BiSURE: Toward a Unified Neural Representation for Visual Encoding and Decoding

Anonymous authors
Paper under double-blind review

Abstract

If visual encoding and decoding truly meet in the same interchangeable latent representation Z, then a neural response from one participant should be readable, through Z, as the response of another. We build and test such a representation. **BiSURE** (**Bi**directional **SU**fficient **RE**presentation) maps images and EEG, MEG, and fMRI responses into a common latent, from which the stimulus can be retrieved and every participant's measured response predicted, while retaining private interfaces for each instrument and participant. A single consistency term ties the image-derived and brain-derived routes, making their latents interchangeable. A response from one participant, passed through Z to another participant's response head, predicts that participant's response at 87–96% of the level reached from the image itself, with no stimulus and no parameter fitted for the pair, and more accurately than a linear map fitted for that pair on paired data. The same latent carries EEG responses into MEG at 84%, despite no paired EEG–MEG recordings, and transmits more than the model's own decoded CLIP features. Without the latent tie, retrieval remains essentially unchanged, yet the two routes become unrelated and translation falls to at most 6%; the tie also improves encoding by 15–31%. The same model matches or exceeds the strongest specialist in retrieval on every instrument, predicts every participant's responses above an established encoding baseline, and reconstructs images from the frozen latent.

1 Introduction

Neural encoding maps a visual stimulus S to a measured neural response X; decoding maps X back to stimulus information (Naselaris et al., 2011). The two are usually learned by separate models, one instrument at a time: recent systems retrieve or reconstruct images from EEG (Song et al., 2024; Li et al., 2024; Zhang et al., 2026), MEG (Benchetrit et al., 2024), or fMRI (Scotti et al., 2024), while encoding is evaluated by a different family of estimators. A representation that both directions share would do more than save a model. It would make responses from different people and different instruments expressible in one another's terms. Aligning brains has so far required fitting participant-specific transformations on their responses to shared stimuli (Haxby et al., 2011; Chen et al., 2015).

A pretrained visual space does not by itself provide that representation. CLIP is an effective target for retrieval and an interface to image generators (Radford et al., 2021), but its geometry was learned from image-text pairs, not by predicting neural measurements. We therefore use the visual space as supervision rather than as the definition of the neural representation. Recent tri-modal systems share parameters or a CLIP target (Li et al., 2025; Ferrante et al., 2025), but none predicts measured responses (Table 1), and sharing parameters does not by itself make an image-derived and a brain-derived latent the same object. We instead learn one latent that both an image and a response enter (Figure 1).

Our learning principle is that the two directions should meet in a common latent:

$$S \longrightarrow Z \longrightarrow X, \qquad X \longrightarrow Z \longrightarrow S_{\text{recoverable}}. \tag{1}$$

We call the target property *approximate bidirectional sufficiency*. The latent should retain the stimulus information required to predict the response and the response information from which the stimulus can be recovered:

$$p(X \mid S, m, s) \approx p(X \mid Z, m, s), \qquad p(S \mid X) \approx p(S \mid Z). \tag{2}$$

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.

Table 1: Scope of representative systems under our definitions of decoding (neural → stimulus information) and encoding (stimulus → measured neural response). ✓: evaluated; ○: a related evaluation that does not meet the definition; –: not reported. neural-CLIP's encoding retrieves the nearest stored neural embedding in CLIP space; ReAlnet compares its representational geometry with fMRI; SynBrain's synthesized fMRI trains a separate decoder. †Our implementation on frozen CLIP features (Radford et al., 2021).

	EEG		MEG		fMRI	
method	Dec	Enc	Dec	Enc	Dec	Enc
NICE / ATM (Song et al., 2024; Li et al., 2024)	✓	–	–	–	–	–
CognitionCapturer (Zhang et al., 2025)	✓	–	–	–	–	–
CogCapPro (Zhang et al., 2026)	✓	–	✓	–	–	–
Benchetrit et al. (Benchetrit et al., 2024)	–	–	✓	–	–	–
MindEye2 (Scotti et al., 2024)	–	–	–	–	✓	–
neural-CLIP (Ferrante et al., 2025)	✓	○	✓	○	✓	○
BrainFLORA (Li et al., 2025)	✓	–	✓	–	✓	–
BERG (Gifford et al., 2025)	–	✓	–	✓	–	✓
CLIP ridge†	–	✓	–	✓	–	✓
ReAlnet (Lu et al., 2026)	–	✓	–	–	–	○
SynBrain (Mai et al., 2026)	–	–	–	–	○	✓
BiSURE (ours)	✓	✓	✓	✓	✓	✓

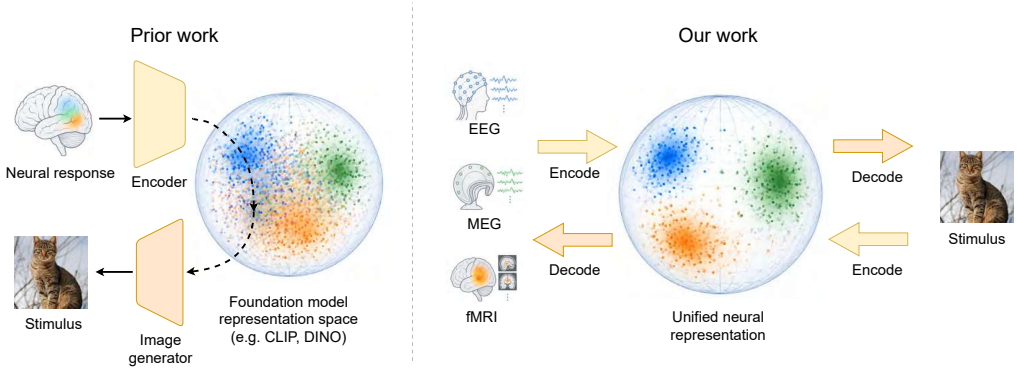

Figure 1: **From separate pipelines to one stimulus-response representation.** Prior systems typically treat encoding and decoding as separate mappings around an external visual space. We instead learn one neuro-representational latent Z that is entered from either an image or an EEG, MEG, or fMRI response. CLIP anchors the visual interface; it is not itself the neural representation. Modality-specific interfaces retain how each instrument measures the shared content.

Equation 2 turns encoding and decoding into complementary tests of one latent rather than two unrelated benchmarks, and its operational consequence is what we measure: whether that latent is interchangeable across routes and participants (Section 3.3).

Heterogeneous instruments should not be forced to observe Z identically. We posit

$$X_{m,s} = h_{m,s}(Z) + \epsilon_{m,s}, \tag{3}$$

where m and s index modality and participant. EEG, MEG, and fMRI can share what is represented while retaining different temporal, spatial, and participant-specific observation functions. In short, *we unify what is represented, not how it is measured*.

One term ties the two routes: the image-derived latent is pulled toward the brain-derived latent of the same stimulus. Our contributions are:

- **Stimulus-blind translation between brains.** A response encoded from one participant drives another participant's response head at 87 to 96% of the image-driven level, and EEG drives MEG at

84%, with no stimulus, no pair-specific parameter and no paired EEG-MEG recording. It exceeds a ridge map fitted for each pair on paired data in 559 of 564 pair-seed cells.

- **The mechanism.** The latent tie is what makes the image-derived and brain-derived latents interchangeable: without it the model still performs both tasks, but no head can read the other route's latent. The tie also raises encoding by 15 to 31%.
- **One model at specialist level in both directions.** Retrieval is level with the strongest specialist on EEG and exceeds it by 6.4 and 13.9 top-1 points on MEG and fMRI; encoding exceeds BERG for all 17 participants at every seed; the frozen latent reconstructs images that lead every evaluated system on CLIP-H for EEG and fMRI.

2 Related Work

Neural encoding and decoding. On EEG, NICE aligns neural and image encoders contrastively (Song et al., 2024; Oord et al., 2018); ATM adds retrieval and diffusion-based reconstruction (Li et al., 2024); and CognitionCapturerPro combines image, depth, edge, and text supervision (Zhang et al., 2026). On MEG, contrastive and regression objectives recover visual features at high temporal resolution (Benchetrit et al., 2024). MindEye2 uses cross-participant fMRI alignment and an unCLIP generator (Scotti et al., 2024). These methods establish strong modality-specific decoders; none predicts measured responses.

Shared models across neural measurements. Recent work has begun to connect measurements and participants. BrainFLORA trains modality-specific adapters and task decoders over EEG, MEG, and fMRI (Li et al., 2025). Ferrante et al. align the same three modalities with CLIP (Ferrante et al., 2025); their encoding and modality conversion retrieve the nearest stored neural embedding in that space rather than predicting measured responses. Task coverage alone does not establish that image-derived and brain-derived latents coincide; we make that distinction measurable and ablate the only term that couples the two directions. Route swaps and translation are defined only for a model that has both an image route and a response head, which none of these systems has (Table 1).

Aligning participants. Hyperalignment maps each participant's responses into a common space through transformations fitted on responses to shared stimuli (Haxby et al., 2011), and the shared response model learns such a space jointly with participant-specific maps (Chen et al., 2015). Both fit their maps on every participant's responses to the same stimuli and have been applied within one instrument. Our translation fits nothing for a pair of participants, crosses from EEG to MEG without paired recordings, and uses a latent that also predicts responses from images.

