D}_{\mathrm{rel}}(Z_{\mathrm{e}}, \widetilde{Z}_{\mathrm{e}}) \\
& + 0.04\,\mathcal{D}_{\mathrm{ens}}(Z_{\mathrm{e}}, \{\widetilde{Z}_{\mathrm{e},r}, V_{\mathrm{e},r}\}_{r=1}^{M}) + 0.05\,\mathcal{L}_{\mathrm{rce}} + 0.04\,\mathcal{L}_{\mathrm{trace}}.
\end{aligned} \tag{S11}$$

The relational and candidate-distribution temperatures are both $\tau = 0.07$, and the router-trace target temperature is 5. Table S8 lists the 15 terms and their fixed weights. Training and evaluation use PyTorch (Paszke et al., 2019) and NumPy (Harris et al., 2020).

Table S8: Training-loss terms in CORTIVA.

Loss term	Weight	Function
SAR symmetric contrastive	0.20	Encourages agreement between SAR EEG and projected CLIP-RN50 image representations.
Image–text contrastive	0.03	Encourages image–text agreement in the trainable semantic image projection.
SAR reference-direction cosine	0.08	Penalizes deviation from the normalized SAR reference direction.
SAR relational KL	0.03	Penalizes deviations from the SAR reference ensemble's within-batch relational geometry.
SAR candidate-logit KL	0.03	Penalizes divergence from SAR reference candidate distributions.
Image-projection preservation	0.02	Penalizes deviation from the reference semantic image projection.
CVR contrastive	0.18	Encourages agreement between the CVR residual route and the frozen composite visual target.
CVR cosine	0.05	Penalizes deviation from the normalized CVR target direction.
CVR relational	0.04	Penalizes deviations from within-batch composite-target geometry.
EPR contrastive	0.18	Encourages agreement between the independent EPR route and the frozen projection ensemble.
EPR reference cosine	0.05	Penalizes deviation from the normalized EPR reference-ensemble EEG direction.
EPR relational	0.04	Penalizes deviations from EPR reference relational geometry.
EPR ensemble-logit KL	0.04	Penalizes divergence from candidate distributions of the frozen EPR projection ensemble.
Routed symmetric cross-entropy	0.05	Trains the fused route-score matrix in both EEG-to-image and image-to-EEG directions.
Router trace cross-entropy	0.04	Encourages router weights to follow route-indexed reference-score margins on training rows.

Table S9: Five-seed validation-selected score-margin confidence parameters. Parameter columns give the range across seeds 2026–2030; validation metrics are mean ± SD across seeds.

Participant	α_s range	$T_{c,s}$ range	Val. Top-1 (%)	Val. mean rank
Sub01	0.0	2.5	99.5 ± 0.0	1.010 ± 0.000
Sub02	0.0	2.5	100.0 ± 0.0	1.000 ± 0.000
Sub03	0.0–0.1	2.5	98.9 ± 0.7	1.011 ± 0.007
Sub04	0.0	2.5	99.4 ± 0.2	1.006 ± 0.002
Sub05	0.0–0.4	2.5–5.0	98.5 ± 0.5	1.015 ± 0.005
Sub06	0.0	2.5	100.0 ± 0.0	1.000 ± 0.000
Sub07	0.0	2.5	100.0 ± 0.0	1.000 ± 0.000
Sub08	0.0	2.5	99.5 ± 0.0	1.005 ± 0.000
Sub09	0.0	2.5	99.9 ± 0.2	1.001 ± 0.002
Sub10	0.0	2.5	100.0 ± 0.0	1.000 ± 0.000

The 2,000 validation observations comprise 200 concepts with ten images per concept. Candidate features and score rows are averaged within concept to form the 200×200 validation matrix. The final test contains one held-out image per concept and one query per candidate.

Concept-level aggregation produces validation Top-1 values of 98.5–100.0% and mean ranks of 1.000–1.015 across participants. Eight of ten participants select $\alpha_s = 0$ for all five seeds; Sub03 and Sub05 select a nonzero confidence contribution in only a subset of seeds. Thus, the score-margin term is inactive for most EEG configurations.

S10 EEG INPUT AND INDEX ALIGNMENT

The full input contains all 63 channels over 0–500 ms at 500 Hz, represented by 250 samples. Session-level trial means are averaged within image, producing 16,540 image-level development inputs: 14,540 for fitting and 2,000 for validation. The primary held-out evaluation averages four test-session means into one 63×250 query tensor per image; the repetition analysis evaluates every combination of one through four sessions.

Table S10: EEG input surfaces and orthogonal negative controls for 60-epoch CORTIVA retrieval.

Condition	**Input timing**	**Channels and processing**	**Role**
Full CORTIVA input	0–500 ms; 500 Hz; 250 samples	All 63 channels; four-session average before scoring	Primary input
Candidate-label permutation	Same tensor as the full CORTIVA input	Candidate identities permuted after scoring	Indexing check
Pre-stimulus input	-500 to 0 ms; 500 Hz; 250 samples	Real duration-matched EEG from all 63 channels	Temporal control
Frontal-only input	0–500 ms; 500 Hz; 250 samples	Frontal sensors retained; all other channels set to zero	Spatial control

Together, the controls disrupt query–candidate identity, post-stimulus timing, or posterior spatial information while preserving the frozen CORTIVA endpoint and its 200-way candidate bank.

