BEYOND HIERARCHICAL ALIGNMENT: DISTILLING TEACHER DIVERSITY FOR BRAIN VISUAL DECODING

Anonymous authors
Paper under double-blind review

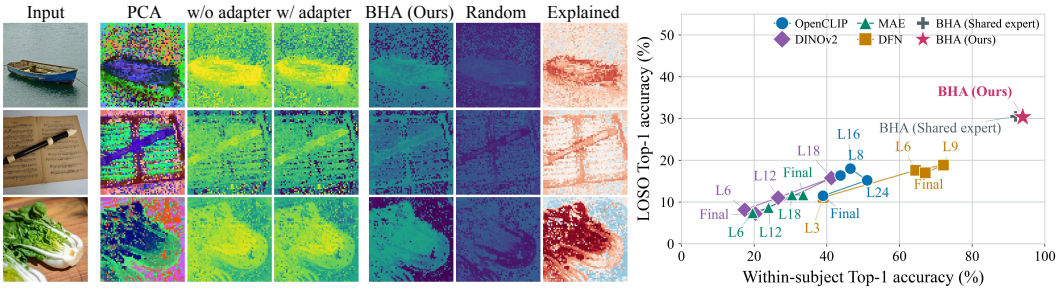


Figure 1: Visual representations shape brain–image alignment. Left: Input; visual-feature PCA; global-to-patch similarity before/after adaptation; BHA-to-CLIP-patch similarity; random-query reference (Random); BHA-minus-random difference (Explained). **Right:** CSLS-2 Top-1 on THINGS-EEG2 for VFM layers and BHA readouts. The BHA LOSO point uses Stage 1; training settings differ across configurations.

ABSTRACT

A strong visual representation is not necessarily an effective target for brain decoding. Different visual models capture complementary stimulus properties, but their usefulness depends on what can be predicted from limited, noisy brain recordings. Fusing these representations provides shared supervision, yet still defines a single similarity geometry with its own representational bias. The fused space also need not retain each teacher's visual structure or match the conditioning space of a pretrained generator. We introduce **BHA (Beyond Hierarchical Alignment)** to learn shared brain–visual alignment while composing predictions across distinct visual spaces. BHA jointly learns a brain encoder and a fused visual target, allowing visual supervision to adapt to brain signals. To avoid relying exclusively on the resulting shared geometry, teacher-specific Group Mixture-of-Experts (MoE) readouts learn separate visual predictions on the frozen backbone. Learnable gating and query-dependent weighting compose their matching scores with the shared score, allowing complementary notions of visual similarity to contribute to retrieval. Separately trained semantic and spatial adapters reuse the same backbone for reconstruction without fine-tuning the pretrained generator. BHA improves within-subject retrieval on EEG and MEG, with gains persisting under reduced EEG training data. To examine what distinct visual spaces contribute, we analyze the learned representations and their matching behavior. Experts can assign opposing contributions to the same image regions, while learned model neurons exhibit semantic selectivity. These findings suggest that shared alignment can support distinct visual matching functions that can be composed for brain decoding.

1 INTRODUCTION

Decoding visual stimuli from non-invasive brain recordings requires learning from limited, noisy neural signals (Simanova et al., 2010; Kaneshiro et al., 2015; Horikawa & Kamitani, 2017; Kavasidis et al., 2017; Schirrmeister et al., 2017; Spampinato et al., 2017). Pretrained vision foundation models (VFMs) provide visual targets that let brain encoders predict image representations rather than pixels. This is particularly useful for EEG and MEG, given their low signal-to-noise ratios and limited spatial resolution (Benchetrit et al., 2024; Song et al., 2024; Bai et al., 2024; Li et al., 2024; Fu et al., 2025). Different models and network depths emphasize different stimulus properties, motivating

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

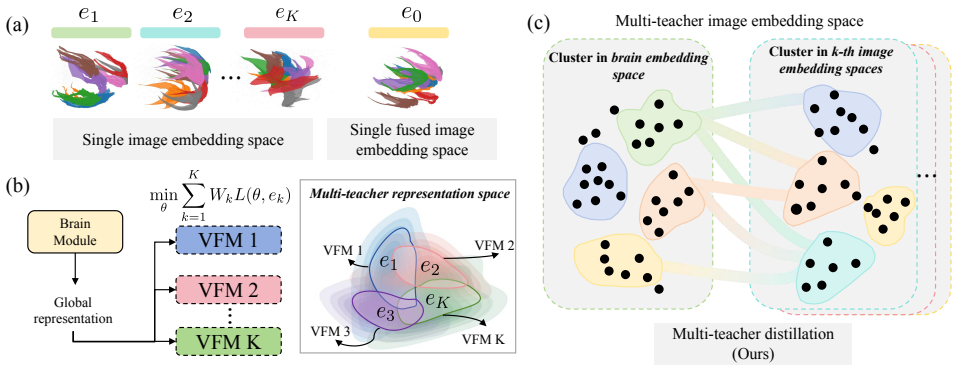


Figure 2: **Shared alignment need not erase teacher-specific structure.** (a) Different VFMs define distinct visual geometries; fusion combines them into a single target space. (b) Multi-teacher supervision allows one brain representation to support predictions in several visual spaces. (c) Teacher-specific readouts retain distinct brain–image correspondences, motivating score composition alongside shared alignment. The manifolds and clusters illustrate the design principle schematically.

complementary visual supervision (Li et al., 2025a; Wu et al., 2025; Du et al., 2026). Figure 1 illustrates how teacher choice and adaptation affect retrieval and spatial matching.

Recent work shows that the choice of visual supervision matters: intermediate features benefit downstream multimodal and spatial tasks (Bolya et al., 2025) and brain-to-image retrieval (Du et al., 2026), while fusing global CLIP embeddings with VAE latents provides a shared alignment target spanning semantics and structure (Zheng et al., 2026). These findings motivate a question beyond feature selection or fusion: *which differences between visual teachers are predictable from brain activity and support complementary forms of image matching?* We hypothesize that teacher diversity benefits decoding when its distinct information is both learnable from neural signals and useful beyond a shared alignment target.

Retrieval and reconstruction place different demands on visual supervision. Adapting features toward brain-predictable information can improve matching while changing the structure expected by a frozen generator. This motivates sharing a brain representation while predicting features in the spaces required by each task.

Our approach. We introduce **BHA (Beyond Hierarchical Alignment)** to combine shared brain–visual alignment with task-specific prediction spaces (Figure 2). BHA jointly learns a brain encoder and a fused visual target. Teacher-specific Group MoE readouts then map the frozen brain representation into distinct visual spaces; input-dependent gating combines their matching scores with the shared score for within-subject retrieval. Separately trained semantic and spatial adapters predict conditions in the generator's native feature spaces, supporting reconstruction without updating the generator. Across teacher families and network depths, we examine which visual distinctions remain predictable from brain activity and how they appear in matching behavior and semantic selectivity.

Our main contributions are:

- We formalize how complementary teacher supervision can reduce alignment estimation error, while adapting visual targets can change their compatibility with a frozen generator.
- We introduce BHA, which jointly learns a brain encoder and a visual fusion head, and unifies retrieval and reconstruction through a shared backbone with task-specific readouts. Group MoE readouts and learnable gating substantially improve within-subject visual retrieval with limited training data.
- Our empirical analyses reveal visual and semantic selectivity in learned brain representations, with distinct concept associations and spatial contributions to image matching across experts.

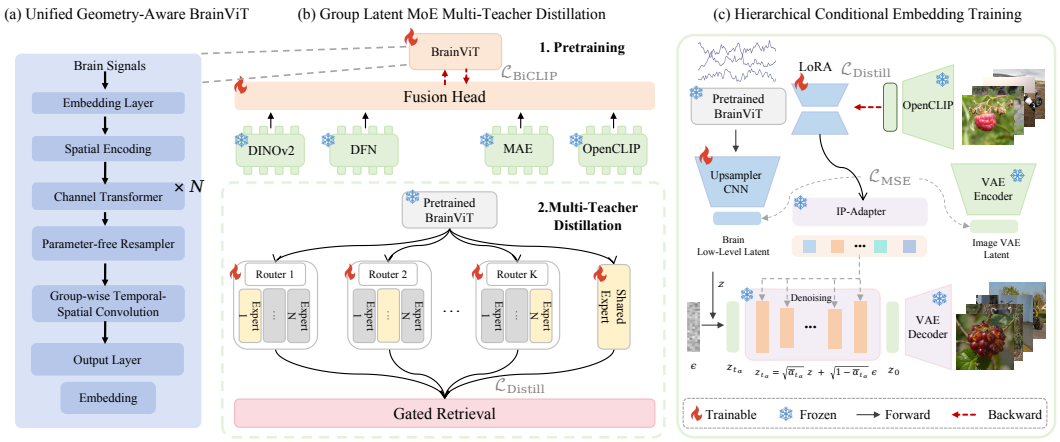


Figure 3: **BHA learns shared representations and task-specific predictions.** (a) BrainViT combines channel attention and feature–sensor convolutions. (b) Stage 1 jointly learns BrainViT and visual fusion; Stage 2 freezes both and trains shared and teacher-specific readouts, visual adapters, and routing. (c) Semantic and spatial predictors condition the frozen diffusion generator, VAE, and IP-Adapter. The schematic uses a scaled latent $z = a\hat{\mathbf{z}}$ and normalized noisy coordinates (Appendix C.2.2).

2 Related Work

Representation Mixture for Visual Understanding VFM complementarity benefits visual understanding across layers and models. MMFuser (Cao et al., 2026) queries shallow and intermediate details with deep features. AM-RADIO (Ranzinger et al., 2024) and RADIOv2.5 (Heinrich et al., 2025) distill VFMs; Swiss Army Knife (Lu et al., 2025) mixes teacher-specific features; and Theia (Shang et al., 2025) and SigLino (Chaybouti et al., 2025) compress heterogeneous teachers into compact students. Multi-encoder VLMs also combine visual features: BRAVE (Kar et al., 2024) compresses features from five encoders, Cambrian-1 (Tong et al., 2024) aggregates spatial features, and Eagle (Shi et al., 2025) finds simple concatenation competitive with complex mixers. MoME (Shen et al., 2024), MoVA (Zong et al., 2024), and PhotoEye (Qi et al., 2025) condition expert fusion on tasks or instructions. Together, these studies report downstream gains from VFM mixtures. BHA transfers this principle to neural decoding, mixing models and depths for retrieval while retaining generator-native targets.

Brain Visual Decoding and Reconstruction Brain-to-image reconstruction combines neural feature prediction with generative priors (Lin et al., 2022; Chen et al., 2023; Takagi & Nishimoto, 2023; Zeng et al., 2024; Bai et al., 2024; Scotti et al., 2024), including methods that keep generators frozen (Scotti et al., 2023; Li et al., 2024; Mayo et al., 2024). Projection modules connect predicted embeddings to generation (Li et al., 2024; Dai et al., 2026; Zheng et al., 2026), and fused features can require a separately trained conditioning prior (Zheng et al., 2026). ENIGMA (Kneeland et al., 2026) also predicts CLIP features for SDXL-Turbo's IP-Adapter. BHA uses a multi-teacher-trained brain backbone with separate semantic and spatial readouts, and optionally refines noisy latents toward the predicted VAE condition during inference.

Hierarchical Brain–Vision Modeling Intermediate visual features can outperform final-layer representations (Bolya et al., 2025). Neural decoding studies examine feature granularity (Du et al., 2026), visibility-guided layer fusion (Tang et al., 2026), and modality-specific routing (Li et al., 2025a). In fMRI, VINDEX (Xia & Oztireli, 2025) compares feature organizations, while MindHier (Zhang et al., 2026b) aligns brain and visual hierarchies for reconstruction. MBP (Zheng et al., 2026) trains a Fusion Prior, including IP-Adapter, while freezing the diffusion backbone. BHA uses fusion for retrieval and native generator conditions for reconstruction, avoiding that additional adaptation stage. Teacher-specific Group MoE outputs preserve separate alignment spaces, while the shared expert predicts the fused visual target.

3 Problem Setup and Preliminaries

Given B brain–image pairs $\{(\mathbf{X}_i, \mathbf{I}_i)\}_{i=1}^{B}$, a brain encoder and readout predict $\hat{\mathbf{e}}_i = H(f_\theta(\mathbf{X}_i, s_i)) \in \mathbb{R}^d$, where s_i optionally identifies the subject. A frozen visual encoder provides targets $\mathbf{e}_i = \phi(\mathbf{I}_i) \in \mathbb{R}^d$. For nonzero feature vectors, let $\mathrm{norm}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$, $\mathrm{sim}(\mathbf{x}, \mathbf{y}) = \mathrm{norm}(\mathbf{x})^\top \mathrm{norm}(\mathbf{y})$, and $\ell_{ij} = \mathrm{sim}(\hat{\mathbf{e}}_i, \mathbf{e}_j)/\tau$ with temperature $\tau > 0$. Feature regression and bidirectional contrastive alignment (BiCLIP) (Radford et al., 2021) use

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{Bd} \sum_{i=1}^{B} \|\hat{\mathbf{e}}_i - \mathbf{e}_i\|_2^2, \tag{1}$$

$$\mathcal{L}_{\mathrm{BiCLIP}}(\boldsymbol{\ell}) = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{e^{\ell_{ii}}}{\sum_j e^{\ell_{ij}}} + \log \frac{e^{\ell_{ii}}}{\sum_j e^{\ell_{ji}}} \right]. \tag{2}$$

BiCLIP averages the brain-to-image and image-to-brain contrastive losses. Retrieval ranks images by embedding similarity. Reconstruction predicts semantic conditions and spatial latents in the pretrained generator's native feature spaces. We use these losses within the corresponding prediction spaces, with task-specific weights.

4 Joint Brain–Visual Adaptation and Multi-Teacher Distillation

4.1 Complementary Supervision, Distinct Prediction Spaces

Visual supervision must be informative and predictable from limited brain recordings. The following linear model shows how complementary teachers reduce estimation error without removing irreducible uncertainty.

Proposition 1 (Alignment risk under complementary teacher supervision)**.** Let $\mathbf{b} \in \mathbb{R}^p$ be brain features with $\mathbb{E}[\mathbf{b}\mathbf{b}^\top] = I_p$, with visual target $\mathbf{y} = \mathbf{M}_\star \mathbf{b} + \boldsymbol{\xi} \in \mathbb{R}^{d_0}$ and common-space teacher features $\mathbf{t}_k = \mathbf{y} + \boldsymbol{\delta}_k$, $k = 1, \ldots, K$. The residual $\boldsymbol{\xi}$ and joint teacher-error vector have zero means, finite second moments, and are independent of $\mathbf{b}$ and of each other. Suppose $\mathbb{E}\|\boldsymbol{\xi}\|_2^2/d_0 = \sigma_{\mathrm{br}}^2$, $\mathbb{E}\|\boldsymbol{\delta}_k\|_2^2/d_0 = \sigma_{\mathrm{T}}^2 > 0$, and $\mathbb{E}[\boldsymbol{\delta}_k^\top \boldsymbol{\delta}_\ell]/d_0 \le \rho_{\mathrm{T}} \sigma_{\mathrm{T}}^2$ for $k \neq \ell$, where $0 \le \rho_{\mathrm{T}} < 1$. Fit a linear predictor $\widehat{g}_K$ by least squares to $K^{-1} \sum_k \mathbf{t}_k$ on n i.i.d. samples with design matrix B satisfying $\mathsf{B}^\top \mathsf{B} \succeq n\kappa I_p$, $\kappa > 0$. Define test risk $\mathcal{R}_{\mathrm{vis}}(g) = \mathbb{E}\|g(\mathbf{b}) - \mathbf{y}\|_2^2/d_0$ and $\chi_n = \mathrm{tr}[(\mathsf{B}^\top \mathsf{B})^{-1}] \le p/(n\kappa)$. Then

$$\sigma_{\mathrm{br}}^2 \le \mathbb{E}_{\mathrm{train}}[\mathcal{R}_{\mathrm{vis}}(\widehat{g}_K) \mid \mathsf{B}] \le \sigma_{\mathrm{br}}^2 + \chi_n \left[\sigma_{\mathrm{br}}^2 + \sigma_{\mathrm{T}}^2 \left(\rho_{\mathrm{T}} + \frac{1 - \rho_{\mathrm{T}}}{K} \right) \right]. \tag{3}$$

For $K > 1$, this upper bound is strictly below the expected risk of any single-teacher least-squares predictor conditional on the same design. With uncorrelated teacher errors, the teacher contribution decreases by a factor of K.

Proposition 1 links complementary supervision to lower estimation risk and, under a target-norm condition, a tighter angular-error bound (Appendix B.1). This linear result motivates, but does not guarantee, gains from learned nonlinear fusion. Improving shared alignment does not ensure that the adapted space preserves the visual content or feature geometry needed for reconstruction.

Proposition 2 (Alignment to adapted targets and reconstruction discrepancy)**.** Let $\mathbf{e}$ be a native visual feature, $\mathbf{t} = V(\mathbf{e})$ an adapted target, and $\mathbf{h} = f(\mathbf{X})$ a brain prediction, with $\|\mathbf{h}\|_2 = \|\mathbf{t}\|_2 = 1$ almost surely. For fixed sampling noise, let G be a frozen L_G-Lipschitz decoder on the feature domain. Define angular error $\delta = \mathbb{E}[1 - \mathbf{h}^\top \mathbf{t}]$, adaptation-induced output drift $\Delta_G = \mathbb{E}\|G(\mathbf{t}) - G(\mathbf{e})\|_2^2$, and native-output discrepancy $\mathcal{R}_G = \mathbb{E}\|G(\mathbf{h}) - G(\mathbf{e})\|_2^2$. Assuming finite second moments,

$$\left| \sqrt{\mathcal{R}_G} - \sqrt{\Delta_G} \right| \le L_G \sqrt{2\delta}. \tag{4}$$

Proposition 2 distinguishes alignment to a learned target from compatibility with a pretrained generator: even perfect alignment leaves $\mathcal{R}_G = \Delta_G$ relative to the native-conditioned output

(Appendix B.2). This motivates prediction spaces tailored to each task. BHA uses teacher-specific Group MoEs for distinct matching functions and semantic and spatial adapters for generator-native conditions (Figure 3), all drawing on information retained in a shared brain backbone.