3 Two Routes into One Latent, Coupled by One Term

3.1 Two routes into one latent

Let S be an image and $X_{m,s}$ the response recorded with modality $m \in \{\mathrm{EEG}, \mathrm{MEG}, \mathrm{fMRI}\}$ from participant s. A measurement interface T_m and shared neural encoder E_θ produce a brain-derived latent

$$Z_{\mathrm{brain}} = E_\theta(T_m(X_{m,s})) \in \mathbb{R}^{L \times d}. \tag{4}$$

A frozen visual encoder v and image map g_ϕ produce the corresponding image-derived latent

$$Z_{\mathrm{stim}} = g_\phi(v(S), m, s) \in \mathbb{R}^{L \times d}. \tag{5}$$

The two routes end in arrays of the same shape and coordinate system. Conditioning g_ϕ on the requested modality and participant accounts for the response to be predicted; it does not create a separate neural trunk.

Matching shapes are not enough: the routes can still learn unrelated codes. We remove that freedom with the conceptual objective

$$\mathcal{L} = \mathcal{L}_{\mathrm{decode}}(R(Z_{\mathrm{brain}}), v(S)) + \mathcal{L}_{\mathrm{encode}}(D(Z_{\mathrm{stim}}; q_{m,s}), X_{m,s}) + \lambda_{\mathrm{tie}} d(Z_{\mathrm{stim}}, \mathrm{sg}[Z_{\mathrm{brain}}]), \tag{6}$$

where the retrieval head R supports brain-to-image retrieval, the response head D predicts requested neural coordinates $q_{m,s}$, and sg stops gradients through the neural route. The third term is the only

term that explicitly makes the inference and generative directions identify the same latent. D is trained only on Z_{stim} and R only on Z_{brain}, so neither head sees the other route's latent during training. The exact modality-specific forms are given in Appendix B.

3.2 Shared representation, private measurement interfaces

A shared Perceiver maps every instrument to 256 latent tokens of width 256 (Jaegle et al., 2021). The image-to-Z map, latent shape, spatial visual target, and coordinate decoder are also shared. Subject identity cannot enter the retrieval readout and appears only on the response-prediction side. **CLIP is the visual anchor and interface, not the neural representation itself**: Z lies upstream of the CLIP readout and must also predict measured neural activity.

The boundary remains measurement-specific. EEG and MEG require temporal sensor tokenizers, whereas fMRI requires a dense subject-specific voxel interface. Participant residual heads restore per-unit degrees of freedom after the shared coordinate decoder. Alignment and tie losses also differ where the instruments demand different statistical contracts. Appendix B gives the implementation and the matched ablations that establish this boundary.

3.3 A shared latent must be interchangeable, not only aligned

Shared parameters are not evidence that paired inputs reach the same state. We therefore define

$$\Delta_Z = \mathbb{E}_{(S,X)}[\cos(Z_{\text{stim}}, Z_{\text{brain}})] - \mathbb{E}_{(S,X')}[\cos(Z_{\text{stim}}, Z'_{\text{brain}})]\,, \tag{7}$$

where cosine is evaluated along the feature dimension for each latent token and then averaged over pairs and tokens; X' is shuffled across stimuli. The subtraction removes common directions such as modality or participant means. $\Delta_Z = 0$ means that matching a stimulus to its response provides no additional agreement; positive values identify a stimulus-specific shared component. Agreement is necessary but not sufficient: a positive Δ_Z does not show that a head can use the other route's latent. Because Equation 6 trains each head on one route only, swapping routes at test time gives two stricter tests,

$$\text{swap}_R = \text{ret}\big(R(c_m\, Z_{\text{stim}})\big), \qquad \text{swap}_D = r\big(D(E_\theta(T_m(X^{(1)}_{m,s})); q_{m,s}),\, X^{(2)}_{m,s}\big), \tag{8}$$

where ret is retrieval in the test gallery and c_m is one scalar per modality, the ratio of the mean token norms of Z_{brain} and Z_{stim} on training data. The tie aligns the direction of the two latents and leaves their norm free; c_m restores it for fMRI, whose retrieval head does not normalise its input, and changes EEG and MEG top-1 by at most .0005. $X^{(1)}, X^{(2)}$ are the averages of two disjoint halves of each test stimulus's trials, so the encoder's input and its target are independent measurements. The strictest test removes the stimulus altogether. The response of participant a is encoded and read out through the response head of participant b,

$$\widehat{X}_{b\leftarrow a} = D\big(E_\theta(T_{m_a}(X_{m_a,a})); q_{m_b,b}\big), \tag{9}$$

with no parameter fitted for the pair. We score it by per-unit correlation with b's measured response, as a fraction of b's image-driven prediction $D(Z_{\text{stim}}; q_{m_b,b})$.

4 Experiments

4.1 Data

We use the THINGS image collection (Hebart et al., 2019), THINGS-EEG2 (Gifford et al., 2022), and the THINGS MEG/fMRI release (Hebart et al., 2023). The three datasets contain no overlapping participants. Joint training creates a split-linkage risk if each release's native partition is used independently. We therefore rebuild MEG with the same concept-level partition used for EEG (1,654 training and 200 held-out concepts), following BrainFLORA (Li et al., 2025); fMRI retains its released image split (8,640 training and 100 test images). Batches are modality-homogeneous and never construct paired observations across instruments. The image collections overlap across datasets; single-modality models, which never see another dataset's images, retrieve as well as the joint model or better (Section 5.4), so the overlap does not account for the scores. Appendix Table 12 records the actual estimator in each test set, which matters because the MEG prototype averages different images of the same concept rather than repetitions of one image.

4.2 Training and evaluation protocol

We train one 268.11M-parameter model jointly on homogeneous EEG, MEG, and fMRI batches, with three independent seeds. Every reported checkpoint is the final checkpoint after 60 training epochs. None of the three datasets defines a validation split, so every model, ours and the baselines, is scored at a fixed final checkpoint rather than one selected on test data. Our retrieval, encoding, agreement and swap results are means over the three seeds (standard deviations in Appendix Table 13); baselines are single released or retrained checkpoints. Figures and reconstruction use seed 2. Architecture and optimisation details are in Appendix A.

We report four complementary measurements. **Retrieval** matches each brain response against its gallery and is reported by Top-k accuracy. Information gain (bits), which also credits the rank of the correct image, is defined in Appendix C and reported for every gallery size in Appendix Tables 8 and 9. **Encoding** is the Pearson correlation across test images for every neural unit, averaged within participant and then across participants. **Latent agreement, route swaps and translation** are Equations 7–9. **Reconstruction** reads the existing retrieval head into the frozen unCLIP-SDXL generator used by MindEye2 (Scotti et al., 2024); no neural parameter is retrained (Section 5.5).

Every retrieval and encoding result is recomputed under one evaluator. We compare only within a modality because gallery size, participant count, and prototype construction differ across instruments.

5 Results

5.1 A latent tie makes the image-derived and brain-derived latents interchangeable

Sharing an encoder and training both directions does not by itself make the two routes meet. We therefore train the same joint model with $\lambda_{\text{tie}} = 0$, a change of one term, at the same three seeds. Its retrieval is unchanged: top-1 is .5988, .1796 and .8189 on EEG, MEG and fMRI, against .5942, .1792 and .8222 with the tie (Table 2), and information gain is 5.53, 2.55 and 5.84 bits against 5.61, 2.56 and 5.84. The tie aligns the two latents on training pairs, but that alone does not let a head trained on one route read the other: R never sees an image-derived latent, D never sees a brain-derived one, and every test below uses held-out test responses. Figure 2 applies three tests of increasing strictness.

Agreement. The tie raises Δ_Z from $|\Delta_Z| \leq .026$ in all nine untied seed-modality cells, where matched image-response pairs are no closer than shuffled pairs, to .344, .212 and .516 on EEG, MEG and fMRI (Figure 2a).

Swapped retrieval head. R is trained only on brain-derived latents. Applied to image-derived latents (Equation 8), it retrieves the correct image at top-1 .52, .11 and 1.00 in the full gallery (chance .005, .005 and .010; Figure 2b). The image-derived latent carries no measurement noise, so the swapped head can exceed the native one, as it does on fMRI. Without the tie it stays near chance (top-1 at most .019, .014 and .023 at any seed, with the same scalar fitted for that model). At every seed the tied value is at least $4\times$ the largest untied value (EEG ranges from .32 to .83 across seeds).

Swapped response head. D is trained only on image-derived latents. Applied to latents encoded from half of the test trials, it predicts the held-out half at per-unit $r = .093$, .090 and .143, which is 114%, 86% and 92% of its image-driven prediction of the same target and above a model-free baseline that copies one half to the other (Figure 2c). Without the tie, $r \leq .009$ at every seed.

The tie is therefore what lets each head read the other route's latent: without it the model performs both mappings, but neither head can. The tie also raises encoding, from .0862 to .1012 on EEG, .1087 to .1254 on MEG and .1553 to .2029 on fMRI (per-unit r; $+17\%$, $+15\%$ and $+31\%$), and it is what lifts fMRI above BERG (.1673; Table 3).