S11 VISUAL TARGET SPECIFICATIONS

CLIP-RN50 supplies SAR and EPR features (Radford et al., 2021); OpenCLIP ViT-B/32, SynCLR, and SDXL-VAE form the composite visual target (Ilharco et al., 2021; Cherti et al., 2023; Schuhmann et al., 2022; Tian et al., 2024; Kingma & Welling, 2014; Podell et al., 2024).

Table S11: Frozen visual sources used in CORTIVA. Development rows comprise 14,540 fitting and 2,000 validation images.

Source key	**Layer**	**Dev. rows**	**Test rows**	**Role**
`clip-rn50-features`	Candidate source	16,540	200	1,024-d source for SAR and EPR candidate construction
`CLIP-ViT-B-32-laion2B-s34B-b79K`	Composite-target input	16,540	200	OpenCLIP ViT-B/32, LAION-2B-en, 512-d
`synclr_vit_b_16`	Composite-target input	16,540	200	SynCLR ViT-B/16, 768-d
`SDXL-VAE`	Composite-target input	16,540	200	SDXL AutoencoderKL; $16 \times 16 \times 4$ latent flattened to 1,024-d

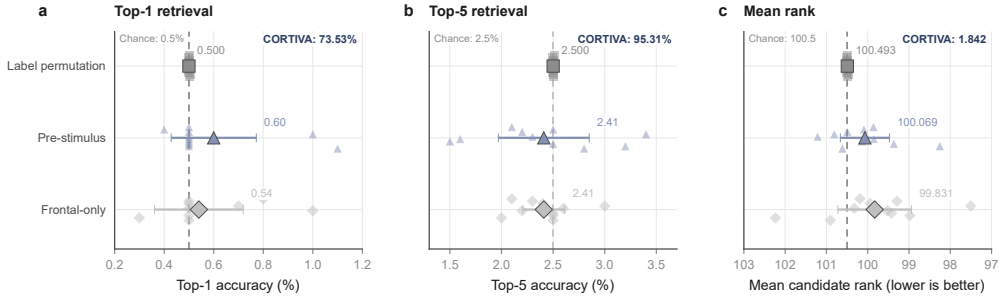

Figure S1: **Three orthogonal negative controls at the analytical-chance scale. a–c**, Participant-level Top-1, Top-5, and mean-rank results for candidate-label permutation, duration-matched pre-stimulus EEG, and frontal-only post-stimulus input. Small markers denote participant means after five-seed averaging; large markers and whiskers denote group means and 95% t-intervals. Dashed lines mark analytical chance, and navy annotations give the corresponding CORTIVA endpoints. The controls isolate query–candidate identity, post-stimulus timing, and posterior spatial information. Holm-adjusted exact sign-flip tests separate CORTIVA from candidate-label permutation across all three metrics ($p_{\mathrm{Holm}} = 0.0059$).

S12 Retrieval summary

SD and range quantify between-participant variation, whereas intervals quantify uncertainty in the group means.

Table S12: CORTIVA aggregate statistics over participant-level summaries.

Metric	n	**Mean**	**SD**	**95% CI**	**Min–Max**
Top-1 retrieval (%)	10	73.53	6.59	69.62–77.33	63.90–81.50
Top-5 retrieval (%)	10	95.31	2.37	93.81–96.56	89.80–97.60
Mean rank (candidates)	10	1.842	0.372	1.654–2.083	1.493–2.749

Table S13: Orthogonal negative-control summary for 60-epoch CORTIVA retrieval. Values are means over ten participant-level five-seed summaries; candidate-label values additionally average 10,000 permutations per participant and seed.

Condition	**Information disrupted**	**Top-1 (%)**	**Top-5 (%)**	**Mean rank**
CORTIVA	None	**73.53**	**95.31**	**1.842**
Candidate-label permutation	Query–candidate identity	0.500	2.500	100.493
Pre-stimulus input	Post-stimulus timing	0.60	2.41	100.069
Frontal-only input	Posterior sensor information	0.54	2.41	99.831
Analytical chance	—	0.50	2.50	100.5

S13 CORTIVA COMPONENT COMPARISON

Table S14: CORTIVA component comparison. Values are ten-participant means on the same 200-way test; deltas are CORTIVA minus the row value.

Configuration	Family	Top-1 (%)	Top-5 (%)	ΔT1 (pp)	ΔT5 (pp)	Definition
Semantic Alignment Route (SAR)	route output	63.81	90.48	9.72	4.83	Graph-token EEG encoder scored against projected CLIP-RN50 candidates.
Composite Visual Route (CVR)	route output	57.54	88.19	15.99	7.12	Residual CVR head scored against the frozen composite visual target.
Frozen composite visual target	reference diagnostic	49.55	81.00	23.98	14.31	Frozen target-side output scored against its matching visual target.
Ensemble Projection Route (EPR)	route output	62.63	89.45	10.90	5.86	Independent EPR encoder scored against its image-projection ensemble.
Frozen EPR projection ensemble	reference diagnostic	65.00	91.45	8.53	3.86	Frozen projection-ensemble scores.
Fixed-weight fusion	fusion variant	70.57	94.56	2.96	0.75	Route fusion with a prespecified fixed weight vector.
Validation-rank fusion	fusion variant	70.84	94.61	2.69	0.70	Query-weighted negative-rank fusion before final candidate-score fusion.
CORTIVA	**fused model**	**73.53**	**95.31**	**0.00**	**0.00**	Validation-selected fusion of three 200-way score vectors.