To preserve this common representation during specialization, we learn it before fitting the task-specific predictors. The brain encoder and visual fusion first adapt jointly using frozen VFM features $\mathbf{e}_{ik} = \phi_k(\mathbf{I}_i) \in \mathbb{R}^{d_k}$. We then freeze the backbone and fusion, allowing the prediction branches to specialize without their objectives reshaping the shared representation.

4.2 Two-Stage Training

Retrieval follows this separation in two stages: Stage 1 learns shared alignment; Stage 2 learns how shared and teacher-specific predictions contribute to matching. Reconstruction predictors are trained separately on the pretrained backbone (Section 4.3).

Stage 1: Shared representation learning. A learned fusion combines complementary visual features into a target that adapts with the brain encoder. Let $\mathbf{h}_i = f_\theta(\mathbf{X}_i, s_i) \in \mathbb{R}^{d_0}$. Projections $W_k \in \mathbb{R}^{d_0 \times d_k}$ map teacher features to a common space, and $F : \mathbb{R}^{d_0} \to \mathbb{R}^{d_0}$ forms the fused target:

$$\mathbf{e}_{i0} = \mathrm{norm}\left[F\left(\sum\nolimits_{k=1}^{K} \mathrm{norm}(W_k \mathbf{e}_{ik})\right)\right], \qquad \mathcal{L}_{\mathrm{pre}} = \mathcal{L}_{\mathrm{BiCLIP}}(\boldsymbol{\ell}_{\mathrm{pre}}). \tag{5}$$

Here $\ell_{\mathrm{pre},ij} = \mathrm{sim}(\mathbf{h}_i, \mathbf{e}_{j0})/\tau_{\mathrm{pre}}$ with temperature $\tau_{\mathrm{pre}} > 0$. Joint brain–visual adaptation couples encoder learning with target learning: the brain encoder predicts visual features, while fusion adapts the supervisory target to neural signals.

Stage 2: Teacher-specific prediction. Shared brain features can support predictions in multiple visual spaces. With BrainViT and fusion frozen, teacher-specific Group MoEs learn separate mappings into their teachers' feature spaces, allowing visual distinctions recoverable from the backbone to contribute to matching. A retrieval adapter yields $\mathbf{h}'_i = A(\mathrm{sg}(\mathbf{h}_i)) \in \mathbb{R}^{d_0}$, where sg denotes stop-gradient. Each group $k = 1, \ldots, K$ combines a base projection $P_k \in \mathbb{R}^{d_k \times d_0}$, an always-active expert U_k, and R input-dependent residual experts, each mapping $\mathbb{R}^{d_0}$ to $\mathbb{R}^{d_k}$:

$$\hat{\mathbf{e}}_{ik} = P_k \mathbf{h}'_i + U_k(\mathbf{h}'_i) + \sum\nolimits_{r=1}^{R} w_{ikr} E_{kr}(\mathbf{h}'_i) \in \mathbb{R}^{d_k}. \tag{6}$$

Here $w_{ikr} \geq 0$ and $\sum_r w_{ikr} = 1$; the shared readout E_0 predicts the fused target as $\hat{\mathbf{e}}_{i0} = E_0(\mathbf{h}'_i)$. Groups retain their teachers' feature dimensions d_k, accommodating heterogeneous visual spaces. Trainable visual adapters $\tilde{\mathbf{e}}_{ik} = V_k(\mathrm{sg}(\mathbf{e}_{ik})) \in \mathbb{R}^{d_k}$ adjust teacher-specific targets without updating the source VFMs; $\tilde{\mathbf{e}}_{i0} = \mathrm{sg}(\mathbf{e}_{i0})$ remains frozen. Each readout uses a positive temperature τ_k and is aligned through

$$\ell^k_{ij} = \mathrm{sim}(\hat{\mathbf{e}}_{ik}, \tilde{\mathbf{e}}_{jk})/\tau_k, \qquad k = 0, \ldots, K. \tag{7}$$

Readouts, visual adapters, and routing are jointly trained with

$$\begin{aligned} \mathcal{L}_{\mathrm{distill}} = \sum\nolimits_{k=0}^{K} \lambda_k \mathcal{L}_{\mathrm{BiCLIP}}(\boldsymbol{\ell}^k) + \lambda_{\mathrm{ret}} \mathcal{L}_{\mathrm{ret}} \\ + \lambda_{\mathrm{aux}} \mathcal{L}_{\mathrm{aux}} + \lambda_{\mathrm{drift}} \mathcal{L}_{\mathrm{drift}}. \end{aligned} \tag{8}$$

The objective combines per-space alignment and retrieval-score learning ($\mathcal{L}_{\mathrm{ret}}$), with regularization for expert usage ($\mathcal{L}_{\mathrm{aux}}$) and visual drift ($\mathcal{L}_{\mathrm{drift}}$). Appendix C.2.1 details these objectives.

Gated score composition. Although fusion combines multiple teachers, it defines a single similarity geometry that can attenuate distinctions preserved by individual teachers. We compose matching scores to retain these teacher-specific distinctions alongside shared alignment. For brain query i and candidate image j, define $s^k_{ij} = \mathrm{sim}(\hat{\mathbf{e}}_{ik}, \tilde{\mathbf{e}}_{jk})$. The composed score is

$$S_{ij} = s^0_{ij} + g_i \sum\nolimits_{k=1}^{K} \pi_{ik} s^k_{ij}. \tag{9}$$

The shared score s^0_{ij} supplies the common matching component. Query-dependent weights $\pi_{ik} \geq 0$, $\sum_k \pi_{ik} = 1$, form a convex combination of teacher-specific scores, while the gate $g_i \in [0, 1]$ sets its overall contribution. Learning both separates the relative weighting of teachers from the strength of their combined contribution to shared matching.

4.3 Reconstruction without Generator Fine-Tuning

Retrieval alignment alone does not ensure reconstruction fidelity (Proposition 2). We therefore train task-specific semantic and spatial predictors in the generator's native feature spaces, using the pretrained BrainViT backbone with frozen base weights and bypassing retrieval readouts. The semantic prediction $\hat{\mathbf{e}}^{\mathrm{sem}}$ provides the condition $\mathbf{c}_{\mathrm{sem}}$ for denoising, while the predicted VAE latent $\hat{\mathbf{z}}$ supplies spatial initialization. Both paths condition a pretrained latent diffusion generator (Rombach et al., 2022) whose weights and conditioning interface remain frozen.

Generation from semantic and spatial predictions. A mixing parameter $\alpha \in [0, 1]$ controls the transition from decoding predicted spatial structure to generation under semantic conditioning. For a schedule $t_0 > \cdots > t_{N_{\mathrm{diff}}-1}$, let $k_\alpha = \lfloor \alpha N_{\mathrm{diff}} \rfloor$ be the number of denoising updates. We initialize the latent as

$$\mathbf{z}^{\mathrm{init}}(\alpha) = \begin{cases} a\hat{\mathbf{z}}, & k_\alpha = 0, \\ a\hat{\mathbf{z}} + \sigma_{t_{N_{\mathrm{diff}}-k_\alpha}} \boldsymbol{\epsilon}, & 0 < k_\alpha < N_{\mathrm{diff}}, \\ \sigma_{\mathrm{init}} \boldsymbol{\epsilon}, & k_\alpha = N_{\mathrm{diff}}, \end{cases} \tag{10}$$

where a is the VAE scaling factor, $\boldsymbol{\epsilon}$ is standard Gaussian noise, and σ_t and σ_{init} are scheduler noise scales. Let $\mathcal{S}_\alpha$ apply the final k_α denoising updates under $\mathbf{c}_{\mathrm{sem}}$, acting as the identity when $k_\alpha = 0$. The reconstructed image is

$$\hat{\mathbf{I}} = D_{\mathrm{VAE}}\big(\mathcal{S}_\alpha\big(\mathbf{z}^{\mathrm{init}}(\alpha); \mathbf{c}_{\mathrm{sem}}\big)/a\big)\,, \tag{11}$$

with frozen VAE decoder D_{VAE}. At $\alpha = 0$, generation reduces to decoding the spatial prediction; at $\alpha = 1$, it starts from noise. Intermediate values combine spatial initialization with denoising under semantic conditioning, with larger α allowing greater departure from the predicted structure.

Training-free guidance. Optional guidance refines the noisy latent to bring its estimated clean latent closer to the brain-predicted VAE latent, maintaining a spatial constraint beyond initialization. This uses neither target images nor weight updates. Only the brain-to-condition predictors require training; objectives are given in Appendix C.2.2, and conditioning interfaces and sampling settings in Appendix C.5.

Table 1: **200-way brain-to-image retrieval (Top-1 / Top-5, %).** Means over ten EEG and four MEG subjects. BHA uses Stage 2 gated scores within subject and Stage 1 outputs under LOSO. CSLS uses all 200 queries as unlabeled references; "–": unreported.

Method	EEG				MEG			
	Within-subject		LOSO		Within-subject		LOSO	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
NICE (Song et al., 2024)	16.1	43.6	6.2	21.4	12.8	36.0	–	–
ATM-S (Li et al., 2024)	28.5	60.4	11.9	33.8	–	–	–	–
CC-All (Zhang et al., 2025)	35.6	80.2	–	–	–	–	–	–
UBP (Wu et al., 2025)	50.9	79.7	12.4	33.4	26.7	55.2	2.2	10.4
HyFI (Jo et al., 2026)	68.2	91.9	15.1	37.2	35.8	64.6	3.2	11.5
NeuroBridge (Zhang et al., 2026a)	63.2	89.9	19.0	45.9	32.2	60.8	3.4	12.8
MBP (Zheng et al., 2026)	75.7	94.6	20.0	44.1	33.9	60.7	**5.4**	15.2
LVBP (Liu et al., 2026)	80.0	96.9	20.0	48.0	44.0	72.0	5.3	15.9
DMSA (Du et al., 2026)	82.6	97.7	21.8	49.4	48.0	74.4	3.7	13.0
BHA (Ours) / Raw	87.00	98.45	22.50	51.60	55.63	85.75	3.63	12.75
BHA (Ours) / CSLS-2	**93.90**	**99.70**	**30.35**	**63.35**	**68.13**	**91.38**	5.25	**16.50**

5 Experiments

5.1 Experimental Setup

We evaluate 200-way brain-to-image retrieval on THINGS-EEG2 (Gifford et al., 2022) and THINGS-MEG (Hebart et al., 2023) under within-subject and leave-one-subject-out (LOSO) settings, and image reconstruction on THINGS-EEG2. Results average ten EEG and four MEG subjects. BrainViT combines channel attention with a feature–sensor convolutional readout; its teachers are OpenCLIP

Table 2: **Visual reconstruction on THINGS-EEG2.** Ten-subject BHA means. Pink: image-conditioned and permutation controls (Appendix D.3.1). Baselines retain their reported generation settings (Appendix D.3.1).

Methods	Low-level Metrics				High-level Metrics			
	PixCorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	EffNet-B↓	SwAV↓
w/ GT image feature ($\alpha = 1.0$)	0.2877	0.4429	0.9472	0.9919	0.9985	0.9998	0.4064	0.2649
w/ GT latent ($\alpha = 0.0$)	**0.9971**	**0.9588**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**0.0168**	**0.0118**
w/ GT feature + latent ($\alpha = 0.5$)	0.8868	0.6840	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.2198	0.1310
Chance (non-target BHA)	0.1115	0.4004	0.5010	0.4750	0.4976	0.4910	0.9782	0.6698
Within-subject (ten-subject mean)								
ATM-S (Li et al., 2024)	0.130	0.339	0.740	0.822	0.685	0.716	-	0.599
CC-All (Zhang et al., 2025)	0.143	0.347	0.754	0.623	0.669	0.715	-	0.590
AVDE (Dai et al., 2026)	0.147	0.366	0.766	0.835	0.724	0.747	-	0.587
MBP (Zheng et al., 2026)	0.195	0.336	0.843	0.905	0.756	0.808	-	0.554
BHA (Ours)	0.277	0.454	**0.874**	**0.922**	**0.791**	**0.830**	**0.845**	0.516
BHA + VAE-guided diffusion	**0.283**	**0.464**	**0.874**	0.921	0.790	**0.830**	**0.845**	**0.515**

ViT-H/14, DFN-6, MAE-18, and DINOv2-12 (Appendices C.3 and C.4). Within-subject retrieval uses Stage 2 gated scores; LOSO matches Stage 1 outputs to fused visual features. Raw scores each query independently. Cross-domain similarity local scaling (CSLS) (Conneau et al., 2017) uses all 200 test queries as unlabeled references, making it transductive. Preprocessing and training are detailed in Appendix C.

5.2 Retrieval Results

BHA achieves within-subject Raw Top-1 accuracy of 87.00% on EEG and 55.63% on MEG, rising to 93.90% and 68.13% with CSLS-2 (Table 1). These Raw scores exceed the reported within-subject baselines in both modalities. Under LOSO, pretrained BrainViT achieves 22.50% and 3.63% Raw Top-1, respectively, identifying cross-subject transfer as a remaining limitation. CSLS rescoring uses unchanged predictions and depends on the reference cohort; $k = 2$ was assessed after benchmark inspection. Subject-wise EEG and MEG results and scoring sensitivity appear in Appendices D.1.1, D.1.2, and D.1.3.

5.3 Reconstruction Results

BHA supports reconstruction with separately pretrained BrainViT weights and a frozen generator (Table 2). Across ten EEG subjects, it achieves PixCorr 0.277, SSIM 0.454, and CLIP 0.830, improving on MBP's reported 0.195, 0.336, and 0.808, respectively (Zheng et al., 2026). VAE guidance further improves pixel and structural fidelity while preserving CLIP identification. These comparisons reflect complete systems, including differences in generators, conditioning interfaces, and sampling settings (Appendix D.3.1). Permuting the predicted conditions reduces identification to near chance, supporting stimulus-specific information in the reconstruction predictions. Evaluation details and subject-wise results appear in Appendix D.3, with examples in Appendix E.3.

5.4 Ablation Studies

The gated readout improves within-subject retrieval over both pretrained BrainViT and the shared expert across training-data scales (Figure 5). Under LOSO, the gated readout underperforms pretrained BrainViT (Table 3). Readout specialization thus benefits within-subject discrimination without necessarily improving cross-subject transfer, motivating our choice of readout for each evaluation setting. Independent readout controls also favor the combined architecture: BHA reaches 87.00% Raw Top-1, versus 85.00% for linear heads and 82.05%/79.60% for shared-only/teacher-only variants (Appendix Table A.4). Encoder search favors local–global sensor convolution, while the main results retain the default feature–sensor readout (Appendix C.3).

Reconstruction benefits depend on the evaluation metric. In matched pretraining controls, fused visual targets yield higher SSIM, whereas OpenCLIP yields higher CLIP identification at stronger generation settings (Appendix D.3.2). These results motivate evaluating structural fidelity and

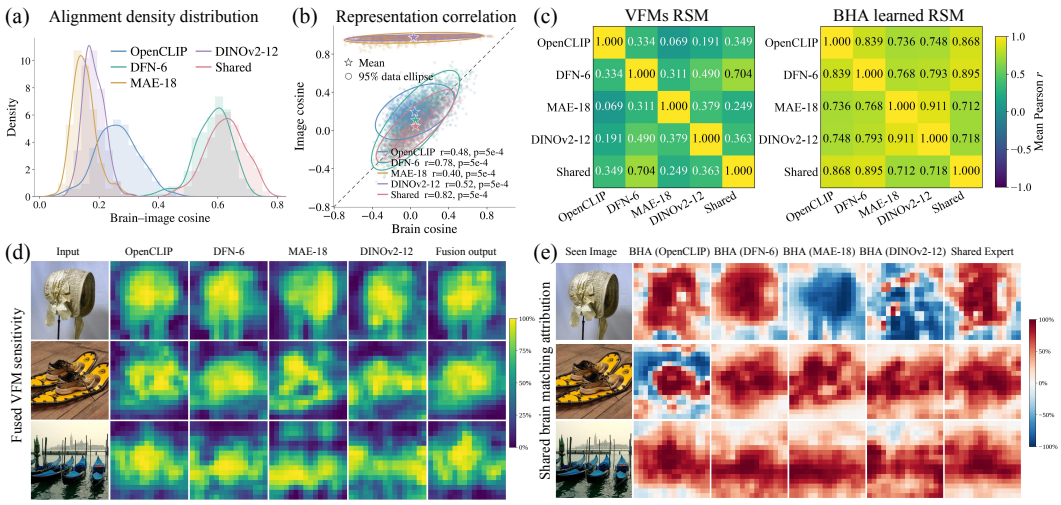


Figure 4: **Shared geometry with distinct matching contributions.** (a) Paired brain–image cosine similarity. (b) Agreement between inter-image similarities in student and visual spaces. (c) Correlations among representational similarity matrices for VFMs (left) and student outputs (right), averaged over ten subjects. (d) Visual sensitivity maps. (e) Occlusion attributions for each teacher branch and the shared expert: red/blue indicate positive/negative contributions to matching. Colors denote within-map percentiles, not comparable effect magnitudes.

Table 3: **Retrieval readouts across training stages.** THINGS-EEG2 Top-1 (%, ↑), full training data.