5.2 One participant's response predicts another's, without the stimulus

If Z is one representation, a response encoded from one participant should be readable by another participant's response head without the stimulus (Equation 9). No parameter is fitted for the pair, and the stimulus is never used. Averaged over all ordered participant pairs and three seeds, the translated response reaches 91% (EEG→EEG, 90 pairs), 87% (MEG→MEG, 12 pairs) and 96%

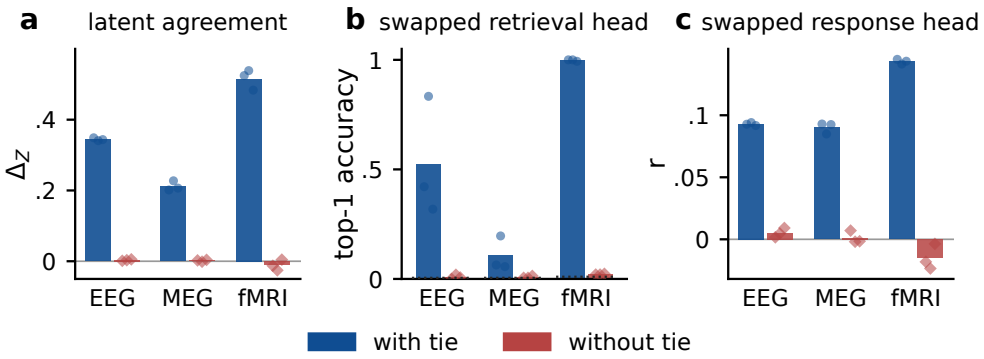

Figure 2: **The two directions meet in one latent only when they are tied.** Blue: BiSURE; red: the identical model trained without the latent tie. Bars are means over three seeds; dots are individual seeds. **(a)** Latent agreement Δ_Z (Eq. 7). Without the tie it is zero at every seed, although both directions are trained. **(b)** Swapped retrieval head: the head trained only on brain-derived latents is applied to image-derived latents, rescaled by one norm ratio fitted on training data (top-1 accuracy on the full test gallery; dotted: chance). **(c)** Swapped response head: the head trained only on image-derived latents is applied to latents encoded from half of the test trials and scored against the held-out other half (per-unit correlation). With the tie both heads work on the other route's latent; without it both are at or near chance.

(fMRI→fMRI, 6 pairs) of the target participant's image-driven prediction. The latent also crosses instruments: EEG→MEG translation reaches 84%, and MEG→EEG 43% (Figure 3c–f; Appendix Figure 6).

The zero-parameter route through Z outperforms a linear map fitted for the pair on paired data, the kind of fitted alignment that hyperalignment relies on (Haxby et al., 2011). Ridge regression from the source's to the target's training responses to the images both saw reaches 62%, 50% and 79% within EEG, MEG and fMRI, and 59% and 26% from EEG to MEG and back; Z is higher in 270/270, 36/36, 18/18, 117/120 and 118/120 pair-seed cells (Appendix F). Z also carries more than a semantic summary of the stimulus: routing the source response through the CLIP route instead (its decoded CLIP features, mapped to the target) gives 82%, 81% and 70% on EEG→EEG, MEG→MEG and EEG→MEG, below Z in 269 of 270 EEG→EEG and 103 of 120 EEG→MEG pair-seed cells. The translation also keeps the temporal and spatial structure of the target: it follows the target's time course (Figure 3) and its sensor topography (Appendix Figures 7 and 8). Averaged over sensors and 50–400 ms, EEG and MEG translation reach 90% and 88% of the image-driven prediction, with the occipital (EEG) and posterior (MEG) distribution of the split-half noise ceiling (Appendix Figure 9). The information is stimulus-specific: pairing the source response with a different stimulus's target gives $|r| \leq .002$ in every direction (seed 2). On fMRI it reproduces the target participant's cortical pattern (Figure 4). Without the tie, translation falls to at most 6% of the image-driven prediction, and none of the 564 pair-seed cells exceeds the CLIP route.

5.3 The shared latent costs neither direction anything

Tables 2 and 3 evaluate the two directions, with every encoding metric in Appendix Table 10; every BiSURE row is the same model (mean over three seeds). In retrieval, BiSURE matches the strongest EEG specialist, CogCapPro, which fuses several modality-specific branches, on top-1 (.5942 vs .5945) and exceeds it on top-5 (.9000 vs .8905); it exceeds the strongest MEG and fMRI specialists by 6.4 and 13.9 points of top-1. In encoding, its per-unit correlations exceed BERG by 31%, 121% and 21%, and every participant clears its BERG value at every seed (51/51 participant-seed cells; Appendix Table 11). The untied model retrieves as well but encodes worse in every modality (Section 5.1).

The gains are largest where encoding is hardest. On MEG, whose training targets are single trials, every participant reaches 1.8 to 2.7 times its BERG value. On EEG and MEG, the image-driven prediction reaches 58% and 49% of the split-half noise ceiling at 100–200 ms after onset (Figure 3a,b),

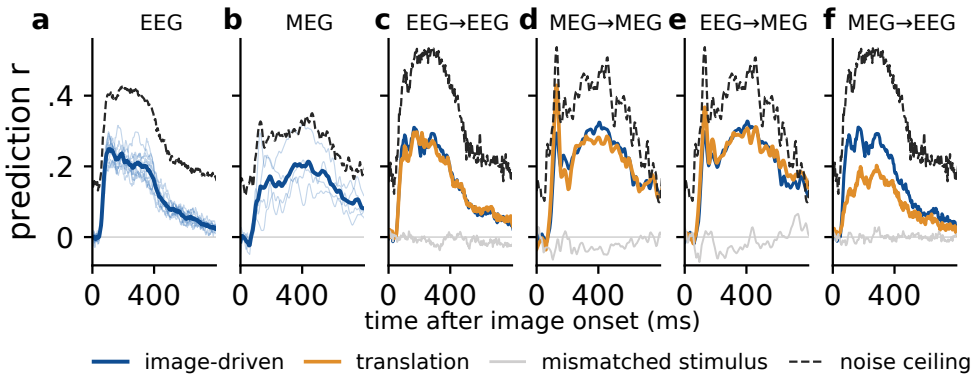

Figure 3: **The same latent predicts time-resolved EEG and MEG responses, from an image or from another brain.** Per-unit correlation over test stimuli, averaged over sensors. Dashed: split-half noise ceiling. **(a, b)** From the image: thin lines are participants (10 EEG, 4 MEG); the thick line is their mean. At 100–200 ms the model reaches 58% (EEG) and 49% (MEG) of the ceiling. **(c–f)** Stimulus-blind translation of another participant's measured response through Z, within EEG (c), within MEG (d) and across instruments (e, f). Orange: translation; blue: the target's image-driven prediction; grey: mismatched stimulus. Each panel shows the best participant pair of its direction.

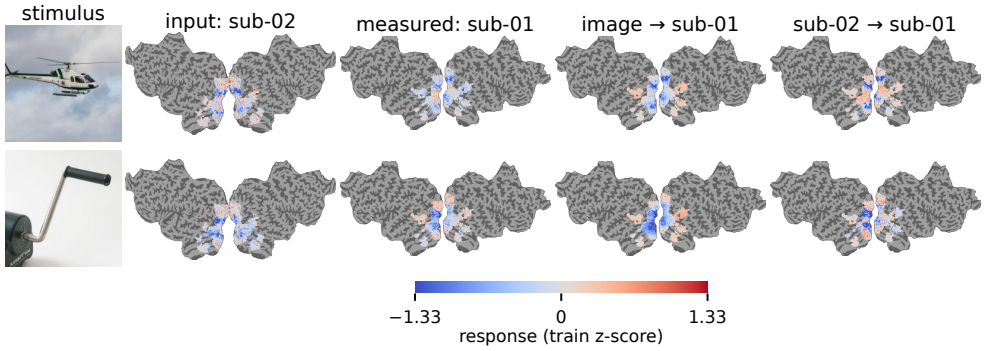

Figure 4: **Translating one participant's fMRI response into another's.** The model receives only sub-02's measured response (second column), encodes it into Z and reads it out through sub-01's own response head (last column). Third and fourth columns: sub-01's measured response and its prediction from the image. Over all 100 test images, sub-02's response predicts sub-01 as well as the image does (per-voxel $r = .216$ in both cases). Both stimuli are among the three best of this pair; the median image is predicted equally well by both routes (pattern $r = .30$). Each prediction column uses one display gain over the test set.

so much of the remaining gap is measurement noise rather than missing stimulus information. These comparisons use one protocol for every row: each method is rescored by the same evaluator at a final or released checkpoint. CogCapPro's released alignment stage selects its epoch on the test loader (Appendix D), whereas no BiSURE checkpoint is chosen with test data.

5.4 Joint training raises encoding in every modality

Against the same model trained on one modality (three seeds each; Appendix Figure 10), joint training raises encoding in all three modalities ($+.005$, $+.042$ and $+.008$; MEG $+51\%$). It leaves MEG and fMRI retrieval unchanged (top-1 $+.002$ and $-.002$). On EEG, top-1 is .026 lower (.602 vs .628), but information gain, which scores the rank of the correct image rather than only first place, is unchanged ($5.64 \pm .10$ vs $5.73 \pm .16$ bits, within one seed standard deviation). Joint training helps most where data are scarcest: MEG, with four participants and single-trial targets, gains 51% in encoding and $1.5\times$ in latent agreement (Appendix Figure 10).

Table 2: **Decoding/retrieval on the complete test gallery** (200-way for EEG and MEG, 100-way for fMRI) under the unified final-checkpoint protocol (↑). Our row is the mean over three seeds (standard deviations in Appendix Table 13); baselines are single checkpoints. Bold and underline mark best and second best within a modality; "–" marks a modality the method does not report. CognitionCapturer has no fused embedding, so its three branches are listed; its own combination rule (intersecting the branches' top-5 lists) gives top-1 .2445, the same as its image branch. Every candidate count (2- to 100-way) is in Appendix Tables 8 and 9.