Relative to negative-rank fusion, CORTIVA gains 2.69 pp Top-1 and 0.70 pp Top-5. The route-weight controls below test whether this advantage depends on query-aligned weights.

S14 MATCHED STRUCTURAL ABLATIONS

Each leave-one-route-out model is independently retrained under the matched design used for full CORTIVA. Participant metrics average five seeds; losses are full minus ablated for accuracy and ablated minus full for mean rank.

Table S15: Five-seed 60-epoch structural ablation statistics. Each interval is a simultaneous 95% participant-bootstrap interval within the corresponding metric; p-values are exact two-sided sign-flip tests Holm-adjusted across all nine route-by-metric tests. Top-1 paired standardized effects d_z are 1.04, 1.88, and 1.42 for SAR, CVR, and EPR removal.

Removed route	Top-1 loss [95% CI] (pp)	(p)	Top-5 loss [95% CI] (pp)	(p)	Mean-rank loss [95% CI]	(p)
SAR	(0.95 [-1.31,3.21])	0.055	(0.70 [-0.47,1.87])	0.055	(0.074 [-0.177,0.326])	0.055
CVR	(5.47 [3.21,7.73])	0.018	(3.03 [1.86,4.20])	0.023	(0.585 [0.334,0.837])	0.018
EPR	(5.40 [3.14,7.66])	0.018	(1.26 [0.09,2.43])	0.055	(0.257 [0.005,0.508])	0.029

CVR and EPR show robust removal costs, with simultaneous Top-1 intervals of 3.21–7.73 and 3.14–7.66 pp, respectively. SAR removal yields 0.95 pp with a simultaneous 95% interval of -1.31 to 3.21 pp.

S15 ROUTE-WEIGHT CONTROLS

Table S16 quantifies the sensitivity of query-aligned weights to four controls across the 50 route-score matrices. Accuracy effects are query-aligned minus control accuracy, and mean-rank effects are control minus query-aligned rank. The shuffled control averages 2,000 within-matrix permutations per participant. Intervals use 50,000 participant bootstrap resamples, and exact sign-flip tests are Holm-adjusted within each metric.

Table S16: Paired route-weight controls. Effects compare query-aligned weights with each control after within-participant seed averaging. Values are means with 95% participant-bootstrap intervals; positive effects favor query alignment.

Weight control	Metric	Aligned	Control	Effect [95% CI]	p_{Holm}
Shuffled within participant	Top-1 (%)	73.53	73.42	$0.11\ [-0.08, 0.31]$	0.70
	Top-5 (%)	95.31	95.42	$-0.11\ [-0.33, 0.15]$	0.96
	Mean rank	1.842	1.844	$0.002\ [-0.012, 0.020]$	1.00
Participant-mean weights	Top-1 (%)	73.53	73.76	$-0.23\ [-0.47, 0.01]$	0.47
	Top-5 (%)	95.31	95.63	$-0.32\ [-0.58, -0.04]$	0.27
	Mean rank	1.842	1.822	$-0.020\ [-0.035, -0.002]$	0.23
Uniform weights	Top-1 (%)	73.53	74.22	$-0.69\ [-1.37, 0.09]$	0.47
	Top-5 (%)	95.31	95.49	$-0.18\ [-0.63, 0.25]$	0.96
	Mean rank	1.842	1.834	$-0.008\ [-0.062, 0.057]$	1.00
Fixed prior weights	Top-1 (%)	73.53	73.72	$-0.19\ [-1.56, 1.33]$	0.82

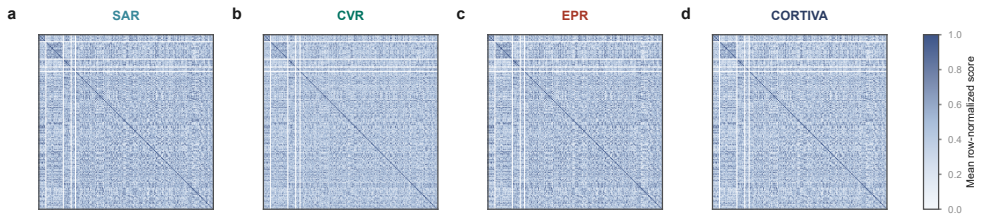

Figure S2: **Route-specific candidate-score structure. a–d**, Mean row-normalized score matrices for the SAR, CVR, EPR, and fused CORTIVA outputs. Queries and candidates use the same label-derived high-level group ordering; white lines mark group boundaries.

Table S16: Paired route-weight controls (Continued).

Weight control	Metric	Aligned	Control	Effect [95% CI]	p_{Holm}
	Top-5 (%)	95.31	94.92	0.39 $[-0.22, 1.07]$	0.96
	Mean rank	1.842	1.906	0.064 $[-0.020, 0.170]$	1.00

Uniform weighting gives the highest Top-1 point estimate, 74.22% versus 73.53% for query-aligned weighting, and all four paired Top-1 intervals span zero. Selecting the maximum-weight route yields 59.91% Top-1, whereas the any-correct-route oracle reaches 83.62%; mean normalized weight entropy is 0.938. Uniform summation gives a similar Top-1 estimate with the simplest weighting rule and falls within the same inferential range as the validation-fixed query-aligned configuration.

S16 Input sensitivity and practical sampling

Input sensitivity is measured at the selected 60-epoch participant checkpoints, holding visual targets and fusion parameters fixed across every perturbation.