Stage / readout	Raw	CSLS-2	CSLS-10
Within-subject			
Stage 1 / pretrained	80.80	90.95	87.85
Stage 2 / shared expert	82.20	91.95	88.65
Stage 2 / gated	**87.00**	**93.90**	**91.35**
LOSO			
Stage 1 / pretrained	**22.50**	30.35	**31.65**
Stage 2 / shared expert	22.05	**30.50**	31.55
Stage 2 / gated	20.20	28.15	29.90

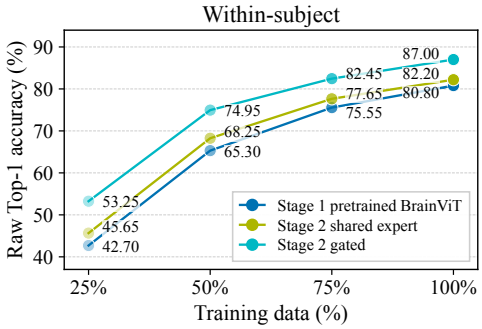


Figure 5: **Readout gains across training-data scales.** THINGS-EEG2 within-subject Raw Top-1 (%); ten-subject means. Each fraction trains a separate backbone.

semantic identification separately. Under BiCLIP, visual adaptation improves paired cosine (0.217 → 0.481) and Raw Top-1 (30.05% → 37.80%), but lowers CLIP identification (0.754 → 0.706). Under MSE–BiCLIP, adaptation lowers both retrieval and reconstruction (Appendix D.3.2). Joint brain–visual adaptation can therefore improve matching without preserving generator compatibility.

5.5 Representation Analysis

5.5.1 Representational Alignment and Matching

We examine whether teacher-specific readouts learn distinct matching functions through paired brain–image similarity, inter-image geometry, and spatial attributions. Student outputs have more similar geometries than their visual targets, with cross-expert correlations of 0.712–0.911 versus 0.069–0.704 (Figure 4(c)). The shared expert also best preserves inter-image relationships ($r = 0.82$; Figure 4(a–b)). Yet overlapping visual sensitivity can yield opposite signed contributions to matching across experts (Figure 4(d–e)). Similar embedding geometry therefore does not imply interchangeable matching functions. The shared geometry may reflect common inputs, backbone, and supervision;

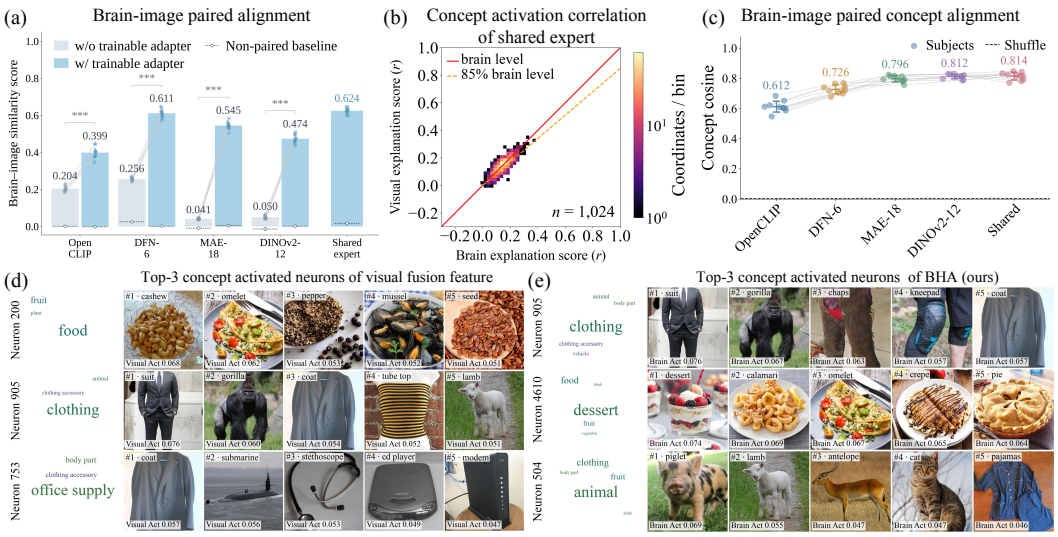


Figure 6: **Feature alignment and semantic selectivity capture different structure.** (a) Paired brain–image similarity before and after visual adaptation. (b) Shared-expert concept correlations on test images; concepts are selected separately for student and visual features using training images (Appendix Figure A.6). (c) Brain–visual concept-profile alignment versus neuron shuffling on training images. (b–c) report subject means. (d–e) Concept associations and activating test images for training-selected visual-fusion and BHA neurons.

these model-level analyses do not establish its neural origin or biological causality. Appendix D.2 details the comparisons.

5.5.2 Concept Associations in Model Outputs

To test whether feature alignment preserves concept structure, we use CLIP-Dissect (Oikarinen & Weng, 2023) to associate model neurons with 27 THINGS categories, selecting concepts on training images and evaluating correlations on held-out images (Figure 6(b)). On training images, the shared expert and DINOv2-12 show the strongest brain–visual concept-profile alignment (mean cosine 0.814 and 0.812), and all branches exceed neuron shuffling (Figure 6(c)). This ordering differs from paired feature alignment (Figure 6(a)), suggesting that concept correspondence captures a distinct property of the learned space. Pooled activations illustrate food-, clothing-, and animal-related associations (Figure 6(d–e)). These results characterize semantic selectivity within learned models; they do not establish corresponding neuron identities across subjects. Appendices D.2 and E.1 provide subject-wise scores and additional galleries. Feature inversion examines recoverable image content under different priors (Appendices D.2.3 and E.2); temporal masking on four datasets probes model dependence on response windows (Appendix D.1.4).

6 Conclusion

In this work, we identify a mismatch between the representational requirements of visual retrieval and reconstruction from brain signals: features that support image discrimination need not preserve the visual structure required by a pretrained generator. This motivates Beyond Hierarchical Alignment (BHA), a multi-teacher framework that learns shared brain–visual alignment while maintaining distinct prediction spaces for teacher-specific retrieval and semantic and spatial reconstruction. Composing shared and teacher-specific scores improves within-subject retrieval, while generator-native conditions enable reconstruction without generator fine-tuning. Representation analyses reveal distinct matching contributions and concept associations despite overlapping feature geometry. The gains are task-dependent: readout specialization does not consistently improve cross-subject transfer, and richer pretraining does not improve every reconstruction metric. These findings motivate preserving complementary prediction spaces while evaluating discrimination, reconstruction fidelity, and cross-subject generalization separately.

REPRODUCIBILITY STATEMENT

Proofs appear in Appendix B; datasets, preprocessing, architecture, objectives, and inference settings are specified in Appendix C. Appendix D provides subject-wise results, ablations, and analysis protocols. Source code will be released publicly upon acceptance.

ETHICS STATEMENT

This study uses only existing recordings from four publicly available neural datasets: THINGS-EEG1, THINGS-EEG2, THINGS-MEG, and Alljoined-1.6M. We did not recruit participants, conduct new human-subject experiments, or collect new neural recordings.

AI USE STATEMENT

Large language model tools assisted manuscript drafting and revision, literature search and comparison, experimental design and methodological feedback, implementation of experimental and analysis code, and result interpretation. AI assistance also contributed to the formulation of mathematical claims and the development of supporting derivations and proofs. These tools further helped check consistency among the manuscript, experimental settings, model outputs, reported results, and their sources. The authors remain responsible for all manuscript content, results, and scientific claims, including the correctness of the mathematical statements and proofs developed with AI assistance.

REFERENCES

Yunpeng Bai, Xintao Wang, Yan-Pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: High-quality eeg-to-image generation with temporal masked signal modeling and clip alignment. In *European Conference on Computer Vision*, pp. 472–488. Springer, 2024.

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.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 60884–60937. Curran Associates, Inc., 2025. doi: 10.52202/085713-2036. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/57bc0a850255e2041341bf74c7e2b9fa-Paper-Conference.pdf.

Yue Cao, Yong Huang, Wei Zhu, Yangzhou Liu, Zhe Chen, Guangchen Shi, Yong Fa, Yujie Yang, Song Mei, and Tong Lu. Mmfuser: Multimodal multi-layer feature fuser for fine-grained vision-language understanding. In *International Conference on Pattern Recognition*, pp. 406–421. Springer, 2026.

Sofian Chaybouti, Sanath Narayan, Yasser Dahou, Phúc H. Lê Khac, Ankit Singh, Ngoc Dung Huynh, Wamiq Reyaz Para, Hilde Kuehne, and Hakim Hacid. Siglino: Efficient multi-teacher distillation for agglomerative vision foundation models, 2025. URL http://arxiv.org/abs/2512.20157v2.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22710–22720. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.02175. URL http://dx.doi.org/10.1109/cvpr52729.2023.02175.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.00276. URL http://dx.doi.org/10.1109/cvpr52729.2023.00276.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data, 2017. URL http://arxiv.org/abs/1710.04087v3.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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, Sept 2023. ISSN 1939-3539. doi: 10.1109/tpami.2023.3263181. URL http://dx.doi.org/10.1109/tpami.2023.3263181.

Yang Du, Siyuan Dai, Yonghao Song, Paul M. Thompson, Haoteng Tang, and Liang Zhan. Deep models, shallow alignment: Uncovering the granularity mismatch in neural decoding, 2026. URL http://arxiv.org/abs/2601.21948v2.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations*, volume 2024, pp. 36221–36237, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Honghao Fu, Hao Wang, Jing Jih Chin, and Zhiqi Shen. Brainvis: Exploring the bridge between brain and visual signals via image reconstruction. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, Apr 2025. doi: 10.1109/icassp49660.2025.10889805. URL http://dx.doi.org/10.1109/icassp49660.2025.10889805.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 50742–50768. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2208. URL http://dx.doi.org/10.52202/075280-2208.

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, Dec 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2022.119754. URL http://dx.doi.org/10.1016/j.neuroimage.2022.119754.

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), Jan 2022. ISSN 2052-4463. doi: 10.1038/s41597-021-01102-7. URL http://dx.doi.org/10.1038/s41597-021-01102-7.

Matthias Guggenmos, Philipp Sterzer, and Radoslaw Martin Cichy. Multivariate pattern analysis for meg: A comparison of dissimilarity measures. *NeuroImage*, 173:434–447, June 2018. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2018.02.044. URL http://dx.doi.org/10.1016/j.neuroimage.2018.02.044.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988. IEEE, June 2022. doi: 10.1109/cvpr52688.2022.01553. URL http://dx.doi.org/10.1109/cvpr52688.2022.01553.

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, feb 2023. ISSN 2050-084X. doi: 10.7554/eLife.82580. URL https://doi.org/10.7554/eLife.82580.

Greg Heinrich, Mike Ranzinger, Hongxu Danny Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2.5: Improved baselines for agglomerative vision foundation models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22487–22497. IEEE, June 2025. doi: 10.1109/cvpr52734.2025.02094. URL http://dx.doi.org/10.1109/cvpr52734.2025.02094.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1), May 2017. ISSN 2041-1723. doi: 10.1038/ncomms15037. URL http://dx.doi.org/10.1038/ncomms15037.

Sangmin Jo, Wootaek Jeong, Da-Woon Heo, Yoohwan Hwang, and Heung-Il Suk. Hyfi: Hyperbolic feature interpolation for brain-vision alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(7):5575–5583, Mar 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i7.37476. URL http://dx.doi.org/10.1609/aaai.v40i7.37476.

Blair Kaneshiro, Marcos Perreau Guimaraes, Hyung-Suk Kim, Anthony M. Norcia, and Patrick Suppes. A representational similarity analysis of the dynamics of object processing using single-trial eeg classification. *PLOS ONE*, 10(8):e0135697, Aug 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0135697. URL http://dx.doi.org/10.1371/journal.pone.0135697.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, pp. 113–132. Springer, 2024.

Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1809–1817, 2017.

Reese Kneeland, Wangshu Jiang, Ugo Bruzadin Nunes, Paul Steven Scotti, Arnaud Delorme, and Jonathan Xu. Enigma: Eeg-to-image in 15 minutes using less than 1% of the parameters. *arXiv preprint arXiv:2602.10361*, 2026.

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 37*, NeurIPS 2024, pp. 102822–102864. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-3266. URL http://dx.doi.org/10.52202/079017-3266.

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*, MM '25, pp. 5577–5586. ACM, Oct 2025a. doi: 10.1145/3746027.3754996. URL http://dx.doi.org/10.1145/3746027.3754996.

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 *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025b. doi: 10.1109/ICME59968.2025.11210130.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pp. 29624–29636. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi: 10.52202/068431-2148. URL http://dx.doi.org/10.52202/068431-2148.

Wenchao Liu, Hongwei Li, Zhouyang Xu, Lin Ma, and Haifeng Li. Leveraging visual blur perception characteristics for eeg decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17580–17588, Mar 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i21.38813. URL http://dx.doi.org/10.1609/aaai.v40i21.38813.

Yuxiang Lu, Shengcao Cao, and Yu-Xiong Wang. Swiss army knife: Synergizing biases in knowledge from vision foundation models for multi-task learning. In *International Conference on Learning Representations*, volume 2025, pp. 80538–80574, 2025.

David Mayo, Christopher Wang, Asa Harbin, Abdulrahman Alabdulkareem, Albert Shaw, Boris Katz, and Andrei Barbu. Brainbits: How much of the brain are generative reconstruction methods using? In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 54396–54420. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-1723. URL http://dx.doi.org/10.52202/079017-1723.

Tuomas Oikarinen and Tsui-Wei Weng. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=iPWiwWHc1V.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. 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, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. The photographer's eye: Teaching multimodal large language models to see and critique like photographers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24807–24816. IEEE, June 2025. doi: 10.1109/cvpr52734.2025.02310. URL http://dx.doi.org/10.1109/cvpr52734.2025.02310.

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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12490–12500. IEEE, June 2024. doi: 10.1109/cvpr52733.2024.01187. URL http://dx.doi.org/10.1109/cvpr52733.2024.01187.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685. IEEE, June 2022. doi: 10.1109/cvpr52688.2022.01042. URL http://dx.doi.org/10.1109/cvpr52688.2022.01042.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, Aug 2017. ISSN 1097-0193. doi: 10.1002/hbm.23730. URL http://dx.doi.org/10.1002/hbm.23730.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors.

In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 24705–24728. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-1073. URL http://dx.doi.org/10.52202/075280-1073.

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 Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/scotti24a.html.

Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 724–748. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/shang25a.html.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 42048–42070. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-1330. URL http://dx.doi.org/10.52202/079017-1330.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *International Conference on Learning Representations*, volume 2025, pp. 92628–92646, 2025.

Irina Simanova, Marcel van Gerven, Robert Oostenveld, and Peter Hagoort. Identifying object categories from event-related eeg: Toward decoding of conceptual representations. *PLoS ONE*, 5(12):e14465, Dec 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0014465. URL http://dx.doi.org/10.1371/journal.pone.0014465.

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.

C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah. Deep learning human mind for automated visual classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4503–4511. IEEE, July 2017. doi: 10.1109/cvpr.2017.479. URL http://dx.doi.org/10.1109/cvpr.2017.479.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14453–14463, 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 http://arxiv.org/abs/2603.07077v1.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann Lecun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 87310–87356. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-2771. URL http://dx.doi.org/10.52202/079017-2771.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2246–2257. IEEE, June 2025. doi: 10.1109/cvpr52734.2025.00215. URL http://dx.doi.org/10.1109/cvpr52734.2025.00215.

Weihao Xia and Cengiz Oztireli. Exploring the visual feature space for multimodal neural decoding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4370–4379. IEEE, Oct 2025. doi: 10.1109/iccv51701.2025.00416. URL http://dx.doi.org/10.1109/iccv51701.2025.00416.

Jonathan Xu, Ugo Bruzadin Nunes, Wangshu Jiang, Samuel Ryther, Jordan Pringle, Paul S. Scotti, Arnaud Delorme, and Reese Kneeland. Alljoined-1.6m: A million-trial eeg-image dataset for evaluating affordable brain-computer interfaces, 2025. URL https://arxiv.org/abs/2508.18571.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL http://arxiv.org/abs/2308.06721v1.

Bohan Zeng, Shanglin Li, Xuhui Liu, Sicheng Gao, Xiaolong Jiang, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. Controllable mind visual diffusion model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):6935–6943, Mar 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i7.28519. URL http://dx.doi.org/10.1609/aaai.v38i7.28519.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):14486–14493, Apr 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i13.33587. URL http://dx.doi.org/10.1609/aaai.v39i13.33587.

Wenjiang Zhang, Sifeng Wang, Yuwei Su, Xinyu Li, Chen Zhang, and Suyu Zhong. Neurobridge: Bio-inspired self-supervised eeg-to-image decoding via cognitive priors and bidirectional semantic alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):18028–18036, Mar 2026a. ISSN 2159-5399. doi: 10.1609/aaai.v40i21.38863. URL http://dx.doi.org/10.1609/aaai.v40i21.38863.

Xu Zhang, Ruijie Quan, Wenguan Wang, and Yi Yang. Moving beyond diffusion: Hierarchy-to-hierarchy autoregression for fmri-to-image reconstruction. In *International Conference on Learning Representations*, volume 2026, pp. 68490–68515, 2026b.

Jiawen Zheng, Haonan Jia, MING LI, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 121474–121496, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c5506f154e5f2203a7dca2bee2faeced-Paper-Conference.pdf.

Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 103305–103333. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-3282. URL http://dx.doi.org/10.52202/079017-3282.

SUPPLEMENTARY MATERIAL:

BEYOND HIERARCHICAL ALIGNMENT: DISTILLING TEACHER DIVERSITY FOR BRAIN VISUAL DECODING

A APPENDIX OVERVIEW

The contents below link to the appendix sections and subsections; page numbers appear on the right.