	EEG (200-way)			MEG (200-way)			fMRI (100-way)		
method	Top-1	Top-5	bits	Top-1	Top-5	bits	Top-1	Top-5	bits
BiSURE (ours)	<u>.5942</u>	**.9000**	<u>5.610</u>	**.1792**	**.4483**	**2.557**	**.8222**	**.9711**	**5.836**
CogCapPro, fusion	**.5945**	<u>.8905</u>	**5.660**	<u>.1150</u>	<u>.3200</u>	<u>1.778</u>	–	–	–
CogCapPro, image	.4735	.7865	4.846	.0950	.2725	1.497	–	–	–
MindEye2	–	–	–	–	–	–	<u>.6833</u>	<u>.8833</u>	<u>5.009</u>
BrainFLORA, unimodal	.2675	.5865	3.354	.0813	.2462	1.307	.2567	.5367	2.374
BrainFLORA, multimodal	.2280	.5155	3.012	.0912	.2675	1.470	.2000	.5300	2.300
CogCapturer, image	.2430	.5305	3.055	–	–	–	–	–	–
CogCapturer, depth	.1695	.4345	2.488	–	–	–	–	–	–
CogCapturer, text	.0955	.2740	1.539	–	–	–	–	–	–
ATM	.2275	.5185	3.010	–	–	–	–	–	–
neural-CLIP	.0260	.0965	.270	.0537	.2075	1.084	.1167	.2700	.968

Table 3: **Encoding of measured neural responses** (↑). Per-unit r (primary) correlates prediction and measurement across test images for every unit; centered r removes each unit's mean and correlates per stimulus. Participant-equal means over all units; our row is the mean over three seeds (standard deviations in Appendix Table 13). Bold and underline: best and second best within a modality. The full metric family is in Appendix Table 10.

	EEG		MEG		fMRI	
method	per-unit r	centered r	per-unit r	centered r	per-unit r	centered r
BiSURE (ours)	**.1012**	**.2939**	**.1254**	**.1539**	**.2029**	**.2753**
BERG	<u>.0774</u>	<u>.1366</u>	<u>.0568</u>	.0609	<u>.1673</u>	<u>.2348</u>
CLIP ridge	.0723	.1299	.0566	<u>.0625</u>	.1581	.2199
ReAlnet	.0063	.0121	–	–	–	–
SynBrain	–	–	–	–	.1011	.1496

5.5 THE SAME LATENT SUPPORTS IMAGE RECONSTRUCTION

The frozen latent also reconstructs images. BiSURE leads every evaluated external system on all six semantic columns on EEG and fMRI, including CLIP-H (.8427 vs CogCapPro .8310; .8616 vs MindEye2 .8130). On MEG it leads on AlexNet, EfficientNet-B1, DINOv2, PixCorr and SSIM; CogCapPro, which renders from three visual conditions (image, depth and edge) where ours uses one, is higher on Inception and CLIP-H (Appendix D). Our single condition is read from the brain-derived latent, so these scores measure what Z carries rather than what added depth or edge predictors contribute. On PixCorr, BiSURE also exceeds a render of the training-set mean image by $.068$ on EEG and $.038$ on fMRI (Appendix D).

Reconstruction reuses the frozen model in two steps. First, the visual readout of Z_{brain} is mapped into the generator's image-embedding space by a ridge map fitted on training data and a frozen image-only bridge. The response determines the direction of this embedding, but not its full variance, so we mix the predicted direction in equal parts with the deviation of a sample from a learned diffusion prior conditioned on it. The prior follows MindEye2 (Scotti et al., 2024); it is trained once on the fMRI training set and applied unchanged to EEG and MEG. Second, the frozen unCLIP-SDXL generator renders the embedding. No neural parameter is retrained, and the images in Figure 5 are the ones scored in Table 4.

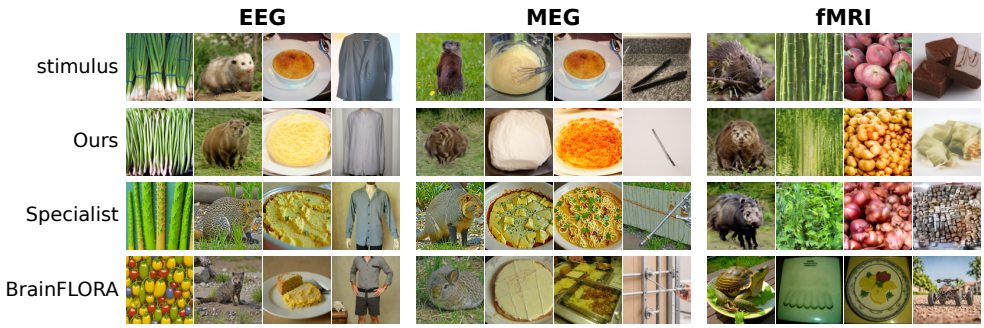

Figure 5: **Reconstruction from the same frozen model.** Four test stimuli per modality (EEG sub-08, MEG P2, fMRI sub-01) whose reconstruction by our method is closest to the stimulus in CLIP-H, at most one per THINGS category. Rows: stimulus; ours; the strongest specialist (CogCapPro for EEG/MEG, MindEye2 for fMRI); BrainFLORA, the only other tri-modal model. Random stimuli: Appendix Figures 11, 12 and 13.

Table 4: **Post-hoc image reconstruction from the same frozen model** (seed 2). Six network columns are two-way identification accuracies; PixCorr and SSIM are low-level similarities (all ↑). Bold/underline: best/second within a modality.

method	Alex-2	Alex-5	Incep	CLIP-H	EffB1	DINOv2	PixCorr	SSIM
EEG								
BiSURE (ours)	**.8835**	**.9163**	**.7529**	**.8427**	**.8229**	**.8074**	**.2272**	.3360
ATM	.6501	.7191	.6228	.7062	.6426	.6344	.1039	.3171
CogCapPro, encoder	.7154	.8014	.7260	.8070	.7567	.7230	.1317	.2701
CogCapPro, align	.7256	.8260	.7398	.8310	.7831	.7383	.1320	.2709
CognitionCapturer, image	.6794	.7382	.5956	.6978	.6395	.6127	.1086	.2958
CognitionCapturer, text	.5802	.6309	.5534	.6049	.5478	.5716	.0932	.2748
CognitionCapturer, depth	.6087	.6732	.5828	.6307	.5876	.5966	.0979	**.3621**
BrainFLORA, multimodal	.6682	.7226	.6072	.6820	.6313	.6186	.0856	.2914
MEG								
BiSURE (ours)	**.7062**	**.7623**	.6451	.7567	**.7033**	**.6970**	**.1169**	**.2682**
CogCapPro, encoder	.6113	.7116	**.6565**	**.7834**	.6702	.6849	.0858	.1906
BrainFLORA, multimodal	.5720	.6364	.5987	.6443	.5938	.5829	.0660	.2225
fMRI								
BiSURE (ours)	**.8839**	**.9241**	**.7808**	**.8616**	**.8568**	**.8231**	**.1683**	**.3651**
MindEye2	.8251	.8488	.7084	.8130	.7539	.7308	.1092	.2862
BrainFLORA, multimodal	.6031	.6326	.5659	.6148	.5834	.5741	.0625	.2869

6 Discussion and Conclusion

Task coverage is not a shared representation. With a shared encoder, matched latent shapes and both objectives, the image route and the brain route still learn unrelated codes; one term makes them one code, and that code is what lets responses move between participants and instruments. Claims that a neural model is unified can be tested the same way, by swapping routes and translating between participants, rather than by counting the tasks it performs.

We unify what is represented, not how it is measured: making every head and loss identical charges one modality for another's convenience, and private tokenizers and residual heads are where each instrument's view of the shared latent belongs (Appendix B). The natural next test is translation into participants the model has never seen.

One latent tie turns encoding and decoding across EEG, MEG and fMRI into one representation that is at specialist level in both directions, and in that representation a response from one brain can be read as the response of another.

AI USE STATEMENT

We used generative AI tools to assist with writing and editing, literature search and discovery, and code development for training, evaluation, and visualization. We did not use generative AI for research ideation or to generate synthetic data or results. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the work.

ETHICS STATEMENT

This work uses only publicly released, de-identified recordings: THINGS-EEG2 (Gifford et al., 2022) and the THINGS MEG and fMRI release (Hebart et al., 2023), collected by their original authors with informed consent and ethical approval. We collected no new data and used the datasets under their terms of use. Models that decode and translate brain responses raise questions of mental privacy. The models studied here require a cooperating participant, hours of that participant's recordings to train a participant-specific response head, and responses to a fixed image set; they decode only which image, among a set, was viewed and do not read unconstrained thought. Translation between participants operates on recordings that both participants provided in the same controlled experiment. We see no other ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

Appendix A gives the complete architecture (Table 5), objective weights and batch sizes (Table 6), and the optimisation schedule; Appendix B gives the modality-specific settings and their ablations. Section 4 and Appendix Table 12 describe the data splits and test prototypes. Every metric is defined in Appendix C, and the protocols for route swaps, translation and the pair-fitted ridge reference are in Appendix F. All results are means over three training seeds, with their variability in Appendix E, and are read from final checkpoints; no checkpoint is selected on test data. All baselines were rescored with the same evaluator from their released or retrained checkpoints (Appendix D). We will release the code, configuration files, evaluation scripts and trained checkpoints.

REFERENCES

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

Po-Hsuan Cameron Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A reduced-dimension fMRI shared response model. *Advances in Neural Information Processing Systems*, 28, 2015.

Matteo Ferrante, Tommaso Boccato, Grigorii Rashkov, and Nicola Toschi. Towards neural foundation models for vision: Aligning EEG, MEG, and fMRI representations for decoding, encoding, and modality conversion. *Information Fusion*, pp. 103650, 2025.

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.

Alessandro T. Gifford, Domenic Bersch, Daniel Janini, Gemma Roig, and Radoslaw Martin Cichy. The brain encoding response generator, 2025. URL `https://github.com/gifale95/BERG`. In preparation.