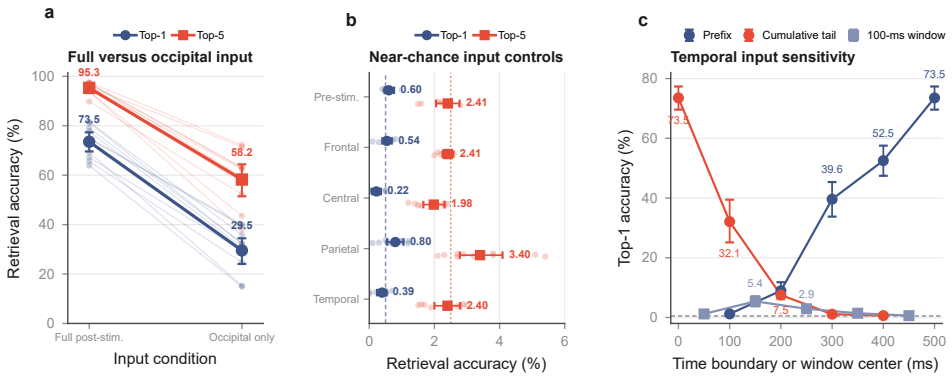

Figure S3: **Sensor and temporal input-ablation sensitivity. a**, Paired participant-level Top-1 and Top-5 accuracy for the full 0–500-ms post-stimulus input and occipital-only input; bold trajectories connect group means. **b**, Top-1 and Top-5 accuracy for duration-matched pre-stimulus input and four additional sensor groups, with metric-specific analytical-chance levels. Small points, large markers, and whiskers in **a,b** denote participant means, group means, and 95% intervals. **c**, Cumulative-prefix, cumulative-tail, and independent 100-ms-window Top-1 sensitivity. Participant means are computed after averaging five optimization seeds within participant. The full input uses all 63 channels; retained-input conditions preserve the named samples and set all others to zero. The pre-stimulus and frontal-only results are also summarized with the orthogonal controls in Fig. S1.

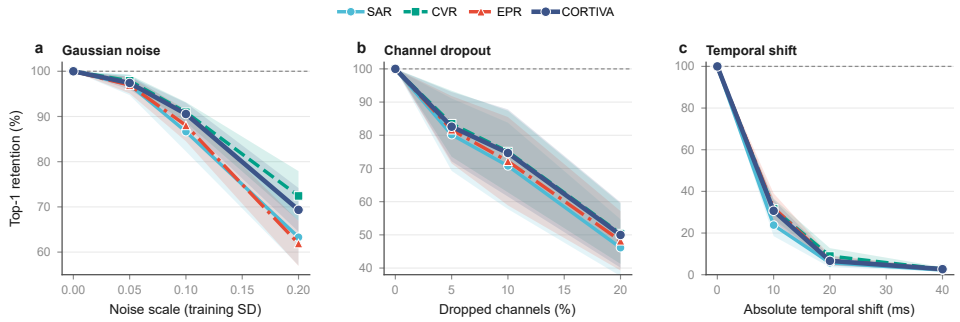

Figure S4: **EEG perturbation sensitivity.** Top-1 retention relative to the clean four-observation input under **a**, Gaussian noise scaled by the participant's training-set channel standard deviations; **b**, nested channel dropout; and **c**, symmetric zero-padded temporal shifts. Four deterministic realizations are averaged for noise and channel dropout; positive and negative shifts are averaged within absolute shift.

Table S17: Input-sensitivity analysis. Values average five optimization seeds within participant and then ten participants. The pre-stimulus row uses real, duration-matched input; sensor-group and temporal rows preserve the named samples at a frozen checkpoint and set all others to zero.

Condition	**Family**	**Top-1 (%)**	**Top-5 (%)**	**Mean rank**
Full 0–500 ms	Full input	73.53	95.31	1.842
Pre-stimulus input	Matched negative	0.60	2.41	100.069
Frontal sensors	Sensor group	0.54	2.41	99.831
Central sensors	Sensor group	0.22	1.98	105.785
Parietal sensors	Sensor group	0.80	3.40	89.868
Temporal sensors	Sensor group	0.39	2.40	101.730
Occipital sensors	Sensor group	29.54	58.23	12.480
0–100 ms	Cumulative prefix	1.19	6.10	79.065
0–200 ms	Cumulative prefix	8.95	24.40	37.238
0–300 ms	Cumulative prefix	39.60	72.16	6.764
0–400 ms	Cumulative prefix	52.52	83.38	3.795
100–500 ms	Cumulative tail	32.09	62.67	10.685
200–500 ms	Cumulative tail	7.53	21.42	42.999
300–500 ms	Cumulative tail	1.10	4.73	81.331
400–500 ms	Cumulative tail	0.62	3.08	95.486
0–100 ms	Independent window	1.19	6.10	79.065
100–200 ms	Independent window	5.40	15.06	51.503
200–300 ms	Independent window	2.91	9.20	67.549
300–400 ms	Independent window	1.42	4.26	84.278
400–500 ms	Independent window	0.62	3.08	95.486

Table S18: CORTIVA repetition and perturbation summary. Values are participant-level means after averaging seeds within participant.