B Theoretical Analysis 16
B.1 Proof of Proposition 1 16
B.2 Proof of Proposition 2 18
C Implementation Details 19
C.1 Datasets and Settings 19
C.2 Detailed Objectives and Training 20
C.3 Model Architecture 21
C.4 Retrieval Implementation 22
C.5 Reconstruction Implementation 24
D Additional Experimental Results 25
D.1 Comprehensive BHA Results 25
D.2 Extended Representation Analysis 29
D.3 Comprehensive Reconstruction Performance 35
E Supplementary Examples 40
E.1 Neuron Concept Associations 40
E.2 Feature Inversion Examples 46
E.3 Reconstruction Examples 49

B THEORETICAL ANALYSIS

We first establish the alignment-risk bound of Proposition 1, then prove Proposition 2 by separating retrieval alignment to adapted targets from fidelity to native generator outputs.

B.1 PROOF OF PROPOSITION 1

General weighted form. Retain the signal model and independence assumptions of Proposition 1, but allow unequal teacher error variances. Define the error covariance matrix through $C_{k\ell} = \mathbb{E}[\boldsymbol{\delta}_k^\top \boldsymbol{\delta}_\ell]/d_0$. Choose weights $\mathbf{w} \in \Delta_K = \{\mathbf{w} \geq 0 : \sum_k w_k = 1\}$ independently of the training samples and fit

$$\widehat{\mathbf{M}}_{\mathbf{w}} = \arg\min_{\mathbf{M}} \frac{1}{nd_0} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k \|\mathbf{M}\mathbf{b}_i - \mathbf{t}_{ik}\|_2^2. \tag{12}$$

With $\widehat{g}_{\mathbf{w}}(\mathbf{b}) = \widehat{\mathbf{M}}_{\mathbf{w}}\mathbf{b}$, we will show

$$\begin{aligned} \mathbb{E}_{\text{train}}[\mathcal{R}_{\text{vis}}(\widehat{g}_{\mathbf{w}}) \mid \mathsf{B}] &= \sigma_{\text{br}}^2 + \chi_n\big(\sigma_{\text{br}}^2 + \mathbf{w}^\top \mathbf{C}\mathbf{w}\big) \\ &\leq \sigma_{\text{br}}^2 + \frac{p}{n\kappa}\big(\sigma_{\text{br}}^2 + \mathbf{w}^\top \mathbf{C}\mathbf{w}\big). \end{aligned} \tag{13}$$

Uniform weights recover $\widehat{g}_K$ in the proposition.

Predictability floor and fused supervision. Since $\mathbb{E}[\boldsymbol{\xi} \mid \mathbf{b}] = 0$, the conditional mean $\mathbf{M}_\star \mathbf{b}$ minimizes squared prediction error, and

$$\mathcal{R}_{\text{vis}}(g) = \sigma_{\text{br}}^2 + \frac{1}{d_0}\mathbb{E}\|g(\mathbf{b}) - \mathbf{M}_\star \mathbf{b}\|_2^2.$$

This proves the lower bound. Define $\bar{\mathbf{t}}_{i,\mathbf{w}} = \sum_k w_k \mathbf{t}_{ik}$. For any prediction $\mathbf{v}$,

$$\sum_k w_k \|\mathbf{v} - \mathbf{t}_{ik}\|_2^2 = \|\mathbf{v} - \bar{\mathbf{t}}_{i,\mathbf{w}}\|_2^2 + \sum_k w_k \|\mathbf{t}_{ik} - \bar{\mathbf{t}}_{i,\mathbf{w}}\|_2^2.$$

The second term is independent of the prediction. Thus multi-teacher least squares is equivalent to regression on the fused target, whose residual is $\boldsymbol{\eta}_{i,\mathbf{w}} = \boldsymbol{\xi}_i + \sum_k w_k \boldsymbol{\delta}_{ik}$. Independence and zero means give $\mathbb{E}\|\boldsymbol{\eta}_{i,\mathbf{w}}\|_2^2/d_0 = \sigma_{\mathrm{br}}^2 + \mathbf{w}^\top \mathbf{C}\mathbf{w}$.

Finite-sample alignment error. Let $\mathsf{N}_\mathbf{w} \in \mathbb{R}^{n \times d_0}$ contain the residuals $\boldsymbol{\eta}_{i,\mathbf{w}}^\top$ as rows. The full-rank normal equations yield

$$\widehat{\mathbf{M}}_\mathbf{w} - \mathbf{M}_\star = \mathsf{N}_\mathbf{w}^\top \mathsf{B}(\mathsf{B}^\top \mathsf{B})^{-1}.$$

For an independent test feature, $\mathbb{E}[\mathbf{b}\mathbf{b}^\top] = I_p$ implies that its prediction error relative to the conditional mean is $\|\widehat{\mathbf{M}}_\mathbf{w} - \mathbf{M}_\star\|_F^2/d_0$. Conditional on B, the training residuals are independent with zero means, so

$$\mathbb{E}[\mathsf{N}_\mathbf{w}\mathsf{N}_\mathbf{w}^\top \mid \mathsf{B}] = d_0(\sigma_{\mathrm{br}}^2 + \mathbf{w}^\top \mathbf{C}\mathbf{w})I_n.$$

Substitution into the squared Frobenius norm gives

$$\frac{1}{d_0}\mathbb{E}_{\mathrm{train}}\Big[\|\widehat{\mathbf{M}}_\mathbf{w} - \mathbf{M}_\star\|_F^2 \mid \mathsf{B}\Big] = (\sigma_{\mathrm{br}}^2 + \mathbf{w}^\top \mathbf{C}\mathbf{w})\chi_n.$$

Adding the independent test residual proves the equality in Eq. (13). Every eigenvalue of $\mathsf{B}^\top\mathsf{B}$ is at least $n\kappa$, hence $\chi_n \le p/(n\kappa)$, proving the upper bound.

When multiple teachers improve alignment. For uniform weights, the covariance assumptions give

$$\mathbf{w}^\top \mathbf{C}\mathbf{w} = \frac{1}{K^2}\sum_{k,\ell} C_{k\ell} \le \sigma_{\mathrm{T}}^2\left(\rho_{\mathrm{T}} + \frac{1-\rho_{\mathrm{T}}}{K}\right) < \sigma_{\mathrm{T}}^2.$$

Since $\chi_n > 0$, Eq. (13) establishes strict improvement over every single teacher with variance σ_{T}^2, completing the proof. For unequal teacher quality, uniform averaging need not improve on the best teacher. Oracle weights instead satisfy $\min_{\mathbf{w}\in\Delta_K} \mathbf{w}^\top\mathbf{C}\mathbf{w} \le \min_k C_{kk}$, because each single-teacher choice belongs to the simplex; this does not guarantee that weights learned from finite data attain the oracle risk.

Relation to angular alignment. For any nonzero vectors u, v, the triangle and reverse-triangle inequalities give

$$\|\operatorname{norm}(u) - \operatorname{norm}(v)\|_2 \le \frac{2\|u-v\|_2}{\|v\|_2}.$$

If $g(\mathbf{b}) \neq 0$ and $\|\mathbf{y}\|_2 \ge r_{\min} > 0$ almost surely, then

$$\mathbb{E}\big[1 - \operatorname{sim}(g(\mathbf{b}), \mathbf{y})\big] \le \frac{2d_0}{r_{\min}^2}\mathcal{R}_{\mathrm{vis}}(g). \tag{14}$$

This follows by squaring the preceding inequality and using $1 - \operatorname{sim}(u,v) = \|\operatorname{norm}(u) - \operatorname{norm}(v)\|_2^2/2$. The same inequality holds after averaging over training samples, conditional on the design. Thus a tighter feature-risk bound also tightens this angular-error bound under the additional norm condition; it need not improve the realized cosine similarity or retrieval ranking, which also depends on similarities to unmatched images.

Implications for brain–vision alignment. Complementarity is useful here when teacher errors are not fully shared: multiple sources reduce estimation error around the brain-predictable visual component, especially when n is small relative to p. They do not reduce the irreducible term σ_{br}^2; identical teacher errors provide no averaging benefit. The result assumes unbiased, commensurate teacher targets and a well-conditioned linear design. It motivates learning complementary visual supervision, but does not establish these assumptions for BHA's learned, normalized fusion or give a guarantee for its BiCLIP objective, retrieval accuracy, or reconstruction quality. In particular, lower alignment risk supports better estimation of recoverable visual information, not complete recovery of stimulus content.

B.2 Proof of Proposition 2

Alignment error and the native reconstruction space. The decoder G includes the frozen conditioning interface and generator; $G(\mathbf{e})$ is its output under the original visual feature, not the ground-truth image. The proposition compares this native output with outputs driven by adapted visual features and brain predictions. All three features must lie in the decoder's input domain. For a stochastic generator, fix its random input, or average over a shared random input independent of the features, assuming the same Lipschitz bound.

Proof. For a square-integrable random vector $\mathbf{A}$, write $\|\mathbf{A}\|_{L^2} = (\mathbb{E}\|\mathbf{A}\|_2^2)^{1/2}$. Applying the reverse triangle inequality in L^2 gives

$$\begin{aligned}\left|\sqrt{\mathcal{R}_G} - \sqrt{\Delta_G}\right| &= \left|\|G(\mathbf{h}) - G(\mathbf{e})\|_{L^2} - \|G(\mathbf{t}) - G(\mathbf{e})\|_{L^2}\right| \\ &\le \|G(\mathbf{h}) - G(\mathbf{t})\|_{L^2} \le L_G\|\mathbf{h} - \mathbf{t}\|_{L^2} = L_G\sqrt{2\delta}. \end{aligned} \tag{15}$$

The last equality uses unit norms. If $\delta = 0$, then $\mathbf{h} = \mathbf{t}$ almost surely and $\mathcal{R}_G = \Delta_G$. □

For unnormalized predictions and targets, the same proof replaces $\sqrt{2\delta}$ by $(\mathbb{E}\|\mathbf{h} - \mathbf{t}\|_2^2)^{1/2}$; cosine similarity alone does not constrain their magnitudes. The bound holds for each jointly learned pair (f, V). Better angular alignment brings the brain-conditioned output closer to the adapted visual output; it does not constrain how far that output moves from the native one. In particular, $\mathcal{R}_G \ge (\sqrt{\Delta_G} - L_G\sqrt{2\delta})_+^2$, where $(x)_+ = \max\{x, 0\}$. To relate the statement to ground-truth reconstruction, let $\mathcal{R}_{\text{native}} = \mathbb{E}\|G(\mathbf{e}) - \mathbf{I}\|_2^2$. Another triangle inequality yields

$$\mathbb{E}\|G(\mathbf{h}) - \mathbf{I}\|_2^2 \ge \left(\sqrt{\Delta_G} - L_G\sqrt{2\delta} - \sqrt{\mathcal{R}_{\text{native}}}\right)_+^2. \tag{16}$$

Thus a sufficiently large adaptation-induced drift cannot be removed by perfecting alignment to the adapted target alone.

Joint brain–visual adaptation and reconstruction. Let C index an image category, $\mathbf{u}_C$ be its standard basis vector, and U an independent uniform sign representing a visual detail. The native feature is $\mathbf{e} = (\mathbf{u}_C, U)/\sqrt{2}$. Let the brain input contain (C, W), where $W = SU$ is a noisy observation of that detail: $S \in \{-1, +1\}$ is independent of (C, U) and $\mathbb{E}[S] = c$. Choose $1/2 < \beta < c < 1$. For $q \in [0, 1]$, define

$$\mathbf{t}_q = \frac{(\mathbf{u}_C, qU)}{\sqrt{1+q^2}}, \qquad \mathbf{h}_q = \frac{(\mathbf{u}_C, \beta q W)}{\sqrt{1+\beta^2 q^2}}, \qquad G(\mathbf{z}) = \mathbf{z}. \tag{17}$$

As q decreases, both sides suppress a detail represented more weakly and noisily in the brain prediction. With the native feature itself as the reconstruction target, the expected angular and reconstruction errors are

$$\delta(q) = 1 - \frac{1 + \beta c q^2}{\sqrt{(1+q^2)(1+\beta^2 q^2)}}, \qquad \mathcal{R}_G(q) = 2 - \frac{\sqrt{2}(1+\beta c q)}{\sqrt{1+\beta^2 q^2}}. \tag{18}$$

Conditional on $S = +1$ or $S = -1$, the paired angle is respectively $\vartheta_+(q) = \arctan(q) - \arctan(\beta q)$ or $\vartheta_-(q) = \arctan(q) + \arctan(\beta q)$. Both increase strictly with q: in particular, $\vartheta_+'(q) = (1-\beta)(1-\beta q^2)/[(1+q^2)(1+\beta^2 q^2)] > 0$. Hence $\delta(q)$ strictly decreases as q decreases from one to zero, whereas $\mathcal{R}_G'(q) = -\sqrt{2}\beta(c - \beta q)/(1+\beta^2 q^2)^{3/2} < 0$ shows that reconstruction error strictly increases. At $q = 0$, alignment is perfect but $\mathcal{R}_G = \Delta_G = 2 - \sqrt{2} > 0$.

To check ranking separately, take a two-image gallery with distinct categories and independent draws of their details and observation noise. At $q = 1$, a brain-to-image query is misranked exactly when its detail is observed incorrectly and the distractor has the opposite detail. An image-to-brain query is misranked exactly when its paired brain detail is incorrect and the distractor brain detail agrees with the query image. Since $\beta > 1/2$, each event makes the unmatched score exceed the matched score; each has probability $(1 - c)/4$. Expected Top-1 accuracy in both directions therefore increases from $(3 + c)/4 < 1$ at $q = 1$ to one at $q = 0$, while reconstruction worsens. At $q = 0$, both embeddings discard U; even an unrestricted decoder then has minimum expected error $1/2$ for recovering $\mathbf{e}$, the variance of its missing last component. For $q > 0$, the corresponding minima are zero from $\mathbf{t}_q$ and $(1 - c^2)/2$ from $\mathbf{h}_q$.

Implications for task-specific prediction. The proposition concerns fidelity to the generator's native output under an explicit regularity assumption; it does not assert that every visual adapter destroys semantic or spatial information. Invertible adaptation can alter generator compatibility without losing information, whereas a many-to-one mapping can remove content altogether. The constructed family shows that alignment improvement and reconstruction degradation can coexist, not that an optimizer must follow this path. For BHA, this motivates predicting native semantic conditions and VAE latents through separate adapters on the shared backbone. These predictors bypass the adapted retrieval spaces and use information retained in backbone features; they cannot recover content already discarded by the backbone.

C Implementation Details

Experiments used eight NVIDIA A800 GPUs. This section describes the datasets, training objectives, architectures, and inference procedures. EEG retrieval hyperparameters are reported with the experimental settings in Section D.1.1 (Table A.5).

C.1 Datasets and Settings

THINGS-EEG2. THINGS-EEG2 contains recordings from 10 participants, with 16,540 training images repeated four times and 200 test images repeated 80 times per participant (Gifford et al., 2022). We band-pass filter the EEG at 0.1–100 Hz, apply baseline correction using the 200 ms pre-stimulus interval, and downsample from 1,000 to 250 Hz. We retain 63 EEG channels and 250 time samples covering the first second after stimulus onset, and apply multivariate noise normalization (Guggenmos et al., 2018).

Training and validation inputs average four trials per image; test inputs average 80 trials within each subject, including the held-out subject in LOSO. Each input contains $C \times 250$ values, with $C \in \{63, 17\}$ depending on the prediction branch. Trial-count ablations vary the number of averaged responses.

Pretraining samples one channel configuration per minibatch; distillation uses all 63 channels. The 17-channel posterior subset comprises **P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, and O2**, in the stored channel order. The low-level reconstruction branch selects these channels before neural encoding, whereas the high-level branch retains the full montage during both training and inference.

THINGS-MEG. THINGS-MEG contains 271-channel recordings from four subjects over 12 sessions (Hebart et al., 2023). The stimulus pool covers 1,854 concepts with 12 images per concept. We exclude the 200 test concepts from training; each test concept has one image presented 12 times. Images are shown for 500 ms with an onset interval of $1{,}000 \pm 200$ ms. We apply a 0.1–40 Hz band-pass filter and baseline correction, downsample to 200 Hz, and use the 0–1,000 ms interval. Inputs contain 271×201 values, including both temporal endpoints. All retrieval models use the full sensor montage and match each response to its presented image.

Additional datasets for temporal analysis. THINGS-EEG1 and Alljoined-1.6M extend the temporal-window comparison to different stimulus timing and recording systems (Appendix D.1.4).

THINGS-EEG1. THINGS-EEG1 contains 64-channel EEG from 50 participants viewing 22,248 images from 1,854 concepts in rapid serial visual presentation at 10 Hz (Grootswagers et al., 2022). We use the 46 participants retained by the dataset's quality criteria. The 200 test images have 12 repetitions each and are distinct from training images, although their concepts can overlap. We use the released average-referenced, 0.1–100 Hz filtered recordings at 250 Hz, apply a $[-100, 0)$ ms baseline correction, and retain $[0, 1000)$ ms epochs, giving 64×250 inputs. Test queries average all 12 repetitions. At this presentation rate, each epoch includes responses to adjacent stimuli; the analysis therefore measures decoding within the rapid presentation stream.

Alljoined-1.6M. Alljoined-1.6M contains over 1.6 million trials from 20 participants, recorded over four sessions per participant with a 32-channel Emotiv Flex 2 system (Xu et al., 2025). It uses

the THINGS-EEG2 stimulus split: 16,540 training images and 200 test images from disjoint concepts. Images are presented for 100 ms followed by a 100 ms blank interval. Recordings are sampled at 256 Hz and resampled to 250 Hz, with a 200 ms pre-stimulus baseline correction and session-wise MVNN estimated from training epochs. We retain 250 post-stimulus samples per channel (32×250 inputs) and average all retained repetitions of each image; test queries contain approximately 80 repetitions. Temporal interpretation follows the timing limitations described in Appendix D.1.4.