James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

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.

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.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver IO: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024.

Dongyang Li, Haoyang Qin, Mingyang Wu, Chen Wei, and Quanying Liu. BrainFLORA: Uncovering brain concept representation via multimodal neural embeddings. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5577–5586, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Zitong Lu, Yile Wang, and Julie D Golomb. Achieving more human brain-like vision via human EEG representational alignment. *Communications Biology*, 9(1):463, 2026.

Weijian Mai, Jiamin Wu, Yu Zhu, Zhouheng Yao, Dongzhan Zhou, Andrew Luo, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. SynBrain: Enhancing visual-to-fMRI synthesis via probabilistic representation learning. *Advances in Neural Information Processing Systems*, 38:29253–29284, 2026.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *International Conference on Learning Representations*, volume 2024, pp. 47648–47665, 2024.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. CognitionCapturer: Decoding visual stimuli from human EEG signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14486–14493, 2025.

Kaifan Zhang, Lihuo He, Junjie Ke, Yuqi Ji, Lukun Wu, Lizi Wang, and Xinbo Gao. CognitionCapturerPro: Towards high-fidelity visual decoding from EEG/MEG via multi-modal information and asymmetric alignment. *arXiv preprint arXiv:2603.12722*, 2026.

A Model and Training

Architecture. Table 5 lists every component. The shared latent has shape 256×256. Retrieval reads Z without any participant-specific parameter, so participant identity enters only on the response-prediction side. The common visual target is a fovea-blurred OpenCLIP ViT-bigG-14 feature map pooled to 4×4 spatial tokens ($16 \times 1{,}664 = 26{,}624$ dimensions). About 74% of the 268.11M parameters belong to participant-specific response heads, which match the unit-specific degrees of freedom of encoding baselines such as BERG.

Table 5: Complete architecture of the joint model. All modality encoders terminate in the same 256-token latent array; their input contracts reflect the geometry of each instrument.

component	implementation
EEG encoder	Dataset-z-scored 63×250 response; length-10 temporal patches with stride 5 give $63 \times 49 = 3{,}087$ value tokens. Each token adds 32-band Fourier embeddings of sensor position and time.
MEG encoder	Dataset-z-scored 271×201 response; a depthwise temporal stem with kernels 7/15/31 precedes length-10 patches with stride 8, giving $271 \times 24 = 6{,}504$ position-time tokens.
fMRI encoder	Participant-z-scored V_s-voxel response; a shared DeepSets patch stem produces 256 spatial tokens, and a participant-specific $\mathrm{Linear}(V_s, 256)$ ridge supplies one dense token. A learned null-time embedding replaces temporal position.
shared neural trunk	256 learned queries; eight Perceiver cross/self-attention blocks, width 256, eight heads, feed-forward multiplier 4, dropout .1, and an input skip. Modality identity enters the decoder query rather than the input tokens.
image encoder g_ϕ	Frozen CLIP feature $\rightarrow$ eight projected image tokens; 256 learned queries attend through two layers and return $Z_{\mathrm{stim}} \in \mathbb{R}^{256\times256}$. Modality and participant embeddings specify the response distribution to predict.
visual readout	$Z \rightarrow 16 \times 1{,}664$ OpenCLIP ViT-bigG-14 tokens. EEG/MEG use mean pooling; fMRI uses a learned mixer token. Sensors also use three 1,024-d rotation-bank heads and a detached-target fusion head.
neural decoder	A shared 1.18M coordinate decoder queries (Z, xyz, t, m, s); participant-specific residual heads add rank-250 sensor maps or $256 \rightarrow V_s$ fMRI maps.

Objective. Sensors use symmetric multi-positive InfoNCE at temperature 0.02 plus a hard-negative term (weight 3.0, margin 0.2). fMRI uses MixCo and SoftCLIP ($\beta = 0.15$, probability 0.5), with temperature 0.006 for the first third of training followed by a $0.004 \rightarrow 0.0075$ schedule. The sensor tie is InfoNCE at temperature 0.07; the fMRI tie is $1 - \cos$. Each EEG training input is the mean of two of the stimulus's four repetitions, redrawn every epoch. Table 6 gives the weights.

Table 6: Exact objective weights and batch sizes.

	λ_{align}	$\lambda_{\mathrm{encode}}$	λ_{tie}	batch	tie
EEG	1.0	1.0	0.1	128	InfoNCE
MEG	1.0	1.0	0.1	64	InfoNCE
fMRI	1.0	1.0	0.2	24	cosine

Schedule and optimization. The model is trained for 60 epochs (172,800 AdamW steps (Loshchilov & Hutter, 2017); about 20 hours on one NVIDIA A800 80GB GPU). Batches never mix modalities, so contrastive negatives stay within an instrument. Each epoch interleaves EEG, MEG and fMRI steps at a 3:2:3 ratio; compared with 1:1:1, this removes a late-training decline of MEG retrieval. Shared parameters are updated by every modality, so their weight decay is divided by the number of shared updates per modality step (2.667), keeping the per-epoch shrinkage of single-modality training. The response-prediction path receives stronger decay because MEG and fMRI targets are single noisy trials. Retrieval and encoding are read from bias-corrected weight EMAs with different horizons: retrieval from 0.999 on EEG and MEG and from 0.999995 on fMRI, encoding from 0.999995–0.999998. Separate horizons improve EEG encoding by $.0077 \pm .0033$ over three seeds at no retrieval cost.

B What Is Shared and What Is Not

The representation core is shared: one trunk, one latent array, one image-to-latent map, one coordinate decoder, and one visual target. A shared fovea-blurred target costs fMRI nothing relative to its own unblurred target (top-1 $.854$ vs $.827$). The measurement boundary is private, for two reasons. First, the instruments differ in geometry: fMRI voxel grids overlap by only 3.9% across participants, and a

dense per-participant voxel input raises fMRI top-1 from .051 to .274, whereas finer shared spatial patches stay between .034 and .054. Participant-specific residual response heads raise encoding by 69% on fMRI, .028 on EEG and .020 on MEG at no retrieval cost. Second, forcing one global setting charges one modality for another's convenience (Table 7): a common readout costs EEG or fMRI retrieval, the sensor alignment objective or tie damages fMRI, auxiliary heads that help sensors hurt fMRI, and MEG needs synthetic corruption because, unlike EEG, it has no repetitions to resample.

Table 7: Direct metrics for matched measurement-interface ablations. Within each adjacent pair, one field changes and the selected setting is bold. Retrieval is top-1 on the complete test gallery; encoding uses all-unit per-unit Pearson r.

modality	setting	top-1↑	encoding r ↑	Δ_Z ↑
EEG	**mean alignment readout**	**.6800**	**.0704**	–
EEG	mixer-token alignment readout	.6133	.0668	–
fMRI	**mixer-token alignment readout**	**.8267**	**.1807**	**.5909**
fMRI	mean alignment readout	.8000	.1807	.4978
fMRI	**MixCo + SoftCLIP alignment**	**.8267**	**.1807**	–
fMRI	sensor-style InfoNCE alignment	.6889	.1755	–
fMRI	**cosine latent tie**	**.8267**	**.1807**	**.5909**
fMRI	sensor-style InfoNCE tie	.8256	.1649	.4535
fMRI	**no auxiliary visual heads**	**.8544**	**.1879**	**.6036**
fMRI	three sensor auxiliary heads	.8178	.1846	.5141
MEG	**synthetic corruption, strength 1.0**	**.2500**	.0665	–
MEG	no synthetic corruption	.1850	**.0735**	–

C Metric Definitions

C.1 Retrieval

For test response i, let $C_i^{(G)}$ contain its true class y_i and $G-1$ distractors sampled uniformly without replacement, and let s_{ij} be the method's native similarity to candidate j. Top-k is

$$\mathrm{Top}\,k(G) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\mathrm{rank}_{j\in C_i^{(G)}}\, s_{ij} \le k\right]. \tag{10}$$

We average 30 seeded candidate draws whenever G is smaller than the gallery; the complete gallery is evaluated once. Information gain uses the same candidates:

$$\mathrm{bits}(G) = \log_2 G + \frac{1}{N}\sum_i \log_2 \frac{\exp(s_{iy_i}/\tau)}{\sum_{j\in C_i^{(G)}} \exp(s_{ij}/\tau)}. \tag{11}$$

The temperature τ is fitted once by minimizing negative log likelihood on the complete gallery and then held fixed for every candidate count. Chance therefore has zero expected bits. Similarity follows each upstream method's convention: cosine for every row except CogCapPro, which normalizes the brain prediction but retains the native gallery norms.

C.2 Encoding

For participant s, test stimulus i, and measured unit u, let X_{siu} and $\widehat{X}_{siu}$ denote the observed and predicted responses. The primary statistic is

$$r_{\mathrm{unit}} = \frac{1}{S}\sum_s \frac{1}{U_s}\sum_u \mathrm{corr}_i(\widehat{X}_{siu}, X_{siu}), \tag{12}$$

excluding a unit only when either vector has zero variance. This averages units within participant and then weights participants equally. The centered per-stimulus statistic first removes each unit's mean

Table 8: **Retrieval accuracy for every candidate count.** Columns report Top-1 accuracy (↑) for every gallery and Top-5 accuracy for galleries of size 10 or larger; smaller galleries are averaged over 30 resamples. Our row is the mean over three seeds (standard deviations in Appendix Table 13); baselines are single checkpoints. Bold and underline mark best and second best within a modality. fMRI has no 200-way result. Information gain is in Table 9.