Analysis	Level	Top-1 (%)	Top-5 (%)	Mean rank
Test repetitions	1	62.71	90.06	2.739
	2	69.27	93.81	2.108
	3	71.75	94.88	1.934
	4	73.53	95.31	1.842
Gaussian noise	0.05 training SD	71.56	94.66	1.970
	0.10 training SD	66.45	92.22	2.357
	0.20 training SD	50.79	82.58	4.243
Channel dropout	5%	60.91	85.01	5.868
	10%	55.24	79.61	8.323
	20%	36.79	62.54	15.678
Temporal shift	10 ms	22.45	50.31	15.018
	20 ms	4.84	16.13	51.050
	40 ms	1.96	6.63	76.122

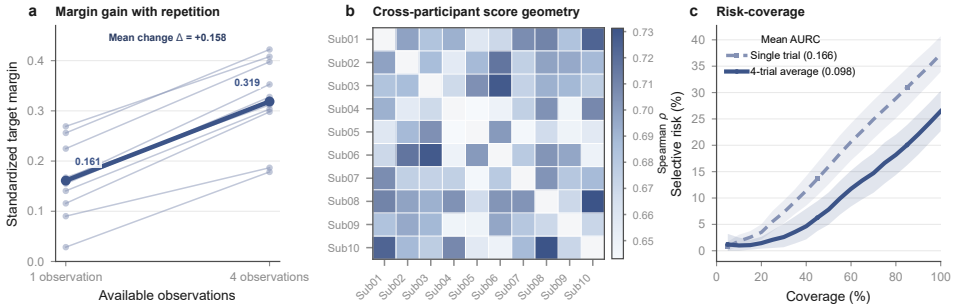

Figure S5: **Repetition-dependent retrieval geometry.** **a**, Participant-paired standardized target-versus-hardest-negative margins for one and four available observations; the navy line is the participant mean, endpoint labels give its values, and the central annotation gives the mean paired change. **b**, Spearman correlations between participant-level four-observation fused-score matrices after averaging five optimization seeds. The off-diagonal mean is $\rho = 0.684$ (range 0.643–0.731), quantifying cross-participant consistency of retrieval geometry; the unit diagonal is omitted and shown as blank cells. **c**, Selective risk as retrievals are retained from highest to lowest confidence; bands are 95% participant-bootstrap intervals, line styles distinguish one versus four observations, and legend values report participant-mean AURC.

S17 Representational analyses

RSA compares neural and model representational geometries (Kriegeskorte et al., 2008). The mean EEG–CLIP Spearman ρ over posterior sensors is 0.047 for 0–500 ms and positive in all ten participants. O1/Oz/O2 have a full-window regional mean of $\rho = 0.064$; the posterior-minus-frontal difference is $\Delta\rho = 0.027$ (one-sided exact sign test, $p = 0.00098$). Among the five 100-ms windows, the largest contrast occurs at 100–200 ms ($\Delta\rho = 0.026$); the 0–100, 100–200, and 400–500 ms contrasts remain positive after Holm correction (Supplementary Figs. S6 and S7).

Time–frequency RSA uses log-bandpower in the 4–8, 8–13, 13–30, and 30–80 Hz bands on the 0–500 ms, 500-Hz analysis grid. FFTs are zero-padded to 512 points for a common frequency grid across windows; the 2-Hz native resolution of a 500-ms window makes the lowest-frequency estimates coarse.

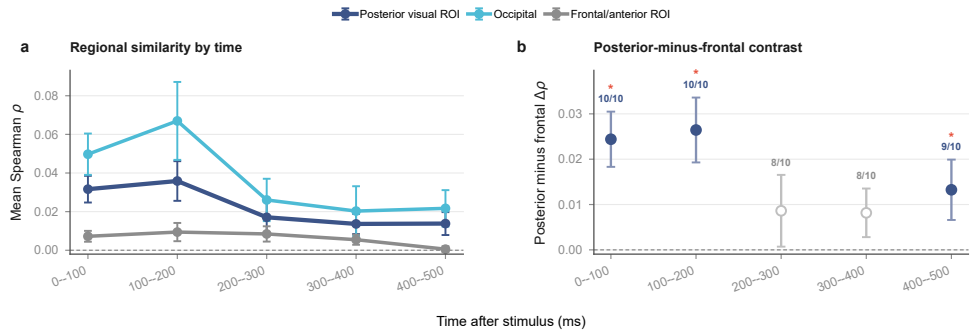

Figure S6: **Regional temporal summary of EEG–CLIP representational similarity. a**, Mean regional Spearman correlations across Sub01–Sub10 in five non-overlapping 100-ms windows. **b**, Posterior-minus-frontal contrasts with 95% t-intervals and positive-participant counts. The 0–100, 100–200, and 400–500 ms contrasts survive Holm correction over the five-window family.