C.2 Detailed Objectives and Training

This section details BHA's retrieval and reconstruction training objectives. We use the MSE and BiCLIP definitions in Eqs. (1)–(2); architecture and numerical settings are given in Appendices C.3–C.5.

C.2.1 Two-stage retrieval training

EEG pretraining uses paired BiCLIP on same-image trial averages. Distillation and semantic reconstruction use image-ID positives: when an image occurs more than once in a minibatch, its positive softmax probabilities are summed before taking the logarithm in each direction of Eq. (2). Distinct images of the same concept remain negatives. MEG-specific pretraining is specified in Appendix D.1.2.

Stage 1: Pretraining. We optimize Eq. (5) jointly over BrainViT, source projections W_k, fusion map F, and contrastive temperature. The source VFMs remain frozen. Both stages use the optimization settings in Table A.5 and the shared data split specified in Section D.1.1; the learned fused representation serves as the distillation target without further adaptation.

Stage 2: Multi-Teacher Distillation. We keep BrainViT, source VFMs, and the pretrained visual fusion frozen, and train teacher-specific visual adapters V_k jointly with the retrieval adapter, shared expert, Group MoEs, outer router, gate, and logit scales. The shared expert matches the frozen fused feature; each native group matches its adapted teacher feature. Both alignment and combined-score losses backpropagate into V_k.

Native alignment averages BiCLIP over teacher groups, with $\lambda_k = 0.25/K$ for $k \geq 1$ and $\lambda_0 = 1$. The combined-score loss is $\mathcal{L}_{\text{ret}} = \mathcal{L}_{\text{BiCLIP}}(\mathbf{S}/\tau_{\text{ret}})$, where $\mathbf{S}$ is defined in Eq. (9) and $\tau_{\text{ret}} > 0$ is learned. To limit changes to teacher geometry, visual adapters receive the regularizer

$$\mathcal{L}_{\text{drift}} = \frac{1}{BK} \sum_{i=1}^{B} \sum_{k=1}^{K} \left[1 - \text{sim}(\tilde{\mathbf{e}}_{ik}, \text{sg}(\mathbf{e}_{ik}))\right]. \tag{19}$$

We use $\lambda_{\text{ret}} = 1$, $\lambda_{\text{aux}} = 0.01$, and $\lambda_{\text{drift}} = 0.1$. Architecture and training settings appear in Tables A.3 and A.5.

For internal router logits $G_k(\mathbf{h}'_i) \in \mathbb{R}^R$, define $\mathbf{w}_{ik} = (w_{ik1}, \ldots, w_{ikR})^\top$; $\mathbf{1} \in \mathbb{R}^R$ is the all-ones vector. The weights and usage penalty are

$$w_{ikr} = (1 - \rho)\,\text{softmax}(G_k(\mathbf{h}'_i)/\tau_g)_r + \rho/R, \tag{20}$$

$$\mathcal{L}_{\text{aux}} = \frac{R}{K} \sum_{k=1}^{K} \|\bar{\mathbf{w}}_k - \mathbf{1}/R\|_2^2, \qquad \bar{\mathbf{w}}_k = B^{-1} \sum_i \mathbf{w}_{ik}. \tag{21}$$

Here $\rho = 0.02$ is the uniform-mixing ratio, giving a minimum routing weight of ρ/R; τ_g controls routing softness. Only the R routed experts enter this normalization and balance penalty; the always-active internal shared expert does not. This soft-usage penalty encourages balanced routing (Shazeer et al., 2017; Fedus et al., 2022). Scoring and evaluation are specified in Appendix C.4.

C.2.2 Reconstruction training

The semantic and spatial predictors use the pretrained backbone with separate trainable parameters ψ_{H} and ψ_{L}:

$$\hat{\mathbf{e}}_i^{\text{sem}} = F_{\text{H}}(\mathbf{X}_i, s_i; \theta, \psi_{\text{H}}), \qquad \hat{\mathbf{z}}_i^{(L)} = F_{\text{L}}(\mathbf{X}_i, s_i; \theta, \psi_{\text{L}}). \tag{22}$$

Backbone weights θ and all generator components remain frozen. Targets are OpenCLIP features $\mathbf{e}_i^{\text{sem}} = \phi_{\text{sem}}(\mathbf{I}_i)$ and unscaled VAE latents $\mathbf{z}_i^{(L)} = E_{\text{VAE}}(\mathbf{I}_i)$, where L is the finest pyramid level. Outside the pyramid we write $\hat{\mathbf{z}}_i = \hat{\mathbf{z}}_i^{(L)}$. Cosine losses average $d_{\cos}(\mathbf{u}, \mathbf{v}) = 1 - \text{sim}(\mathbf{u}, \mathbf{v})$ over samples, flattening tensor-valued features.

High-level conditions. The semantic predictor combines MSE and image-ID BiCLIP on OpenCLIP features with MSE and cosine distance after the frozen IP-Adapter projection (Ye et al., 2023). Architecture and loss weights are given in Appendix C.5.

Low-level conditions. We first train a coarse VAE-latent predictor, then a conditional pyramid that refines its spatial resolution. Area downsampling of the finest VAE target defines $\mathbf{z}^{(p)}$, and bilinear upsampling $\mathcal{U}_p$ defines residual targets $\mathbf{r}^{(p)} = \mathbf{z}^{(p)} - \mathcal{U}_p(\mathbf{z}^{(p-1)})$. At each refinement level $p = 1, \ldots, L$, the predictor produces

$$\hat{\mathbf{z}}^{(p)} = \mathcal{U}_p(\hat{\mathbf{z}}^{(p-1)}) + \gamma_p \hat{\mathbf{r}}^{(p)}, \tag{23}$$

where $p = 0$ denotes the coarse level and γ_p is a learned scalar. Residual supervision is applied to $\hat{\mathbf{r}}^{(p)}$ before this scaling. Loss coefficients, target normalization, and the two-phase training schedule are listed in Appendix C.5.

Frozen-generator inference. An Euler ancestral sampler uses the semantic condition $\mathbf{c}_{\text{sem}} = P_{\text{IP}}(\hat{\mathbf{e}}^{\text{sem}})$ and spatial initialization in Eqs. (10) and (11). Its noise scale satisfies $\sigma_t^2 = (1 - \bar{\alpha}_t)/\bar{\alpha}_t$, where $\bar{\alpha}_t$ is the cumulative signal-retention factor; at interpolated inference timesteps, $\bar{\alpha}_t = (1 + \sigma_t^2)^{-1}$. The denoiser receives the normalized Euler latent, giving

$$\frac{a\hat{\mathbf{z}}_i + \sigma_t \boldsymbol{\epsilon}_i}{\sqrt{1 + \sigma_t^2}} = \sqrt{\bar{\alpha}_t}\, a\hat{\mathbf{z}}_i + \sqrt{1 - \bar{\alpha}_t}\, \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_{d_z}).$$

Here d_z is the latent dimension. Sampling settings appear in Appendix C.5.

C.3 Model Architecture

BrainViT combines channel attention with a convolutional readout: attention models interactions among the observed sensors, and the readout aggregates multiscale features on a common sensor grid. Learned sensor IDs and montage RoPE encode sensor identity and geometry; nearest-neighbor sensor mapping supports different input montages without learned query compression. Table A.1 summarizes the EEG encoder stages, output shapes, and parameter counts for a 63-sensor reference grid, $T = 250$ time samples, ten subject embeddings, and a 1,024-dimensional output. The token count is $N_{\text{tok}} = C + 1$ with a subject token and C without one. MEG uses the same architecture with a 271-sensor reference grid. Subject tokens are enabled only for within-subject models.

Table A.1: **BrainViT architecture for THINGS-EEG2.** Feature convolutions act on Transformer outputs.

Stage	Component	Output shape	Params
1	DCT-initialized projection + sensor/subject IDs	$N_{\text{tok}} \times 250$	81,500
2	L1/H4 channel Transformer + montage RoPE2D	$N_{\text{tok}} \times 250$	379,000
3	Parameter-free sensor mapping	63×250	0
4	Multiscale feature convolution + pooling	$64 \times 63 \times 36$	1,664
5	Local–global sensor conv. + pointwise projection	$40 \times 1 \times 36$	179,512
6	Flatten + residual projection + LayerNorm	1,024	2,527,232
7	Learnable logit scale	scalar	1
	Total BrainViT parameters		**3,168,909**

BrainViT architecture search. We treat the encoder readout as a design choice and examine how strongly it affects retrieval through a search over five predefined candidates. Following a common channel Transformer, the readouts use a flattened MLP, multiscale feature convolution with MLP,

16-query attention pooling, sensor convolution, or combined feature–sensor convolution. Each encoder and its visual fusion are jointly trained under the same Stage 1 settings and selected by validation loss; retrieval directly compares their embeddings without Stage 2 or CSLS. Evaluation uses ten subjects or LOSO folds, 63 sensors, 80-trial averages, and 200 candidates. LOSO excludes the target subject from training and selection and disables subject tokens. Mean Top-1 accuracy spans 16.30 percentage points within subjects and 4.95 points under LOSO (Table A.2), indicating greater sensitivity to readout design in within-subject retrieval. Sensor convolution yields the highest mean Top-1 in both settings; adding multiscale feature convolution provides no further gain. These comparisons characterize the sensitivity of retrieval to encoder design; unequal parameter counts leave architecture and capacity coupled.

Table A.2: **BrainViT architecture search on THINGS-EEG2.** Five readout designs under a common training protocol. Raw retrieval accuracy (%), subject mean ± SD; encoder parameters in millions.

		Within-subject (Raw)		LOSO (Raw)	
Encoder readout	Params	Top-1	Top-5	Top-1	Top-5
Flatten + MLP	17.641	66.80 ± 7.28	92.30 ± 4.36	$\underline{23.35 \pm 8.32}$	$\mathbf{53.50 \pm 11.58}$
Multiscale feature conv. + flatten + MLP	150.150	70.30 ± 10.78	92.75 ± 5.56	19.60 ± 4.85	47.35 ± 8.17
Learned-query attention pooling ($Q = 16$)	6.117	79.10 ± 8.31	96.15 ± 2.39	23.25 ± 7.36	51.25 ± 11.14
Local–global sensor convolution	3.167	$\mathbf{83.10 \pm 5.37}$	$\mathbf{97.35 \pm 2.19}$	$\mathbf{24.55 \pm 7.88}$	$\underline{52.30 \pm 10.49}$
Multiscale feature + sensor conv.	3.169	$\underline{80.90 \pm 7.03}$	$\underline{96.25 \pm 2.67}$	22.50 ± 7.44	51.60 ± 8.60

Teacher-specific experts and routing. The retrieval adapter uses two residual MLP blocks with bottleneck width 512; the shared readout uses width 256. Each teacher group has one always-active expert and four densely routed experts, each implemented as an MLP with bottleneck width 64 and dropout 0.1. They receive the same adapted brain representation and predict the teacher's native feature space (1,024 dimensions for OpenCLIP, DINOv2, and MAE; 768 for DFN). The base projection is the identity except for DFN's learned linear map. Internal routers share a width-256 stem and use separate teacher-specific heads, with weights defined in Eq. (20). Visual adapters use residual MLPs with bottleneck width 512. Tables A.3 and A.5 summarize parameter counts and optimization settings.

Outer teacher weights and residual gate. The outer router combines scores across teacher spaces, whereas the internal routers above combine residual predictions within each space. Both outer weights and the gate depend only on the adapted brain query $\mathbf{h}'_i \in \mathbb{R}^{1024}$:

$$\mathbf{a}_i = W_\pi \,\mathrm{LN}_\pi(\mathbf{h}'_i + \mathbf{t}_0) + \mathbf{b}_\pi, \qquad \pi_{ik} = \frac{\exp(a_{ik})}{\sum_{m=1}^{K} \exp(a_{im})}, \tag{24}$$

$$g_i = \mathrm{sigmoid}(\mathbf{w}_g^\top \mathbf{h}'_i + b_g). \tag{25}$$

Here $K = 4$, $W_\pi \in \mathbb{R}^{4\times 1024}$, $\mathbf{b}_\pi \in \mathbb{R}^4$, and $\mathbf{w}_g, \mathbf{t}_0 \in \mathbb{R}^{1024}$. The learned embedding $\mathbf{t}_0$ is shared across retrieval queries. Both weights depend only on the brain query and are trained through $\mathcal{L}_{\mathrm{ret}}$; the backbone and fused visual target remain frozen. LOSO bypasses these modules.

Retrieval readout ablation. To test the benefit of composing shared and teacher-specific predictions, we independently train five readouts on the same frozen BrainViT and visual fusion (Table A.4). Linear and MLP controls with bottleneck width 256 predict all five targets directly from BrainViT without adapter A, retaining BHA's score composition. Shared-only and teacher-only controls retain A; the latter combines native scores with outer weights but omits the shared score and gate. Applicable visual adapters and Stage 2 training settings are retained, with selection by validation Raw Top-1 and Top-5 as the tie-breaker. Full BHA achieves the highest retrieval accuracy, while linear heads remain competitive after CSLS. Neither shared-only nor teacher-only prediction matches their composition.

C.4 Retrieval Implementation

EEG visual features are extracted with deterministic OpenCLIP preprocessing and frozen VFMs in evaluation mode.

Table A.3: **Retrieval architecture and parameter counts.** Counts are for the THINGS-EEG2 within-subject model. Frozen source VFMs and the image generator are excluded. Stage 2 freezes BrainViT and visual fusion and trains the remaining modules.

Component	Trainable in Stage 2	Frozen in Stage 2
BrainViT	0	3,168,909
Visual fusion	0	6,035,456
Retrieval adapter A	2,101,250	0
Shared readout E_0	528,385	0
Group MoEs: $4 \times (1 + 4)$ rank-64 experts	2,580,500	0
Group MoEs: internal routing	268,560	0
Group MoEs: DFN base projection	786,432	0
Four visual adapters	3,939,844	0
Outer teacher router and scalar gate	9,221	0
Retrieval logit scales	6	0
Total	10,214,198	9,204,365

Table A.4: **Retrieval readout ablation on THINGS-EEG2.** Within-subject accuracy (%), averaged over ten subjects. Light gray marks BHA rows.

	Raw		CSLS-2	
Architecture	Top-1	Top-5	Top-1	Top-5
BrainViT + linear heads	<u>85.00</u>	<u>98.30</u>	<u>93.45</u>	<u>99.60</u>
BrainViT + MLP heads	83.20	97.90	92.15	99.45
BrainViT + shared expert only	82.05	96.95	91.40	99.15
BrainViT + gated teacher experts only	79.60	97.05	87.70	98.55
BrainViT + teacher MoEs + shared expert (BHA)	**87.00**	**98.45**	**93.90**	**99.70**

Model selection and evaluation. Stage 1 selects the minimum validation loss; Stage 2 selects the maximum validation gated Raw Top-1, with Top-5 as the tie-breaker. Both use the image-disjoint split in Section D.1.1. We evaluate 200-way retrieval and average accuracies across ten EEG or four MEG subjects, using 80- and 12-trial test averages, respectively. Within-subject retrieval uses Stage 2 gated scores. EEG and MEG LOSO train on the remaining subjects and directly use $S_{ij}^{\mathrm{LOSO}} = \mathrm{sim}(\mathbf{h}_i, \mathbf{e}_{j0})$, bypassing all Stage 2 modules. Raw scores queries independently; CSLS rescores the same predictions using other test queries. The extended EEG LOSO table also reports image-to-brain retrieval.

Readout choice under LOSO. LOSO requires direct generalization to an unseen subject without target-subject fitting. Query-dependent routing learned on source subjects may encode their subject-specific feature distributions and expert preferences, introducing additional source-subject bias at the readout. We therefore use Stage 1 brain predictions and fused visual targets directly for LOSO, bypassing the specialized Stage 2 readouts and gating. Within-subject evaluation instead uses Stage 2 to learn matching preferences from the same subject's training data. This rationale concerns potential transfer bias; the stage comparison in Appendix D.1.3 does not isolate gating as its cause.

CSLS and reference information. For neural queries indexed by a reference set $\mathcal{R}$ and an image gallery $\mathcal{G}$, CSLS (Conneau et al., 2017) adjusts the raw score matrix $\mathbf{Q}$, where $Q_{ij} = S_{ij}$ for within-subject retrieval and $Q_{ij} = S_{ij}^{\mathrm{LOSO}}$ for LOSO. For a neighborhood size $k \geq 1$, we use $k_{\mathrm{eff}} = \min(k, |\mathcal{R}|, |\mathcal{G}|)$. Let $\mathcal{N}_{k_{\mathrm{eff}}}^{\mathcal{G}}(i)$ index the k_{eff} highest-scoring images for query i, and $\mathcal{N}_{k_{\mathrm{eff}}}^{\mathcal{R}}(j)$ index the k_{eff} highest-scoring neural references for image j. Then

$$r_i(k) = \frac{1}{k_{\mathrm{eff}}} \sum_{j' \in \mathcal{N}_{k_{\mathrm{eff}}}^{\mathcal{G}}(i)} Q_{ij'}, \qquad c_j(k) = \frac{1}{k_{\mathrm{eff}}} \sum_{i' \in \mathcal{N}_{k_{\mathrm{eff}}}^{\mathcal{R}}(j)} Q_{i'j},$$
$$\widetilde{Q}_{ij}(k) = 2Q_{ij} - r_i(k) - c_j(k). \tag{26}$$

Neighborhoods use uncorrected scores. Main CSLS results use all 200 test queries and 200 gallery images, so $k_{\mathrm{eff}} = k$. The reference-size ablation truncates both neighborhoods when fewer queries

are available. For neural-to-image retrieval, $r_i(k)$ is constant across candidates; for the reverse direction, $c_j(k)$ is constant. CSLS therefore changes rankings through the opposite-side correction. It uses no labels or parameter updates, but access to other test queries makes it transductive.