	2-way	4-way	10-way		50-way		100-way		200-way	
method	Top-1	Top-1	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
EEG										
BiSURE (ours)	.9904	.9736	.9335	.9990	**.7954**	.9773	**.7030**	.9491	.5942	**.9000**
ATM	.9391	.8554	.7081	.9866	.4200	.7997	.3149	.6620	.2275	.5185
CogCapPro, fusion	**.9915**	**.9759**	**.9344**	**.9999**	.7884	**.9819**	.6928	**.9502**	**.5945**	.8905
CogCapPro, image	.9833	.9523	.8848	.9991	.6800	.9546	.5734	.8880	.4735	.7865
CogCapturer, image	.9371	.8530	.7082	.9790	.4327	.7921	.3302	.6619	.2430	.5305
CogCapturer, depth	.9114	.8000	.6330	.9651	.3431	.7104	.2458	.5681	.1695	.4345
CogCapturer, text	.8528	.6914	.4857	.9196	.2152	.5502	.1446	.3985	.0955	.2740
neural-CLIP	.6405	.4246	.2425	.6708	.0767	.2700	.0445	.1665	.0260	.0965
BrainFLORA, unimodal	.9482	.8756	.7427	.9884	.4733	.8307	.3613	.7114	.2675	.5865
BrainFLORA, multimodal	.9389	.8520	.7060	.9838	.4210	.7926	.3120	.6567	.2280	.5155
MEG										
BiSURE (ours)	**.9046**	**.7919**	**.6337**	**.9518**	**.3593**	**.7101**	**.2609**	**.5850**	**.1792**	**.4483**
CogCapPro, fusion	.8450	.6985	.5180	.9009	.2505	.5832	.1702	.4395	.1150	.3200
CogCapPro, image	.8258	.6593	.4702	.8842	.2114	.5257	.1431	.3868	.0950	.2725
neural-CLIP	.7933	.6155	.4097	.8492	.1568	.4618	.0958	.3102	.0537	.2075
BrainFLORA, unimodal	.8188	.6492	.4425	.8802	.1905	.5052	.1253	.3540	.0813	.2462
BrainFLORA, multimodal	.8375	.6725	.4751	.8972	.2112	.5333	.1420	.3860	.0912	.2675
fMRI										
BiSURE (ours)	**.9951**	**.9868**	**.9649**	**1.000**	**.8791**	**.9889**	**.8222**	**.9711**	–	
MindEye2	.9792	.9483	.8930	.9951	.7521	.9383	.6833	.8833	–	
neural-CLIP	.7913	.5968	.3822	.8519	.1613	.4232	.1167	.2700	–	
BrainFLORA, unimodal	.9134	.7937	.6111	.9733	.3418	.6857	.2567	.5367	–	
BrainFLORA, multimodal	.9196	.8121	.6283	.9740	.3093	.7206	.2000	.5300	–	

over test stimuli, $\widetilde{X}_{siu} = X_{siu} - N_s^{-1} \sum_j X_{sju}$ (and likewise for $\widehat{X}$), then averages $\mathrm{corr}_u(\widetilde{\widehat{X}}_{si\cdot}, \widetilde{X}_{si\cdot})$ over stimuli and participants. "Pearson" is the same per-stimulus correlation without the across-stimulus unit centering; "cosine" is the raw across-unit cosine. These two compatibility metrics retain the grand mean response pattern and are not primary.

For matrices $A, B \in \mathbb{R}^{N \times U}$ centered over stimuli, linear CKA is

$$\mathrm{CKA}(A, B) = \frac{\|A^\top B\|_F^2}{\|A^\top A\|_F \, \|B^\top B\|_F}, \tag{13}$$

and MSE is $N^{-1}U^{-1}\|A - B\|_F^2$ before centering. MSE is scale-dependent because different baselines predict under different output normalizations; it is reported for completeness, not used as the primary cross-method ranking.

C.3 Reconstruction

The six network-feature columns in Table 4 are two-way identification accuracies computed with the named frozen feature extractor; PixCorr and SSIM compare the generated and target images directly.

Table 9: **Information gain for the retrieval comparisons in Table 8.** Values are in bits (↑); our row is the mean over three seeds. Bold and underline denote best and second best within a modality and candidate count.

method	2-way	4-way	10-way	50-way	100-way	200-way
EEG						
BiSURE (ours)	<u>.9614</u>	<u>1.8920</u>	<u>3.0433</u>	<u>4.7079</u>	<u>5.2254</u>	<u>5.6095</u>
ATM	.7871	1.4620	2.1469	2.8145	2.9409	3.0099
CogCapPro, fusion	**.9677**	**1.9076**	**3.0648**	**4.7343**	**5.2519**	**5.6601**
CogCapPro, image	.9363	1.8176	2.8613	4.2049	4.5720	4.8457
CogCapturer, image	.7803	1.4521	2.1381	2.8277	2.9703	3.0553
CogCapturer, depth	.7021	1.2792	1.8387	2.3394	2.4305	2.4879
CogCapturer, text	.5296	.9149	1.2401	1.4774	1.5161	1.5394
neural-CLIP	.1024	.1747	.2292	.2642	.2683	.2703
BrainFLORA, unimodal	.8178	1.5395	2.2981	3.0900	3.2532	3.3537
BrainFLORA, multimodal	.7815	1.4510	2.1373	2.8082	2.9329	3.0121
MEG						
BiSURE (ours)	**.6815**	**1.2492**	**1.8246**	**2.3825**	**2.4933**	**2.5572**
CogCapPro, fusion	<u>.5198</u>	<u>.9357</u>	<u>1.3447</u>	<u>1.6797</u>	<u>1.7376</u>	<u>1.7777</u>
CogCapPro, image	.4723	.8353	1.1723	1.4303	1.4712	1.4968
neural-CLIP	.3865	.6707	.9037	1.0544	1.0748	1.0842
BrainFLORA, unimodal	.4475	.7744	1.0563	1.2628	1.2889	1.3070
BrainFLORA, multimodal	.4870	.8481	1.1685	1.4112	1.4463	1.4699
fMRI						
BiSURE (ours)	**.9802**	**1.9480**	**3.1802**	**5.1177**	**5.8355**	–
MindEye2	<u>.9228</u>	<u>1.7989</u>	<u>2.8789</u>	<u>4.4541</u>	<u>5.0091</u>	–
neural-CLIP	.3730	.6232	.8169	.9485	.9680	–
BrainFLORA, unimodal	.7141	1.2700	1.7859	2.2879	2.3739	–
BrainFLORA, multimodal	.7143	1.2897	1.8116	2.2350	2.3004	–

Table 10: **Encoding of measured neural responses, every metric.** Values are participant-equal means over all units; our row is the mean over three seeds (standard deviations in Appendix Table 13). Per-unit r, centered per-stimulus r, raw Pearson, raw cosine, and linear CKA are higher-is-better; MSE is lower-is-better. Bold and underline denote best and second best within a modality. Raw Pearson/cosine and MSE are included for compatibility, but the former retain the stimulus-independent mean pattern and MSE is output-scale dependent; per-unit r is primary. Definitions are in Appendix C.

method	per-unit r ↑	centered r ↑	Pearson↑	cosine↑	CKA↑	MSE↓
EEG						
BiSURE (ours)	**.1012**	**.2939**	.5626	.5624	**.4312**	**.01572**
BERG	<u>.0774</u>	<u>.1366</u>	.6084	.6104	<u>.4138</u>	.02453
CLIP ridge	.0723	.1299	**.6667**	**.6687**	.2628	<u>.02086</u>
ReAlnet	.0063	.0121	<u>.6526</u>	<u>.6549</u>	.0412	.02156
MEG						
BiSURE (ours)	**.1254**	**.1539**	**.3969**	**.3934**	.5181	**.08997**
BERG	<u>.0568</u>	.0609	.1983	.1972	**.7143**	5.99206
CLIP ridge	.0566	<u>.0625</u>	<u>.2075</u>	<u>.2065</u>	<u>.5394</u>	<u>5.95422</u>
fMRI						
BiSURE (ours)	**.2029**	**.2753**	.2600	.2528	.5601	.09630
BERG	<u>.1673</u>	<u>.2348</u>	.1610	.1486	<u>.6336</u>	.03404
CLIP ridge	.1581	.2199	**.7167**	**.7455**	.5660	**.00045**
SynBrain	.1011	.1496	<u>.5822</u>	<u>.6003</u>	**.6797**	<u>.00082</u>

D Evaluation Validity

D.1 Within-participant response prediction

Every unit is included: 15,750 channel-time units for each EEG participant, 54,471 for each MEG participant, and 5,238–6,036 voxels for each fMRI participant. Correlating over the pooled participant and image axes would inflate the EEG statistic by $5.7\times$, because it rewards reproducing each participant's mean response; we never use it. Table 11 gives every participant.

Table 11: Within-participant per-unit encoding correlation. Ours is the mean over three seeds; at every seed, every participant is above its BERG value (51/51 participant-seed cells).

EEG	01	02	03	04	05	06	07	08	09	10
BiSURE (ours)	.0876	.0909	.1061	.1161	.0900	.1033	.0949	.1275	.0876	.1081
BERG	.0657	.0738	.0788	.0888	.0693	.0753	.0673	.0974	.0682	.0892

MEG	P1	P2	P3	P4
BiSURE (ours)	.1095	.1879	.1228	.0815
BERG	.0407	.0887	.0679	.0297

fMRI	01	02	03
BiSURE (ours)	.2133	.2027	.1927
BERG	.1902	.1812	.1305

D.2 Checkpoints and test prototypes

All headline values come from final/last checkpoints. Ours are the final checkpoints of three seeds after 60 training epochs; the baselines use their released last checkpoint except for the explicit ATM pooled epoch-35 exception, for which the final epoch-40 file is unavailable. All embeddings are rescored by the common protocol in Appendix C; no row selects an epoch using these results.