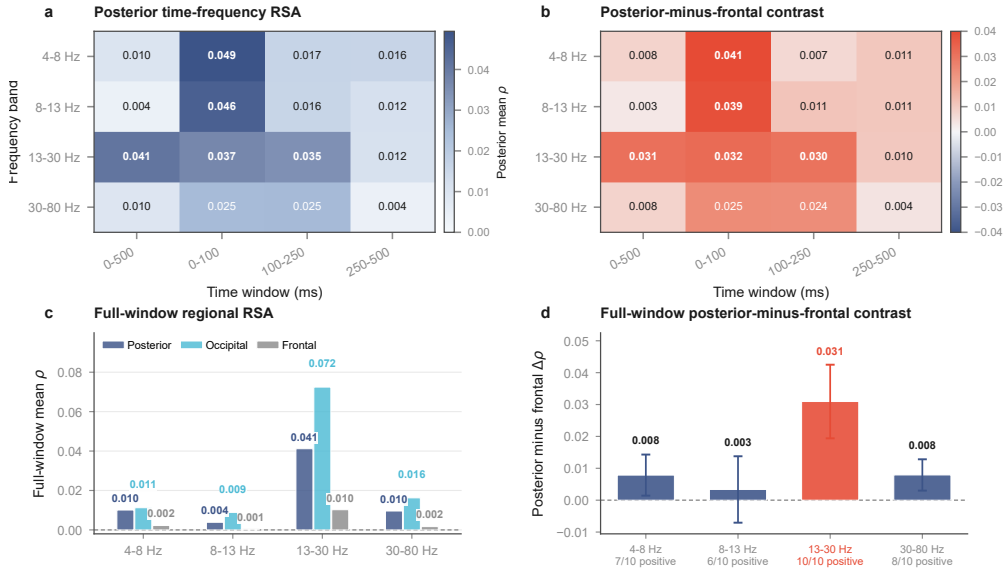

Figure S7: **Time–frequency EEG–CLIP RSA. a**, Posterior mean correlations for 4–8, 8–13, 13–30, and 30–80 Hz log-bandpower over the tested windows. **b**, Corresponding posterior-minus-frontal mean contrasts. **c**, Full-window posterior, occipital, and frontal means. **d**, Full-window posterior-minus-frontal contrasts with 95% t-intervals across Sub01–Sub10; labels report positive-participant counts. Twenty of 56 tested contrasts survive Benjamini–Hochberg false-discovery-rate correction. The 500-ms window provides 2-Hz native resolution; zero padding aligns the frequency grid across windows.

S18 TRAINING DURATION AND OPTIMIZATION STABILITY

Validation-loss minima within the 60-epoch cap determine the participant checkpoints used for the final five-seed evaluation.

Table S19: Validation-loss improvement with extended training. Relative improvement is measured from epoch 18 to the selected concept-disjoint validation minimum.

Participant	Epoch-18 loss	Selected loss	Selected epoch	Relative improvement
Sub01	1.0471	0.9671	60	7.64%
Sub02	1.2298	1.1601	52	5.67%
Sub03	1.1821	1.1110	59	6.02%
Sub04	1.1725	1.0987	60	6.29%
Sub05	1.3888	1.3413	53	3.42%
Sub06	1.0223	0.9366	60	8.39%
Sub07	1.0098	0.9236	60	8.53%
Sub08	0.8270	0.7446	60	9.97%
Sub09	1.1427	1.0772	58	5.73%
Sub10	0.8396	0.7582	60	9.69%
Mean	1.0862	1.0118	–	7.13%

Relative to epoch 18, the selected checkpoints reduce participant-wise validation loss by 7.13% on average, supporting the 60-epoch schedule used for the five-seed evaluation.

Table S20: Multi-seed stability and participant-level uncertainty for CORTIVA on THINGS-EEG2. Seed summaries average Sub01–Sub10; participant intervals follow five-seed averaging.

Statistic	Top-1 (%)	Top-5 (%)	Mean rank
Across-seed mean ± SD	73.53 ± 0.43	95.31 ± 0.25	1.842 ± 0.005
Across-seed range	73.10–74.00	95.05–95.70	1.837–1.849
Participant mean [95% CI]	73.53 [69.62, 77.33]	95.31 [93.81, 96.56]	1.842 [1.654, 2.083]

S19 THINGS-MEG TEMPORAL AND SENSOR ANALYSES

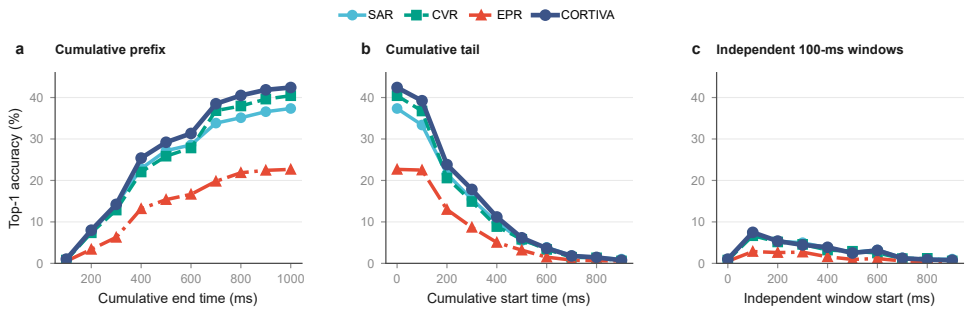

Figure S8: **THINGS-MEG temporal sensitivity. a**, Cumulative-prefix Top-1 accuracy as the input expands from 0 to 1000 ms. **b**, Cumulative-tail accuracy as progressively earlier samples are removed. **c**, Independent 100-ms windows. Curves show four-participant means after averaging six optimization seeds.