Visual targets. Our four retrieval teachers are OpenCLIP ViT-H/14 global features (Cherti et al., 2023), MAE ViT-L (He et al., 2022), DINOv2-L (Oquab et al., 2024), and DFN ViT-B (Fang et al., 2024). The latter three use intermediate features MAE-18, DINOv2-12, and DFN-6, whose suffixes denote zero-based Transformer block indices. DFN-9 is used only for the layer comparison; the main fusion model uses DFN-6. OpenCLIP, MAE-18, and DINOv2-12 provide 1,024-dimensional features; DFN-6 provides 768-dimensional features. Their projections, normalization, and fusion follow Eq. (5), producing $\mathbf{e}_{i0} \in \mathbb{R}^{d_0}$ with $d_0 = 1024$. The individual teacher targets are adapted only during distillation. Reconstruction uses the native OpenCLIP and VAE spaces described in Appendix C.5.

C.5 Reconstruction Implementation

Reconstruction uses the same BrainViT encoder and convolutional readout architecture as primary retrieval. Its base weights and all generator components remain frozen; only the reconstruction predictors are trained.

Semantic and spatial prediction. The semantic branch uses rank-64 LoRA and a two-block residual calibrator with bottleneck width 256 to predict a 1,024-dimensional OpenCLIP ViT-H/14 image embedding. The frozen IP-Adapter (Ye et al., 2023) maps it to four 2,048-dimensional conditioning tokens, following the CLIP-conditioning approach of ENIGMA (Kneeland et al., 2026).

The spatial branch uses the 17 posterior EEG sensor tokens to predict an unscaled $4 \times 16 \times 16$ VAE latent with a convolutional readout. A residual convolutional adapter maps this prediction to the pyramid's coarse target. Cross-attention from 256 learned spatial queries to the EEG tokens supplies a 64-channel conditioning map. Two conditioned residual stages then refine the latent to $4 \times 32 \times 32$ and $4 \times 64 \times 64$, following Eq. (23). The coarse predictor targets the VAE posterior mean of a 128×128 image; pyramid targets are downsampled from its 512×512 encoding.

Training. Semantic losses weight global-feature MSE and BiCLIP by 0.9 and 0.1, and IP-token MSE and cosine distance by 0.1 and 0.01; BiCLIP temperature is 0.07. Coarse MSE and cosine weights are 1 and 0.1. Pyramid MSE weights are $(0.5, 0.5, 1)$ for spatial sizes $(16, 32, 64)$, with 0.2 for each unscaled residual and 0.05 for finest-level cosine distance. Pyramid MSE errors are divided by target-channel training standard deviations. We first train the pyramid with the coarse predictor frozen, then jointly refine the spatial path with an additional unstandardized coarse MSE term (weight 0.5). AdamW uses learning rates 3×10^{-4} for the pyramid and 10^{-5} for the coarse predictor, with weight decay 10^{-4}. Training and validation use four-trial EEG averages.

Generation. We generate 512×512 images with frozen SDXL-Turbo (Sauer et al., 2024) and a ViT-H IP-Adapter, using Euler ancestral sampling, IP-Adapter scale 1.5, and classifier-free guidance scale 0. The main setting is $\alpha = 0.75$: three denoising updates from a five-step schedule. Semantic embeddings enter the IP-Adapter without normalization; spatial latents use VAE scale $a = 0.13025$ before noise addition. Every stimulus uses the same generic quality prompt, without image-specific text. Results average three generations per image. The guided variant uses the same models and sampling settings with the additional updates below.

Training-free guidance. For noisy latent $\mathbf{z}_t$, Euler noise level σ_t, and denoiser condition $\mathbf{c}_{\text{sem}}$ (fixed text conditioning omitted), let $\tilde{\mathbf{z}}_t = \mathbf{z}_t / \sqrt{1 + \sigma_t^2}$. The estimated clean latent and its consistency loss are

$$\hat{\mathbf{z}}_{0|t} = \mathbf{z}_t - \sigma_t \epsilon_{\text{den}}(\tilde{\mathbf{z}}_t, t, \mathbf{c}_{\text{sem}}), \qquad \mathcal{L}_{\text{guide}}(\mathbf{z}_t) = \frac{1}{d_z} \left\| \frac{\hat{\mathbf{z}}_{0|t}}{a} - \hat{\mathbf{z}} \right\|_2^2. \tag{27}$$

Here ϵ_{den} is the frozen denoiser, $a = 0.13025$ is the VAE scaling factor, and $d_z = 4 \cdot 64 \cdot 64$; latent tensors are vectorized in the loss and gradient normalization. The reference $\hat{\mathbf{z}}$ is the unscaled spatial

prediction. An RMS-normalized gradient step refines the noisy latent:

$$\mathbf{g}_t = \nabla_{\mathbf{z}_t} \mathcal{L}_{\text{guide}}(\mathbf{z}_t), \qquad \mathrm{RMS}(\mathbf{g}_t) = \sqrt{\frac{1}{d_z}\sum_{j=1}^{d_z} (\mathbf{g}_t)_j^2},$$
$$\mathbf{z}_t^+ = \mathbf{z}_t - \eta_{\text{g}} \sigma_t \frac{\mathbf{g}_t}{\max\{\mathrm{RMS}(\mathbf{g}_t), \varepsilon\}}. \tag{28}$$

We use $\eta_{\text{g}} = 0.02$ and $\varepsilon = 10^{-12}$, applying one refinement before each of the final two denoising updates, followed by a new noise prediction and an Euler ancestral step. Model weights and predicted conditions remain unchanged. The remaining generation settings are those above.

Image-conditioned and permuted-condition controls are described in Appendix D.3.1.

D Additional Experimental Results

D.1 Comprehensive BHA Results

We examine how retrieval varies across subjects and evaluation settings. The subject-wise tables report individual accuracies and aggregate scores (Avg); "–" denotes unreported results. Baseline aggregates are reproduced from their reporting sources, while BHA averages subjects equally. Across Raw and CSLS, BHA uses identical model outputs and trial averages. Raw scores queries independently, whereas CSLS-2/10 use all 200 test queries as references (Appendix C.4).

Baseline aggregation discrepancies. A dagger marks published aggregates that differ from the arithmetic means of the displayed subject scores beyond rounding. For within-subject EEG, the ATM-S row reproduced from MBP's comparison reports $28.5/60.4$, whereas its subject entries average $27.10/58.08$; MBP reports Top-5 94.6 versus 94.91 from its entries (Zheng et al., 2026). For MEG LOSO, HyFI reports $3.2/11.5$ versus $2.875/10.775$ from its entries (Jo et al., 2026). We retain these published aggregates in the main and supplementary comparisons; without the underlying runs, we cannot resolve the source inconsistencies. CC-All subject values follow its original report (Zhang et al., 2025).

D.1.1 Retrieval Performance in THINGS-EEG2

Training settings. For within-subject retrieval, we fit the two-stage model separately for each EEG participant. We reserve one image per concept (1,654 images) for validation and use 14,886 images for fitting, keeping all four repeats of each image together. Both stages use this split. Table A.5 lists the key training and inference hyperparameters.

Table A.5: **THINGS-EEG2 retrieval hyperparameters.** BS: batch size; GAS: gradient accumulation steps.

Hyperparameter	**Stage 1**	**Stage 2**
Optimizer	AdamW; $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$	
BS / GAS	1024 / 1	
Learning rate / schedule	3×10^{-4}; constant, no warmup	
Weight decay	10^{-2}	10^{-4}
Maximum epochs / patience	100 / 10	160 / 20
Channel sampling	63 or 17; $p = 0.5$ each	–
Expert dropout	–	0.1
Router uniform-mixing ratio	–	0.02
Internal routing schedule	–	Uniform in epoch 1; temperature $1.5 \to 0.8$ over 40 epochs
Native / shared / gated loss weights	–	0.25 (teacher mean) / 1 / 1
Visual drift / router balance weights	–	0.1 / 0.01
CSLS: neighbors / reference queries	$k \in \{2, 10\}$ / $M = 200$	

Within-subject retrieval. Within-subject retrieval uses the Stage-2 gated readout. The subject-wise comparison shows strong within-subject decoding, with CSLS further improving retrieval from the same predictions (Table A.6).

Table A.6: **Within-subject retrieval on THINGS-EEG2.** 200-way Top-1 and Top-5 accuracy (%). Light gray marks BHA rows.

Method	Sub1		Sub2		Sub3		Sub4		Sub5		Sub6		Sub7		Sub8		Sub9		Sub10		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
BraVL (Du et al., 2023)	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE (Song et al., 2024)	13.2	39.5	13.5	40.3	14.5	42.7	20.6	52.7	10.1	31.5	16.5	44.0	17.0	42.1	22.9	56.1	15.4	41.6	17.4	45.8	16.1	43.6
ATM-S (Li et al., 2024)	25.6	60.4	22.0	54.5	25.0	62.4	31.4	60.9	12.9	43.0	21.3	51.1	30.5	61.5	38.8	72.0	34.4	51.5	29.1	63.5	28.5†	60.4†
CC-All (Zhang et al., 2025)	31.4	79.7	31.4	77.8	38.2	85.7	40.4	85.8	24.4	66.3	34.8	78.8	34.7	81.0	48.1	88.6	37.4	79.4	35.6	79.3	35.6	80.2
UBP (Wu et al., 2025)	41.2	70.5	51.2	80.9	51.2	82.0	51.1	76.9	42.2	72.8	57.5	83.5	49.0	79.9	58.6	85.8	45.1	76.2	61.5	88.2	50.9	79.7
NeuroBridge (Zhang et al., 2026a)	50.0	77.6	63.2	90.6	61.6	91.1	61.4	90.0	54.8	85.0	69.7	92.9	62.7	88.8	71.2	95.1	64.0	91.0	73.6	97.1	63.2	89.9
MBP (Zheng et al., 2026)	64.3	88.8	76.3	95.3	74.0	95.0	67.0	91.8	68.0	91.5	81.5	96.3	76.8	96.8	84.8	98.5	76.8	95.8	87.3	99.3	75.7	94.6†
DMSA (Du et al., 2026)	75.0	94.3	87.5	98.9	83.2	98.2	79.5	96.1	74.6	96.4	89.9	99.3	78.5	97.3	86.9	99.4	81.3	97.8	89.3	99.1	82.6	97.7
BHA (Ours) / Raw	90.0	99.5	85.5	97.5	88.0	99.5	83.5	97.5	77.0	96.5	92.5	98.5	88.5	98.5	91.0	99.5	85.0	98.0	89.0	99.5	87.00	98.45
BHA (Ours) / CSLS-2	**98.0**	**100.0**	**92.5**	**100.0**	**95.0**	**100.0**	**91.0**	**99.5**	**85.0**	**99.5**	**96.5**	**99.5**	**96.5**	**99.5**	**96.5**	**100.0**	**91.0**	**99.0**	**97.0**	**100.0**	**93.90**	**99.70**
BHA (Ours) / CSLS-10	93.0	**100.0**	90.5	99.5	93.0	99.5	88.5	98.5	82.5	99.0	94.0	99.0	94.5	99.0	95.0	**100.0**	87.5	**99.0**	95.0	99.5	91.35	99.30

LOSO retrieval comparison. To test whether shared visual supervision supports transfer without target-subject fitting, we train on nine subjects and evaluate the held-out subject in each of ten folds, excluding target-subject data from training and model selection. Retrieval directly matches Stage-1 BrainViT to the frozen fused visual target: $S_{ij}^{\mathrm{LOSO}} = \mathrm{sim}(\mathbf{h}_i, \mathbf{e}_{j0})$. Subject tokens are disabled. The shared representation supports transfer to unseen subjects, although accuracy remains below within-subject retrieval. Table A.7 compares methods; Table A.8 reports both retrieval directions at Top-1 and Top-5. ATM-S subject-wise values follow the comparison reported by NeuroBridge (Zhang et al., 2026a); their means ($11.9/33.8$) are also used in Table 1.

Table A.7: **LOSO retrieval on THINGS-EEG2.** 200-way Top-1 and Top-5 accuracy (%). Light gray marks BHA rows.

Method	Sub1		Sub2		Sub3		Sub4		Sub5		Sub6		Sub7		Sub8		Sub9		Sub10		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
NICE (Song et al., 2024)	7.6	22.8	5.9	20.5	6.0	22.3	6.3	20.7	4.4	18.3	5.6	22.2	5.6	19.7	6.3	22.0	5.7	17.6	8.4	28.3	6.2	21.4
ATM-S (Li et al., 2024)	10.5	26.8	7.1	24.8	11.9	33.8	14.7	39.4	7.0	23.9	11.1	35.8	16.1	43.5	15.0	40.3	4.9	22.7	20.5	46.5	11.9	33.8
Neural-MCRL (Li et al., 2025b)	13.0	31.5	12.0	30.5	14.5	35.5	12.5	35.5	11.5	29.0	13.5	35.5	14.0	36.0	18.5	38.5	13.5	32.5	17.0	39.0	14.0	34.3
UBP (Wu et al., 2025)	11.5	29.7	15.5	40.0	9.8	27.0	13.0	32.3	8.8	33.8	11.7	31.0	10.2	23.8	12.2	32.2	15.5	40.5	16.0	43.5	12.4	33.4
NeuroBridge (Zhang et al., 2026a)	23.2	52.4	21.2	49.3	13.2	36.5	17.0	45.3	14.5	37.7	25.0	55.0	15.3	45.1	20.1	44.9	13.7	36.5	27.2	56.3	19.0	45.9
DMSA (Du et al., 2026)	23.5	53.2	30.6	60.4	10.0	28.1	19.5	48.3	18.1	45.2	22.7	49.8	18.6	46.0	17.3	46.1	23.0	54.8	34.4	62.0	21.8	49.4
BHA (Ours) / Raw	35.0	64.5	29.0	57.5	12.5	38.0	21.0	48.5	14.0	44.5	22.5	51.0	21.0	52.5	17.0	42.5	21.0	54.0	32.0	63.0	22.50	51.60
BHA (Ours) / CSLS-2	43.5	**80.0**	**40.0**	66.0	22.0	48.5	24.0	57.0	20.5	50.5	33.0	62.0	28.0	**66.5**	21.5	**57.5**	27.5	64.0	**43.5**	**81.5**	30.35	63.35
BHA (Ours) / CSLS-10	**47.5**	77.5	39.5	**68.0**	**23.5**	**49.5**	**25.0**	**58.5**	**21.5**	**52.0**	**34.5**	**64.0**	**30.5**	66.0	**22.0**	56.0	**29.0**	**65.5**	**43.5**	77.0	**31.65**	**63.40**

Table A.8: **Bidirectional LOSO retrieval on THINGS-EEG2.** 200-way accuracy (%); columns 01–10 denote subjects. All rows report BHA.

Metric	Scoring	01	02	03	04	05	06	07	08	09	10	Avg
EEG→Image												
	Raw	35.0	29.0	12.5	21.0	14.0	22.5	21.0	17.0	21.0	32.0	22.50
Top-1	CSLS-2	43.5	40.0	22.0	24.0	20.5	33.0	28.0	21.5	27.5	43.5	30.35
	CSLS-10	47.5	39.5	23.5	25.0	21.5	34.5	30.5	22.0	29.0	43.5	**31.65**
	Raw	64.5	57.5	38.0	48.5	44.5	51.0	52.5	42.5	54.0	63.0	51.60
Top-5	CSLS-2	80.0	66.0	48.5	57.0	50.5	62.0	66.5	57.5	64.0	81.5	63.35
	CSLS-10	77.5	68.0	49.5	58.5	52.0	64.0	66.0	56.0	65.5	77.0	**63.40**
Image→EEG												
	Raw	40.0	36.0	25.0	30.0	25.5	38.0	33.5	21.5	31.5	49.0	33.00
Top-1	CSLS-2	46.5	38.0	25.5	30.5	26.5	39.0	36.0	26.0	34.5	53.0	35.55
	CSLS-10	47.0	40.0	27.5	32.5	28.0	40.5	39.0	26.0	36.5	53.5	**37.05**
	Raw	76.5	67.0	58.5	57.0	49.5	70.5	65.5	59.0	63.5	77.0	64.40
Top-5	CSLS-2	79.5	72.0	60.0	59.0	52.0	71.5	68.0	60.0	67.5	81.0	67.05
	CSLS-10	80.0	71.5	62.5	60.5	55.5	73.5	69.0	59.0	69.5	81.0	**68.20**

D.1.2 Retrieval Performance in THINGS-MEG

Within-subject and LOSO retrieval comparison. We evaluate four within-subject and four LOSO MEG models, using the readouts in Appendix C.4. Batch sizes are 256 for pretraining and 1,024 for readout training. MEG pretraining equally weights paired and same-concept BiCLIP, with a four-example concept-average loss weighted by 0.25. LOSO disables subject tokens and uses

source-training normalization statistics without target-subject fitting. CSLS improves retrieval in both settings, but the substantial transfer gap persists across modalities (Tables A.9 and A.10). MBP values follow its published subject-wise table ($33.9/60.7$), also used in Table 1; the source paper's summary table instead reports $33.7/60.5$ (Zheng et al., 2026).

Table A.9: **Within-subject retrieval on THINGS-MEG.** 200-way Top-1 and Top-5 accuracy (%). Light gray marks BHA rows.