Table 12: Datasets and evaluation units. Sensors use 200-way galleries; fMRI uses 100-way.

	subjects	units/subject	pooled train rows	test rows	test prototype
EEG	10	63×250	165,400	2,000	80 repeats, one image
MEG	4	271×201	79,392	800	12 images, one concept
fMRI	3	5,238–6,036	25,920	300	12 repeats, one image

The MEG prototype averages 12 different images of one concept, not repeated measurements of one image. Scoring a single image per class moves the same model from .7647 to .6454 in a diagnostic gallery. This estimand is shared by all methods on that split and is why numerical comparisons are made only down a modality column.

D.3 Reconstruction controls

Rendering the training-set mean for every test stimulus gives PixCorr/SSIM of $.1593/.4030$ on EEG, $.1414/.3844$ on MEG, and $.1300/.4000$ on fMRI. This null is above every evaluated method on SSIM, so the low-level columns in Table 4 rank methods but supply no positive evidence of low-level fidelity. Our PixCorr exceeds the null on EEG ($+.068$) and fMRI ($+.038$). The semantic columns carry the comparison. CogCapPro reconstruction also uses three IP-Adapter conditions (image, depth, and edge), whereas ours uses one. Its released alignment stage selects an epoch using the test loader. We therefore treat specialist reconstruction values as context, not a controlled renderer comparison.

E SEED VARIABILITY

Table 13: Mean ± standard deviation over three training seeds of the reported model. Retrieval is on the complete gallery (200-way for sensors, 100-way for fMRI). The swapped retrieval head varies most across seeds, so its range (minimum to maximum) is given instead of a standard deviation. Without the norm scalar c_m (Equation 8), fMRI's swapped retrieval head gives .184 (.037 to .467); EEG and MEG are unchanged.

	Top-1	bits	per-unit r	Δ_Z	swapped R top-1	swapped D r
EEG	.5942 ± .0086	5.61 ± .11	.1012 ± .0001	.3444 ± .0041	.524 (.319–.834)	.0929 ± .0012
MEG	.1792 ± .0195	2.56 ± .14	.1254 ± .0038	.2118 ± .0137	.105 (.056–.196)	.0901 ± .0045
fMRI	.8222 ± .0269	5.84 ± .05	.2029 ± .0032	.5156 ± .0286	.998 (.993–1.000)	.1433 ± .0019

F ROUTE SWAPS AND TRANSLATION

Protocol. Every test in Section 5.1 and Section 5.2 uses the final checkpoints and fits nothing on test data. The only fitted quantity is the norm scalar c_m of the swapped retrieval head, computed on up to 4,096 training rows per participant. For translation (Equation 9), the source participant's test response is encoded by the shared encoder and read out through the target participant's response head for every unit of the target. We score every ordered participant pair: 90 EEG→EEG, 12 MEG→MEG, 6 fMRI→fMRI, 40 EEG→MEG and 40 MEG→EEG pairs per seed. Each score is the per-unit correlation over test stimuli, averaged over the target's units, divided by the target's image-driven prediction from the same checkpoint. The CLIP route replaces Z with the source response's decoded CLIP features, which are passed through the image map g_ϕ to the target head. A mismatched-stimulus control pairs each source response with a different stimulus's target. The pair-fitted reference is a kernel ridge regression from the source participant's full training response to the target participant's, on the training images both saw (EEG: 16,540 four-trial averages; MEG: 19,848 and fMRI: 8,640 single trials; EEG and MEG: 16,520 shared images). A per-unit penalty is chosen from a 17-point grid on a concept-grouped tenth of those images, the map is refitted on all of them, and it is scored on the same test rows, units and denominator as the route through Z. Extending the grid two decades lower moves the EEG→EEG result from 62% to 61%. Without the tie, the route through Z falls below this map in 563 of 564 cells.

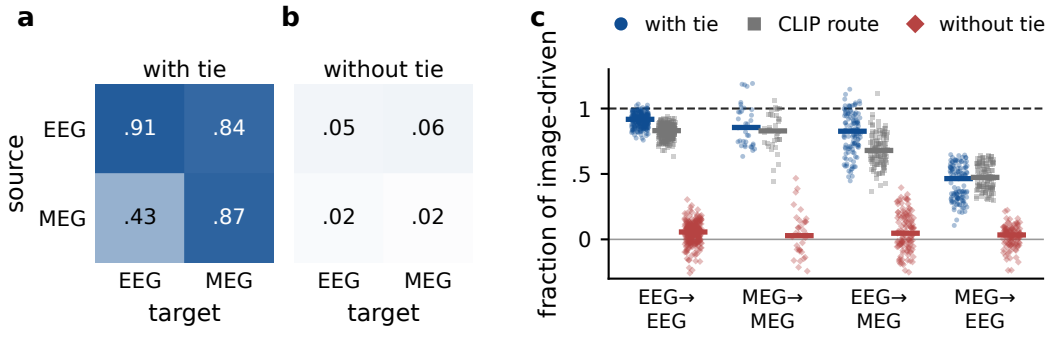

Figure 6: **Translation between EEG and MEG participants, all pairs. (a, b)** Mean fraction of the target's image-driven prediction over ordered participant pairs and three seeds, with and without the latent tie. **(c)** Every ordered pair at every seed (EEG→EEG 270, MEG→MEG 36, EEG→MEG 120, MEG→EEG 120); bars are medians. Blue: through Z with the tie. Grey: the CLIP route (the source response decoded to CLIP, then mapped to the target). Red: through Z without the tie. With the tie, Z carries more than the CLIP route in EEG→EEG, MEG→MEG and EEG→MEG. Without the tie it carries almost nothing.

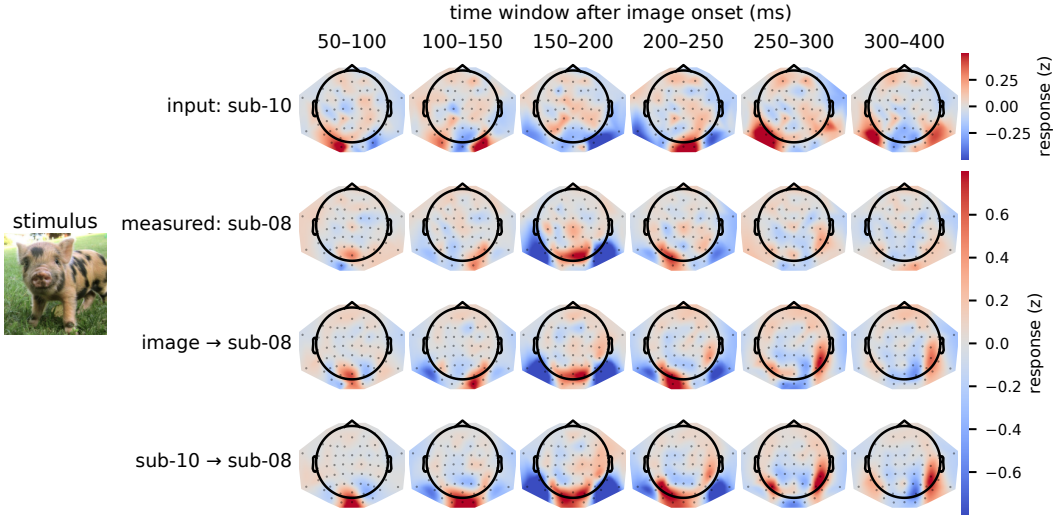

Figure 7: **One EEG translation in detail.** Stimulus: piglet. Rows: participant sub-10's measured response (the only input; own colour scale), participant sub-08's measured response, sub-08 predicted from the image, and sub-08 predicted from sub-10's response through Z. Columns are time windows after image onset; values are response amplitudes (train z-score). This is the best stimulus of the best pair (pattern r: translation .59, image .57). Prediction rows use one display gain each over the whole test set.

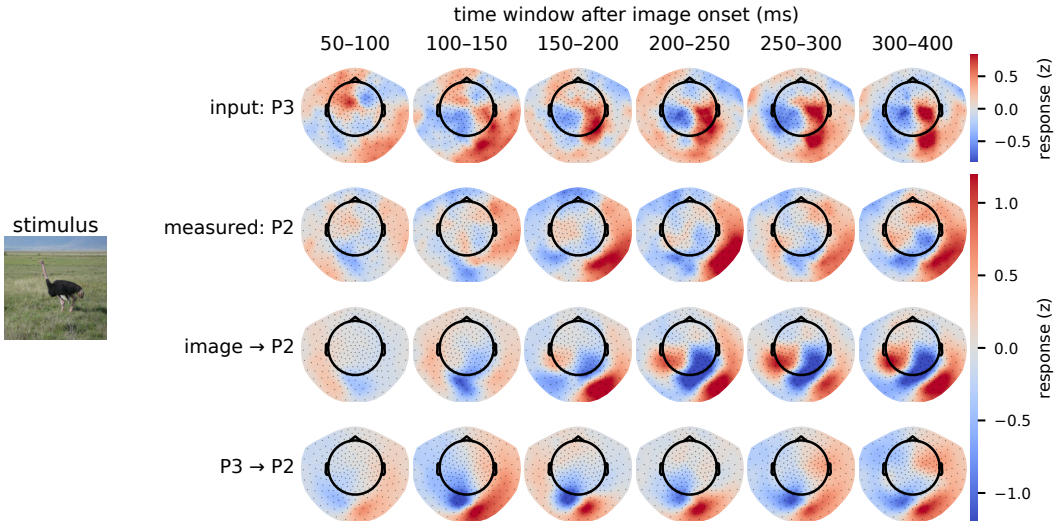

Figure 8: **One MEG translation in detail.** As Figure 7, for MEG participant P3 → P2 (the best pair of MEG→MEG). Stimulus: ostrich, one of the pair's three best stimuli (pattern r: translation .61, image .31).