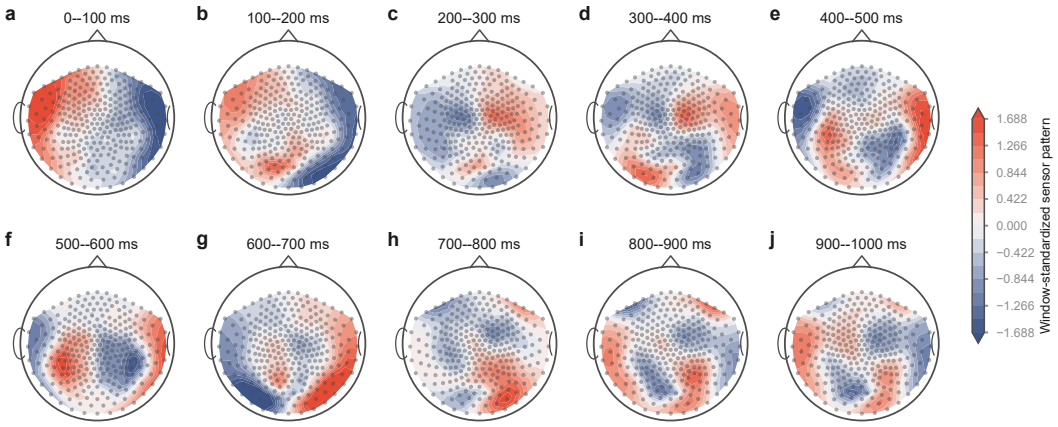

Figure S9: **THINGS-MEG sensor-pattern topography.** The ten panels show the mean within-participant, window-standardized sensor pattern across four participants and 271 magnetometers in non-overlapping 100-ms windows from 0–100 to 900–1000 ms. All panels share one zero-centered scale and use the recorded MEG sensor geometry.

S20 INDEPENDENT DINOV2 VALIDATION

DINOv2 ViT-B/14 (Oquab et al., 2024) is used only for external visual-geometry analyses. Its fixed features for the 200 held-out images support error-neighborhood analysis and EEG representational-similarity analysis. Error-neighborhood rates pool incorrect outcomes over five seeds within each participant before participant-level inference. Fixed 0–500-ms EEG tensors and repeated measurements keep this external validation separate from model fitting and score selection.

Table S21: Independent DINOv2 error-geometry and EEG representational-similarity validation. Intervals are 95% participant-level intervals.

Analysis	Estimate	95% CI	Reference or contrast	Participants; p
A. DINOv2 neighborhood of incorrect Top-1 predictions				
Top-10 error-neighbor rate	21.93%	18.80–25.06%	Random wrong: 5.03%; 4.36×	10/10; 0.00098
Top-25 error-neighbor rate	33.87%	29.44–38.30%	Random wrong: 12.56%; 2.70×	10/10; 0.00098
Top-50 error-neighbor rate	45.77%	41.17–50.38%	Random wrong: 25.13%; 1.82×	10/10; 0.00098
B. EEG–DINOv2 representational similarity				
Posterior, 0–500 ms	$\rho = 0.0316$	0.0258–0.0375	Spearman RSA against zero	10/10; 0.00098
Posterior minus frontal	$\Delta\rho = 0.0108$	0.0067–0.0149	Full-window regional contrast	10/10; 0.00098
Posterior late minus early	$\Delta\rho = 0.0159$	0.0072–0.0247	250–500 vs. 0–100 ms	9/10; 0.0107
Posterior split-half reliability	$\rho = 0.1505$	0.1066–0.1944	Repeated-measure split halves	–

S21 ADDITIONAL VISUAL ANALYSES



Figure S10: **Participant-level consistency of matched route-removal effects. a**, Top-1 loss after removing each route. **b**, Top-5 loss under the same matched retraining protocol. Each cell is the full CORTIVA participant mean minus the corresponding ablated participant mean after averaging five optimization seeds; positive values indicate a retrieval loss after route removal.