Method	Sub1		Sub2		Sub3		Sub4		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
NICE (Song et al., 2024)	9.6	27.8	18.5	47.8	14.2	41.6	9.0	26.6	12.8	36.0
NICE-SA (Song et al., 2024)	9.8	27.8	18.6	46.4	10.5	38.4	11.7	27.2	12.7	35.0
NICE-GA (Song et al., 2024)	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.3	42.3
UBP (Wu et al., 2025)	15.0	38.0	46.0	80.5	27.3	59.0	18.5	43.5	26.7	55.2
MBP (Zheng et al., 2026)	14.0	31.8	63.8	91.8	41.0	78.3	17.0	41.0	33.9	60.7
HyFI (Jo et al., 2026)	17.6	40.1	63.8	91.1	38.0	76.9	23.7	50.2	35.8	64.6
NeuroBridge (Zhang et al., 2026a)	16.5	41.6	53.7	85.3	40.4	73.2	18.1	43.1	32.2	60.8
LVBP (Liu et al., 2026)	25.4	–	78.1	–	37.3	–	35.3	–	44.0	72.0
DMSA (Du et al., 2026)	25.5	54.5	81.9	97.4	56.0	87.5	28.6	58.3	48.0	74.4
BHA (Ours) / Raw	38.0	74.5	78.0	99.5	63.0	89.0	43.5	80.0	55.63	85.75
BHA (Ours) / CSLS-2	**53.0**	**86.0**	**92.5**	**100.0**	**71.5**	**94.5**	**55.5**	**85.0**	**68.13**	**91.38**
BHA (Ours) / CSLS-10	51.0	83.5	89.0	**100.0**	68.5	93.5	52.0	84.0	65.13	90.25

Table A.10: **LOSO retrieval on THINGS-MEG.** 200-way Top-1 and Top-5 accuracy (%). Light gray marks BHA rows.

Method	Sub1		Sub2		Sub3		Sub4		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
UBP (Wu et al., 2025)	2.0	5.7	1.5	17.2	2.7	10.5	2.5	8.0	2.2	10.4
HyFI (Jo et al., 2026)	2.7	6.0	4.3	17.0	3.5	12.5	1.0	7.6	3.2†	11.5†
NeuroBridge (Zhang et al., 2026a)	**4.3**	**13.1**	3.6	15.6	3.0	11.2	2.5	11.3	3.4	12.8
LVBP (Liu et al., 2026)	2.9	–	7.7	–	5.8	–	**4.7**	–	5.3	15.9
DMSA (Du et al., 2026)	1.3	6.9	6.6	18.5	5.4	18.5	1.5	7.9	3.7	13.0
BHA (Ours) / Raw	0.5	5.0	7.5	18.5	5.0	19.5	1.5	8.0	3.63	12.75
BHA (Ours) / CSLS-2	1.0	6.5	10.0	24.5	**7.5**	21.0	2.5	**14.0**	5.25	16.50
BHA (Ours) / CSLS-10	0.5	6.5	**11.5**	**26.0**	**7.5**	**22.5**	2.5	**14.0**	**5.50**	**17.25**

D.1.3 Training-Stage and CSLS Ablations

We examine two sources of retrieval gains: learning task-specific readouts and correcting candidate scores at evaluation. Separate comparisons assess readout performance, training-data requirements, and dependence on the CSLS reference cohort.

Two-stage training and retrieval readouts. To determine whether readout specialization improves both discrimination and transfer, we compare Stage-1 BrainViT with the Stage-2 shared and gated outputs using Top-1 and Top-5 accuracy (Table A.11). Stage 1 learns BrainViT and visual fusion; Stage 2 jointly trains the shared and teacher-specific readouts around the frozen backbone. Both stages use the same 14,886 training images and 1,654 validation images per subject. The shared and gated readouts are evaluated from the same Stage-2 model. All rows use 63 channels, 80-repeat queries, and 200 candidates; results are averaged over ten subjects or LOSO folds. Gating improves both metrics within subjects, whereas pretrained BrainViT generally performs better under LOSO. At full training data, LOSO Raw Top-1 is 22.50% for pretrained BrainViT and 20.20% for the gated readout. This pattern is consistent with the readout rationale in Appendix C.4 and motivates evaluating specialization and cross-subject transfer separately.

Training-data scale. We train separate backbones on nested 25%, 50%, 75%, and 100% subsets of the 14,886 EEG training images, keeping validation data and Stage 2 settings unchanged. Shared and gated scores come from the same trained readouts; the full-data model supplies the main results. Gated retrieval outperforms both pretrained BrainViT and its shared readout at every fraction (Figure 5), showing that the advantage persists with less training data. This comparison does not isolate routing from added capacity.

Table A.11: **Two-stage retrieval comparison on THINGS-EEG2.** Mean Top-1 / Top-5 accuracy (%).

	Within-subject			LOSO		
Stage / readout	Raw	CSLS-2	CSLS-10	Raw	CSLS-2	CSLS-10
Stage 1 / pretrained	80.80 / 96.25	90.95 / 99.05	87.85 / 98.55	**22.50 / 51.60**	30.35 / **63.35**	**31.65 / 63.40**
Stage 2 / shared expert	82.20 / 96.80	91.95 / 99.05	88.65 / 98.60	22.05 / 50.15	**30.50** / 61.15	31.55 / 61.40
Stage 2 / gated	**87.00 / 98.45**	**93.90 / 99.70**	**91.35 / 99.30**	20.20 / 48.35	28.15 / 60.25	29.90 / 60.50

Visual supervision and adaptation. Single-VFM and BHA retrieval comparisons show performance differences across visual spaces (Figure 1), while visual adapters improve paired brain–image alignment (Figure 6(a)). These comparisons motivate adapting visual supervision, but do not isolate fusion from the other components of BHA. The matched pretraining-source control in Appendix D.3.2 separately examines how visual supervision affects reconstruction, using identical reconstruction architectures and training schedules.

CSLS neighborhood size. We test how retrieval depends on the neighborhood size used to correct candidate scores. Each k uses the same raw score matrix, readouts, trial-averaged queries, and 200-image gallery; all 200 queries supply the reference statistics. Results average ten EEG or four MEG subjects. Smaller neighborhoods favor within-subject retrieval, whereas LOSO benefits from larger neighborhoods in several metrics (Table A.12). The preferred correction thus depends on the evaluation setting. This is a post-hoc sensitivity analysis: $k = 2$ was assessed after benchmark inspection, following the initial $k = 10$ setting.

Table A.12: **CSLS neighborhood-size ablation.** Mean 200-way Top-1 / Top-5 accuracy (%).

Method	Setting	Raw	CSLS			
			$k = 2$	$k = 4$	$k = 10$	$k = 20$
THINGS-EEG2						
BHA (Ours)	Within-subject	87.00 / 98.45	**93.90 / 99.70**	92.95 / 99.55	91.35 / 99.30	90.45 / 99.15
BHA (Ours)	LOSO	22.50 / 51.60	30.35 / 63.35	31.40 / 63.45	**31.65** / 63.40	31.50 / **63.55**
THINGS-MEG						
BHA (Ours)	Within-subject	55.63 / 85.75	**68.13 / 91.38**	66.63 / 90.75	65.13 / 90.25	63.25 / 89.25
BHA (Ours)	LOSO	3.63 / 12.75	5.25 / 16.50	5.63 / 16.63	5.50 / 17.25	**5.88 / 17.38**

CSLS reference-set size. We vary the available query cohort from $M = 1$ to 200 while retaining the 200-image gallery (Figures A.1 and A.2). Each subject uses nested subsets from 200 random orderings; the same queries supply CSLS statistics and evaluation, and Raw uses identical subsets. Trial averages and readouts are unchanged, with $k_{\text{eff}} = \min(k, M)$. CSLS preserves Raw rankings at $M = 1$ and recovers the main results at $M = 200$, quantifying its dependence on other test queries.

D.1.4 Temporal Window Ablation

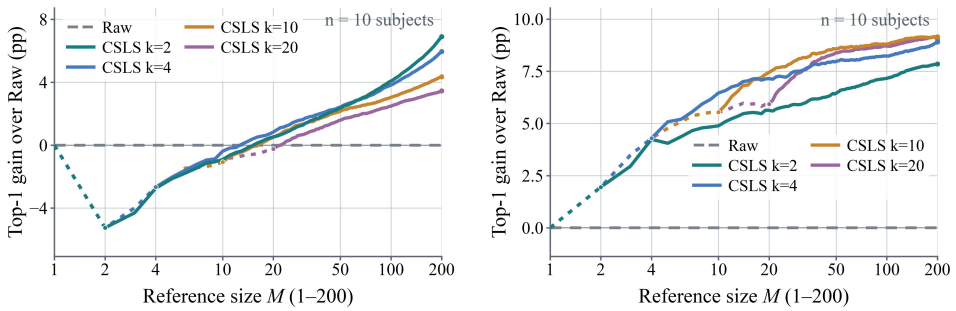
Evaluation details. We measure temporal dependence by deleting or retaining short windows in THINGS-EEG2, THINGS-EEG1, THINGS-MEG, and Alljoined-1.6M (Figure A.3; datasets in Appendix C.1). Each subject's frozen model evaluates its teacher branches, shared expert, and composed output with Raw 200-way retrieval. Trial averages follow the dataset settings: 80 for THINGS-EEG2, 12 for THINGS-EEG1 and MEG, and approximately 80 for Alljoined.

Delete and Keep masks. For $\mathbf{X} \in \mathbb{R}^{C\times T}$, let τ_n be the time of sample n and define $m_w(\tau_n) = \mathbf{1}[a_w \le \tau_n < a_w + 50\,\text{ms}]$. The mask is shared across channels. We compare

$$\mathbf{X}^{\text{Delete},w}_{c,n} = \mathbf{X}_{c,n}\big(1 - m_w(\tau_n)\big), \qquad \mathbf{X}^{\text{Keep},w}_{c,n} = \mathbf{X}_{c,n} m_w(\tau_n). \tag{29}$$

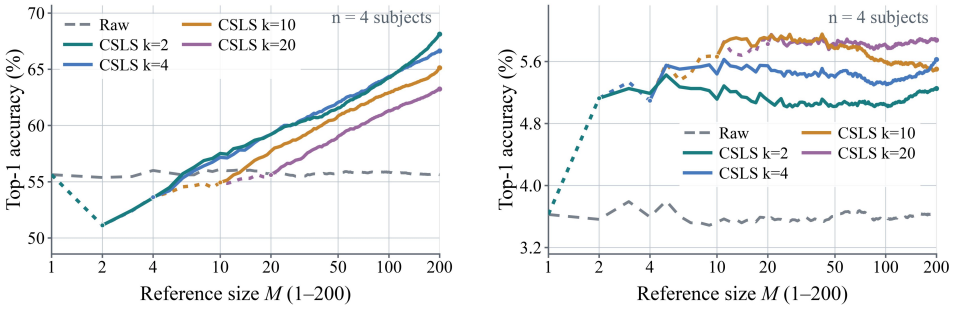
Delete zeros only the selected window; Keep zeros everything outside it. We evaluate each masked input with the frozen model, preserving sequence length and sample times without rescaling or renormalization. We scan 96 windows with starts $a_w = 0, 10, \ldots, 950$ ms; the figure shows window centers up to 795 ms on a 0–800 ms axis.

For subject s, let Acc_s denote Top-1 accuracy as a fraction. The Delete curves in Figure A.3 report the paired decrease $100[\text{Acc}_s(\mathbf{X}) - \text{Acc}_s(\mathbf{X}^{\text{Delete},w})]$ in percentage points; the Keep curves in



(a) EEG within-subject (b) EEG LOSO

Figure A.1: **CSLS reference-size ablation: THINGS-EEG2.** (a) Within-subject Stage-2 gated and (b) LOSO Stage-1 pretrained Top-1. Lines average 200 subset repetitions per subject, then ten subjects; shading denotes between-subject SD. The gallery has 200 images; dotted segments use $k_{\text{eff}} = M < k$.



(a) MEG within-subject (b) MEG LOSO

Figure A.2: **CSLS reference-size ablation: THINGS-MEG.** (a) Within-subject Stage-2 gated and (b) LOSO Stage-1 pretrained Top-1. Sampling and shading follow Figure A.1, averaged over four subjects. Raw and CSLS use the same query subsets and 200-image gallery.

Figure A.3 report $100\,\mathrm{Acc}_s(\mathbf{X}^{\mathrm{Keep},w})$ in percent, with chance at 0.5%. We compute these quantities separately for each subject before reporting their mean and sample standard deviation.

Temporal sensitivity. THINGS-EEG1, THINGS-EEG2, and THINGS-MEG show early sensitivity around 100–125 ms and additional sensitivity later in the response. Later peaks occur near image offset or subsequent stimulus onset, so the masking experiment cannot distinguish these overlapping contributions. Keep and Delete curves capture complementary aspects of model dependence on time; their peaks need not coincide.

Temporal interpretation. Alljoined timestamps contain sample-clock drift and gaps, whereas the preprocessing maps events onto a nominal uniform sample axis. Its plotted latencies are therefore approximate and should not be used for precise cross-dataset timing comparisons. Across datasets, these curves measure the sensitivity of frozen models to temporal masking; overlapping responses and acquisition differences preclude identifying ERP components or causal effects of stimulus duration.

D.2 Extended Representation Analysis

We examine shared geometry, concept associations, and recoverable image content to characterize what the learned embeddings retain beyond retrieval accuracy. The shared expert denotes the outer expert E_0 defined in Section 4.2.

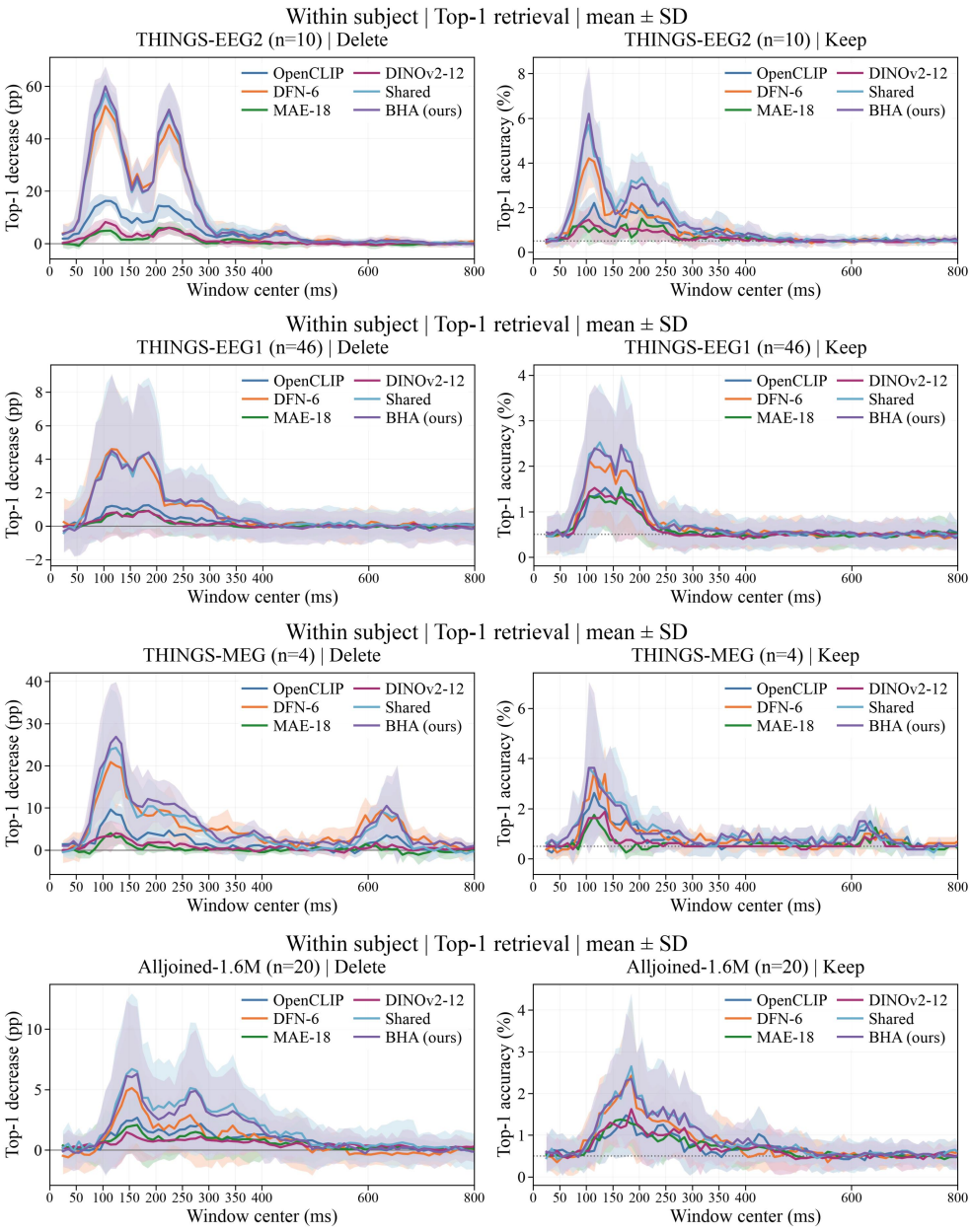


Figure A.3: **Temporal masking of frozen models.** Rows: THINGS-EEG2 ($n = 10$), THINGS-EEG1 ($n = 46$), THINGS-MEG ($n = 4$), and Alljoined ($n = 20$). Left: Full minus Delete Top-1 (percentage points). Right: Keep Top-1 (%). Windows span 50 ms with 10 ms stride; lines/bands show subject mean/± 1 SD. Reference lines mark zero decrease and 0.5% chance. Alljoined timing is provisional (Section D.1.4).

D.2.1 Representational Alignment and Matching

We examine spatial correspondence between brain predictions and image features by comparing OpenCLIP patch-similarity maps for visual and brain queries, alongside a random-query reference (Figure A.4).

To distinguish alignment of individual image pairs from preservation of relations among images, we compare paired brain–image cosine similarity and correlations between representational similarity matrices (Figure 4). For each subject, the matrices use the 19,900 distinct image pairs among 200 test images; displayed correlations are averaged across ten subjects.