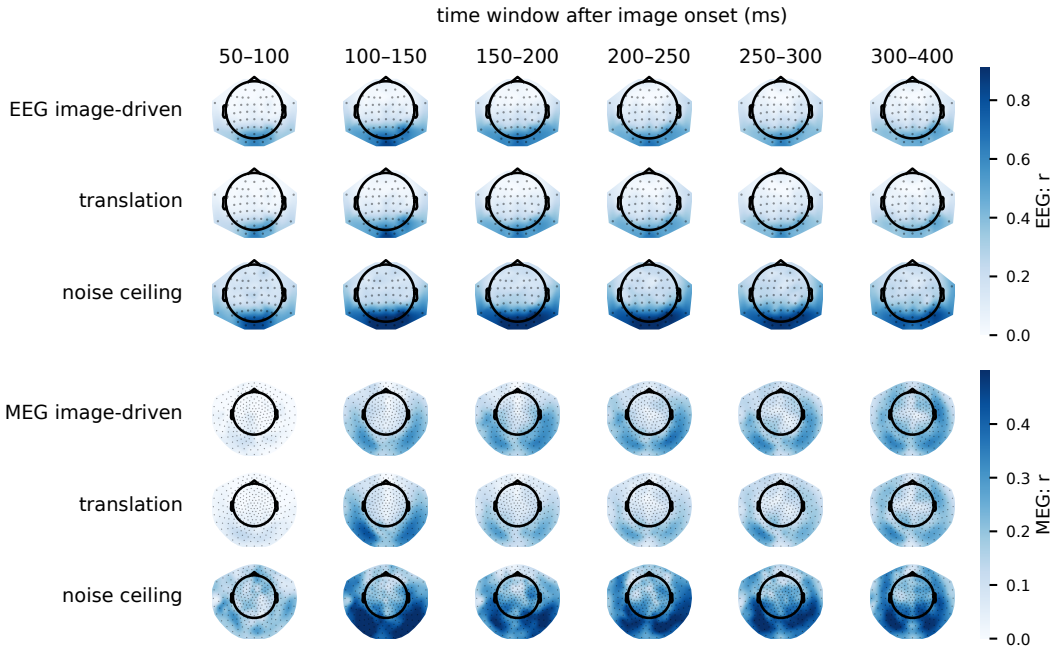

Figure 9: **Where and when the latent predicts the response, against what the data allow.** Per-sensor correlation over all test stimuli, averaged within each time window. Rows: prediction from the image (averaged over participants), stimulus-blind translation from another participant of the same modality (averaged over all ordered pairs: EEG 90, MEG 12), and the split-half noise ceiling (averaged over participants). The three rows of a modality share one colour scale. Averaged over sensors and 50–400 ms, translation reaches 90% (EEG) and 88% (MEG) of the image-driven prediction. Both follow the spatial and temporal pattern of the ceiling: occipital for EEG, posterior for MEG.

G JOINT VERSUS SINGLE-MODALITY TRAINING

Each control is the same architecture and objective trained on one modality for the same number of epochs, three seeds each. Retrieval is the mean of the last three evaluations; encoding and Δ_Z use the final checkpoint.

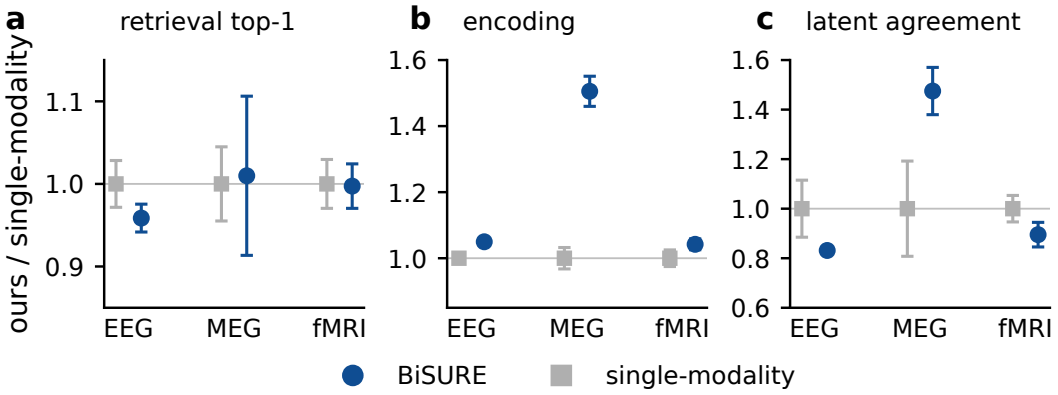

Figure 10: **What joint training changes.** Each modality's joint model (blue), relative to the same architecture trained on that modality alone (grey), mean ± sd over three seeds each. Joint training leaves retrieval unchanged (a; EEG top-1 is .026 lower, but its information gain is unchanged, Section 5.4), raises encoding in all three modalities (b; MEG $+51\%$), and changes latent agreement (c) by modality: MEG, the modality with the least data, gains $1.5\times$, while EEG and fMRI give up 17% and 10%.

H RECONSTRUCTIONS COMPARED WITH EVERY BASELINE

Figures 11, 12 and 13 compare our reconstructions with every baseline that released reconstructions, on eight test stimuli per modality drawn at random. Rows: the stimulus, ours (the same reconstructions scored in Table 4), and each baseline.

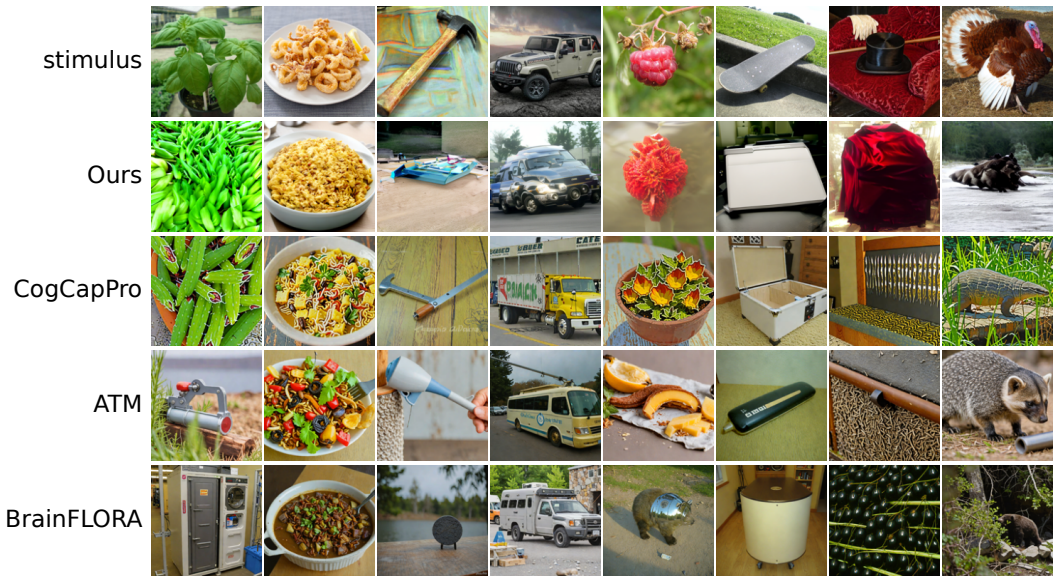

Figure 11: **EEG reconstructions of eight random test stimuli** (participant sub-08).



Figure 12: **MEG reconstructions of eight random test stimuli** (participant P2).

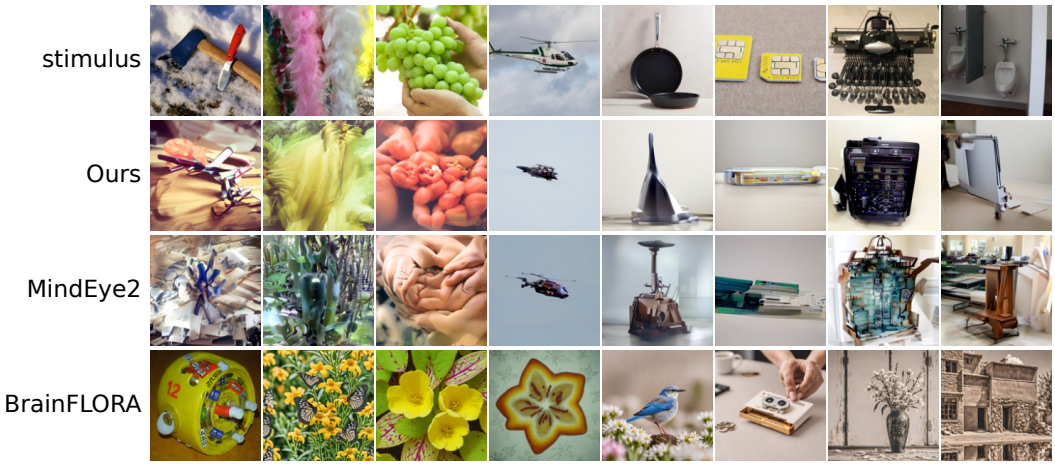

Figure 13: **fMRI reconstructions of eight random test stimuli** (participant sub-01).