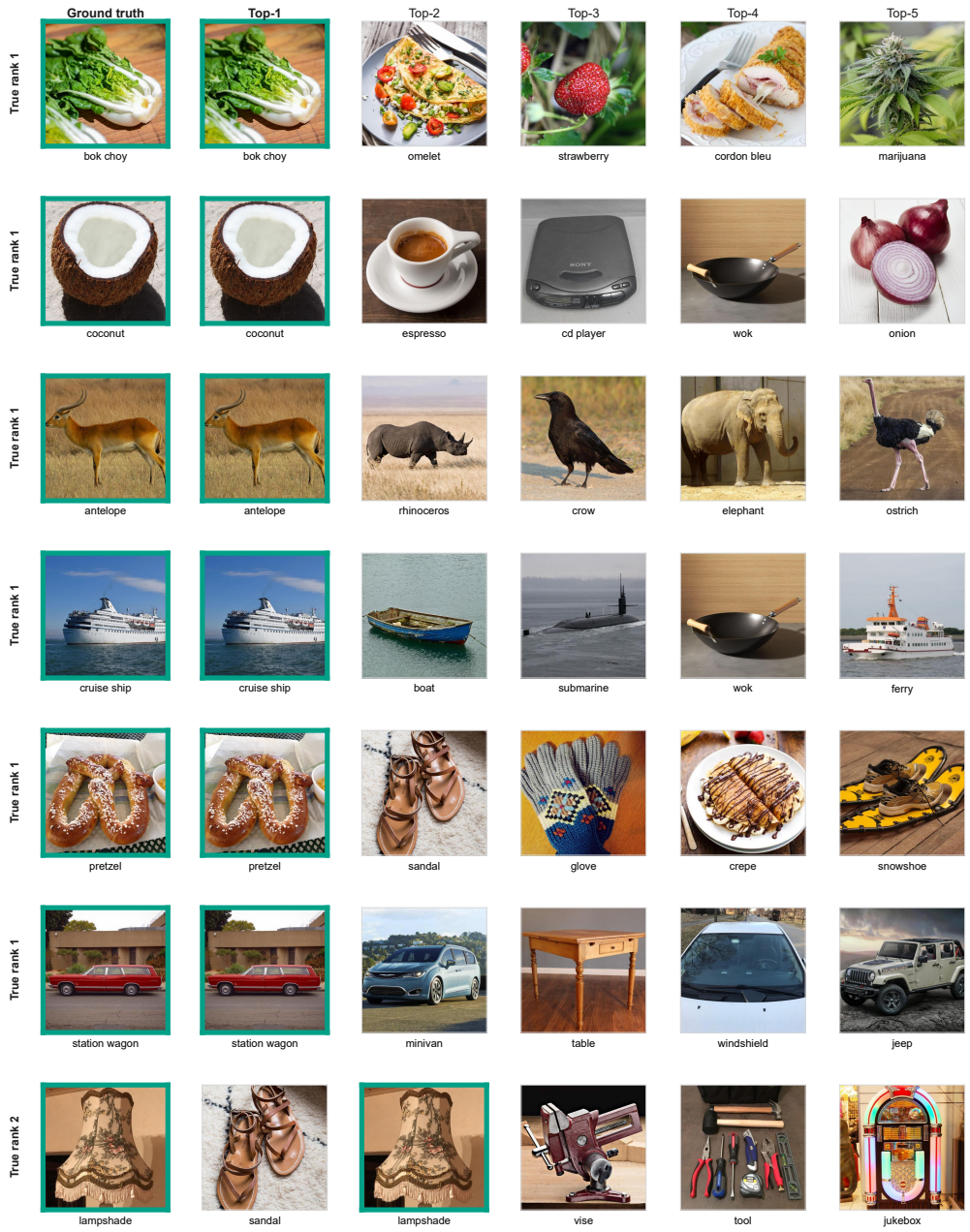

Figure S11: **Representative retrievals.** Six Top-1 outcomes and one Top-2 outcome from Sub10 under seed 2026, selected to span score quantiles and semantic categories. Each row shows the ground-truth image followed by the five highest-scoring candidates; the teal border marks the correct candidate. Stimulus source: the THINGS object-concept image database (Hebart et al., 2019).

Repeated five-fold category readout reaches 40.1–42.3% balanced accuracy across routes against a 20% balanced chance level. The corresponding cosine silhouette coefficients (-0.17 to -0.15) characterize a diffuse, distributed geometry alongside strong linear category readout. Supplementary Fig. S12 visualizes this coexistence across all 200 test queries.

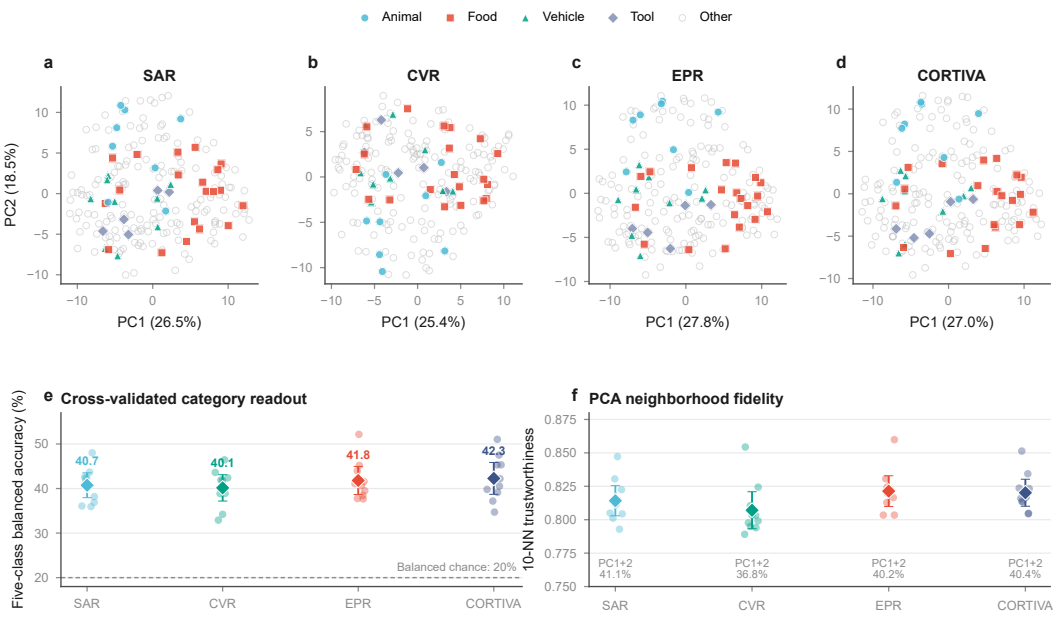

Figure S12: **Score-space projection and broad-category readout. a–d**, Principal-component projections of SAR, CVR, EPR, and CORTIVA score vectors after row standardization within participant and averaging over Sub01–Sub10. Each point is one held-out query; all 200 concepts are shown, and category labels are not used to fit PCA. **e**, Five-class balanced accuracy from five repeats of stratified five-fold cross-validation within each participant. The training-fold pipeline standardizes features, retains 95% variance by PCA, and fits class-balanced logistic regression. **f**, Ten-neighbor trustworthiness of each participant-specific two-dimensional PCA; labels give the participant-mean variance explained by PC1 and PC2. In **e,f**, pale points denote participants and diamonds with bars show participant means and 95% t-intervals.