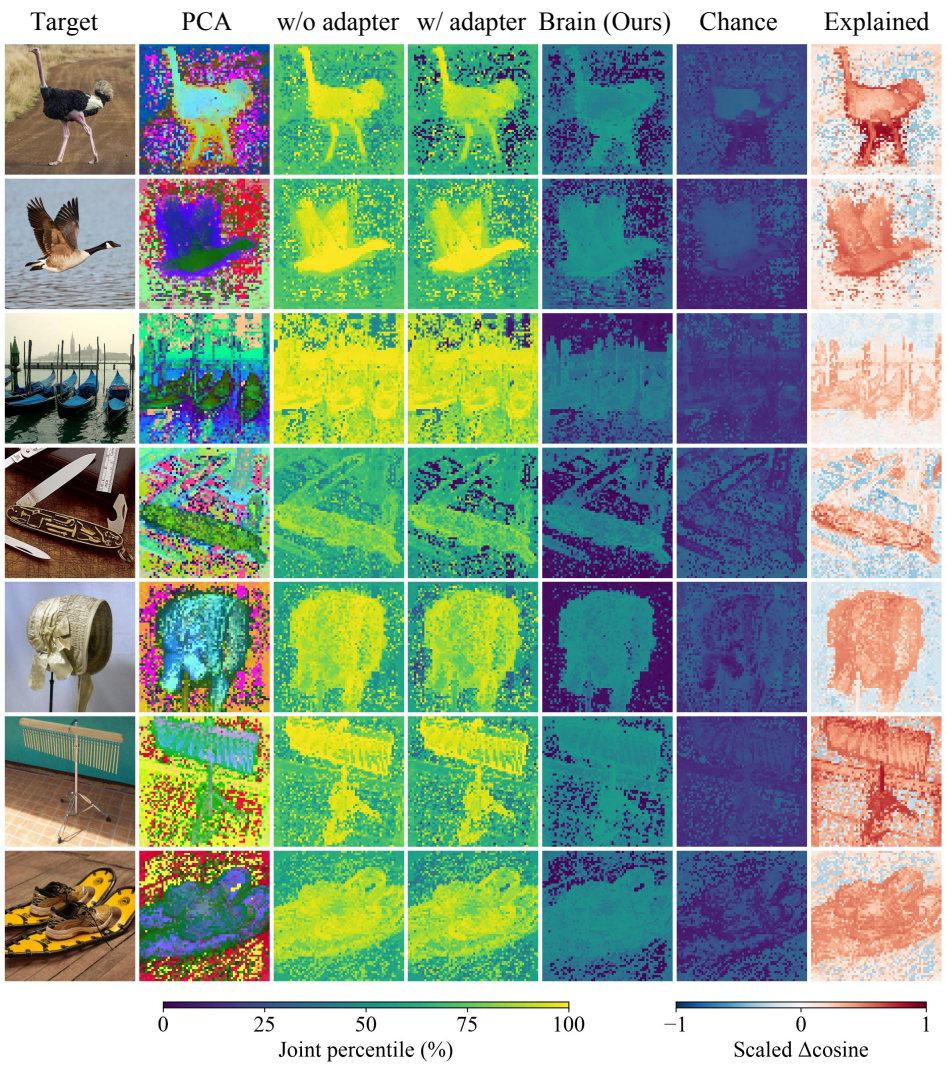


Figure A.4: **Additional OpenCLIP dense-feature query maps.** Supplementing Figure 1, columns show the target image, PCA projections of visual features, global-to-patch cosine similarity without and with the adapter, patch similarity to the matched brain features, a chance-query baseline, and the residual (brain query minus chance). Similarity maps use percentile coloring; the residual preserves the sign of the cosine difference.

The shared expert combines strong paired alignment with close agreement in inter-image geometry. Student outputs also agree more strongly across experts than their visual targets do, consistent with their shared inputs, backbone, and supervision. This comparison does not isolate the contribution of neural information to that agreement. Spatial analyses show a further distinction: experts can respond to similar image regions while assigning opposite signed contributions to a brain–image match.

Spatial attribution. We test whether overlapping visual sensitivity yields similar matching contributions by masking 2×2 regions on a 16×16 patch grid, with stride one patch. Images are resized to 224, 512, 256, and 224 pixels for OpenCLIP, DFN, MAE, and DINOv2, respectively; masks replace pixels with the image-wide mean color. With the brain query unchanged, attribution is the original minus masked-image cosine similarity. Overlapping scores are averaged spatially, then across ten subjects. Colors show signed within-map percentiles, so intensities are not comparable absolute effect sizes. These full-image attributions use different preprocessing from the training crops. Figure A.5 extends the three examples in Figure 4(e) with seven additional images.

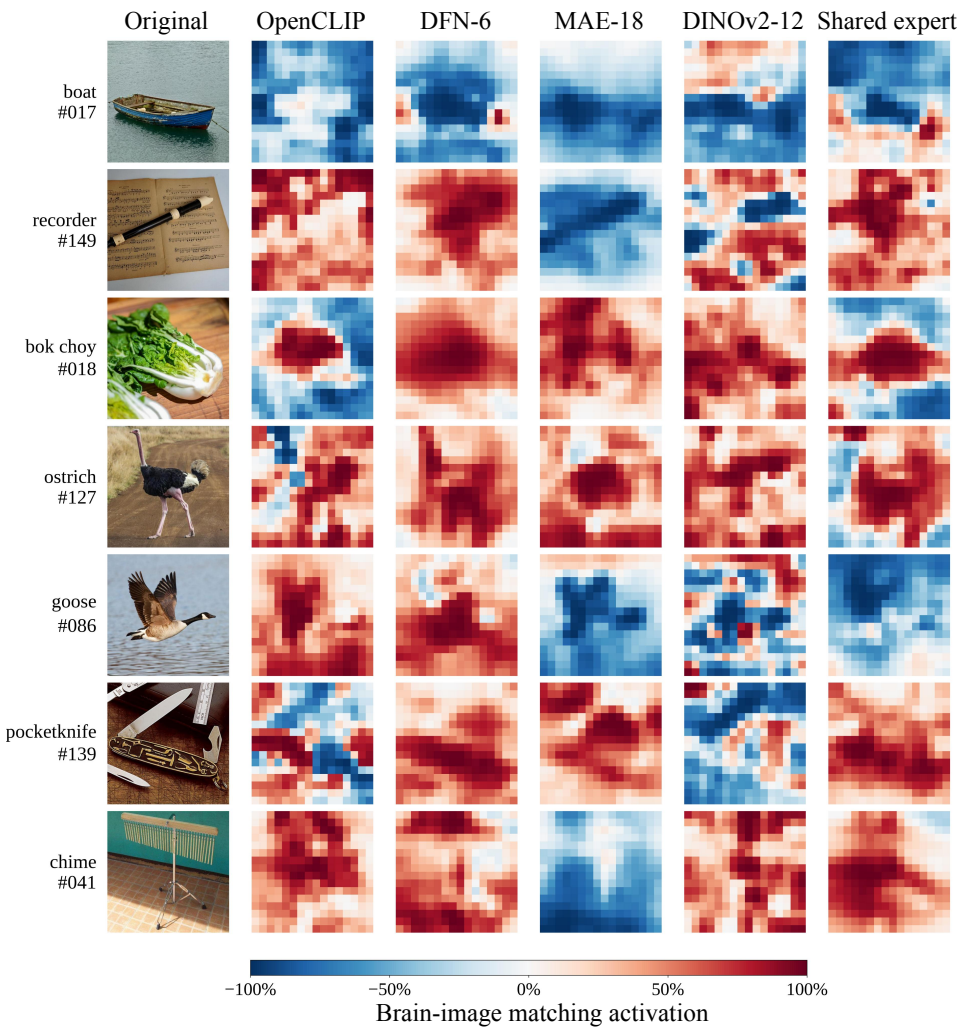


Figure A.5: **Additional brain–image matching attributions.** Each row shows one test image and ten-subject mean maps for four teacher branches and the shared expert. Red/blue indicate positive/negative matching contributions; intensity is a within-map percentile. Masking and averaging follow Section D.2.1.

D.2.2 Concept Associations in Model Outputs

Strong feature alignment need not imply the same concept organization. We therefore compare concept associations within each subject and inspect representative neurons in pooled activations. The concept galleries are collected in Appendix E.1; the selection procedure and quantitative comparisons are given below. Visual adaptation improves paired alignment, especially for MAE and DINOv2, and the shared expert has the strongest paired alignment (Figure 6(a)); the analyses below ask how this ordering relates to semantic correspondence.

Subject-wise concept scores. For each subject and shared-expert neuron, we select the category with the highest signed Pearson correlation on 16,540 training images and evaluate it on 200 test images. Selection is independent for student and visual features, with test scores averaged across subjects (Figure 6(b)); Figure A.6 extends this comparison to all branches. We also compare training-image SoftWPMI concept profiles, centering across neurons within each subject and L2-normalizing before computing paired brain–visual cosine similarity. The shared expert and DINOv2 show the closest correspondence, and all branches exceed the neuron-shuffle reference (Figure 6(c)). This ordering differs from paired feature alignment, distinguishing concept structure from feature similarity.

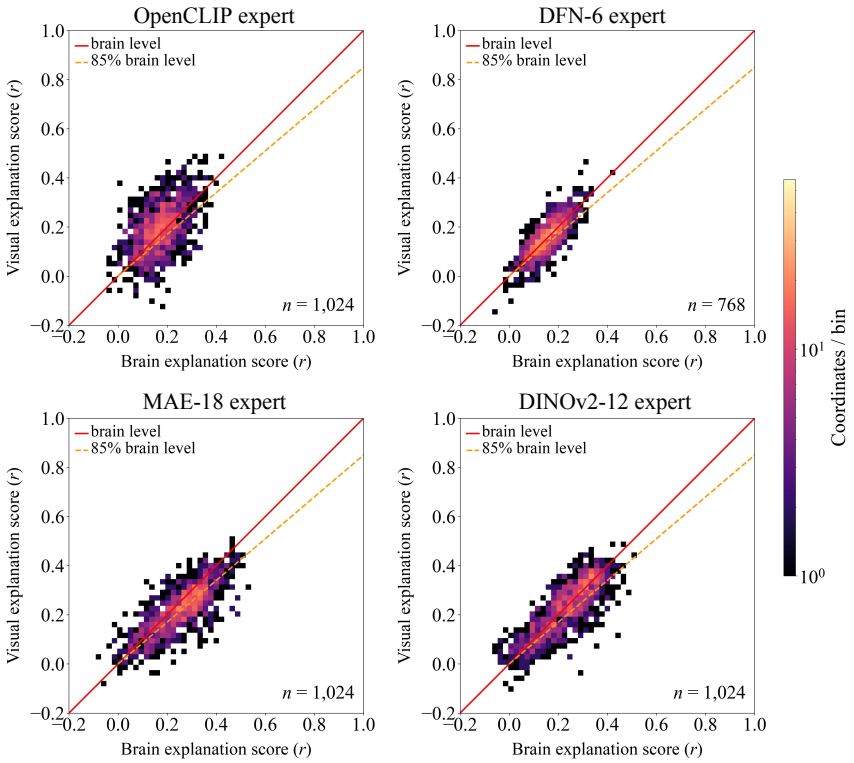


Figure A.6: **Held-out concept explanation scores.** Student-output (horizontal) versus visual-feature (vertical) Pearson scores, averaged over ten subjects after separate training-only concept selection (Section D.2.2). Each branch has 1,024 neurons except DFN-6 (768). Colors show log bin counts; red: $y = x$; orange dashed: $y = 0.85x$.

Concept-gallery construction. Figures A.7–A.12 show six neurons per output, selected by mean activation over their 100 strongest training responses after averaging across ten subjects. For gated outputs, each subject's normalization and gating precede pooling. Using the Cos variant of CLIP-Dissect (Oikarinen & Weng, 2023), we match neuron activation vectors $\mathbf{q}_j^{\mathrm{tr}}$ to CLIP ViT-L/14 image–text response vectors $\mathbf{p}_c^{\mathrm{tr}}$ over the same 16,540 training images:

$$c_j^\star \in \arg\max_{c \in \mathcal{C}} \mathrm{sim}(\mathbf{q}_j^{\mathrm{tr}}, \mathbf{p}_c^{\mathrm{tr}}), \qquad |\mathcal{C}| = 27. \tag{30}$$

Responses remain signed and uncentered. The probing set includes the internal validation split; the 200 test images have disjoint object concepts. Each gallery shows the five highest-scoring categories and five most activating test images. Test images do not select neurons or categories, and word size represents relative association strength.

The 27 labels are *animal*, *bird*, *body part*, *clothing*, *clothing accessory*, *container*, *dessert*, *drink*, *electronic device*, *food*, *fruit*, *furniture*, *home decor*, *insect*, *kitchen appliance*, *kitchen tool*, *medical equipment*, *musical instrument*, *office supply*, *part of car*, *plant*, *sports equipment*, *tool*, *toy*, *vegetable*, *vehicle*, and *weapon*. Pooled galleries do not imply that corresponding neurons share selectivity across subjects. Quantitative scores and category proportions are computed within subject before averaging (Section D.2.2).

D.2.3 Feature Inversion under Increasing Image Priors

We examine which visual properties can be recovered from the retrieval representations using direct pixel optimization, a deep image prior (DIP), and guided diffusion, following DreamSim (Fu et al., 2023). Inversion seeks an image whose features match a fixed EEG prediction or image-derived representation; it does not train a new brain decoder. Appendix E.2 (Figures A.13–A.15) compares subject 08 predictions, obtained by averaging 80 EEG trials before encoding, with image-derived features for the same four stimuli. The figures cover the four teacher spaces, shared fusion, and gated

BHA, with image-derived DreamSim features as an independent reference. Reconstruction accuracy is evaluated separately in Section D.3.

Feature matching in the retrieval spaces. We retain the notation of Section 4.2: $\hat{\mathbf{e}}_{ik}$ is the fixed EEG prediction for stimulus i, $\mathbf{e}_{ik}$ its image-derived feature, and $\tilde{\mathbf{e}}_{ik}$ the visual feature used for retrieval alignment. For a candidate image $\mathbf{I}$, write $\mathbf{e}_k(\mathbf{I}) = \phi_k(\mathbf{I})$ and $\tilde{\mathbf{e}}_k(\mathbf{I}) = V_k(\mathbf{e}_k(\mathbf{I}))$ for $k \geq 1$. For the shared space, $\mathbf{e}_0(\mathbf{I})$ is the fused feature in Eq. (5), and $\tilde{\mathbf{e}}_0(\mathbf{I}) = \mathbf{e}_0(\mathbf{I})$. Thus, evaluating these mappings at $\mathbf{I}_j$ recovers $\mathbf{e}_{jk}$ and $\tilde{\mathbf{e}}_{jk}$ from the main text. EEG inversion matches $\hat{\mathbf{e}}_{ik}$ in the adapted retrieval space; image-feature inversion matches $\mathbf{e}_{ik}$ in its original visual space.

For gated BHA, we preserve the score composition in Eq. (9) by concatenating weighted feature blocks:

$$\begin{aligned}\hat{\mathbf{e}}_i^{\mathrm{BHA}} &= \mathrm{norm}(\hat{\mathbf{e}}_{i0}) \oplus \bigoplus_{k=1}^{K} \sqrt{g_i \pi_{ik}}\ \mathrm{norm}(\hat{\mathbf{e}}_{ik}),\\ \tilde{\mathbf{e}}_i^{\mathrm{BHA}}(\mathbf{I}) &= \mathrm{norm}(\tilde{\mathbf{e}}_0(\mathbf{I})) \oplus \bigoplus_{k=1}^{K} \sqrt{g_i \pi_{ik}}\ \mathrm{norm}(\tilde{\mathbf{e}}_k(\mathbf{I})).\end{aligned} \tag{31}$$

Here $\oplus$ denotes concatenation, and the EEG-dependent g_i and π_{ik} remain fixed. Both vectors have squared norm $1 + g_i$; hence their cosine similarity at $\mathbf{I} = \mathbf{I}_j$ is $S_{ij}/(1 + g_i)$, preserving the retrieval ranking for query i.

For a feature distance d, the matching loss for each displayed branch is

$$\mathcal{D}_{i,d}(\mathbf{I}) = \begin{cases} d(\hat{\mathbf{e}}_{ik}, \tilde{\mathbf{e}}_k(\mathbf{I})), & \text{EEG branch } k,\\ d(\mathbf{e}_{ik}, \mathbf{e}_k(\mathbf{I})), & \text{image branch } k,\\ d(\hat{\mathbf{e}}_i^{\mathrm{BHA}}, \tilde{\mathbf{e}}_i^{\mathrm{BHA}}(\mathbf{I})), & \text{gated BHA},\\ d(\phi_{\mathrm{DS}}(\mathbf{I}_i), \phi_{\mathrm{DS}}(\mathbf{I})), & \text{DreamSim reference}, \end{cases} \tag{32}$$

where $k = 0, \ldots, K$ and ϕ_{DS} is the DreamSim feature ensemble. The first argument of each distance is fixed; only the candidate image changes during inversion. EEG-conditioned branches use no GT-image features or crops. Brain encoders, retrieval readouts, and visual mappings remain frozen; gradients through the visual mappings are retained to optimize the image. All methods produce 256×256 images.

Direct pixel optimization. Starting from random RGB pixels, we minimize cosine distance using sim from Section 3:

$$\min_{\mathbf{I}}\ \mathbb{E}_{\mathcal{T}}[\mathcal{D}_{i,d_{\cos}}(\mathcal{T}\mathbf{I})]\,, \qquad d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \mathrm{sim}(\mathbf{a}, \mathbf{b}). \tag{33}$$

Here $\mathcal{T}$ samples a 224×224 candidate crop. We use 2,500 Adam updates at learning rate 0.01, without an explicit image regularizer.

Deep image prior. We instead parameterize the candidate as $\mathbf{I} = G_\omega(\boldsymbol{\xi})$, where G_ω is a randomly initialized convolutional encoder–decoder with skip connections and $\boldsymbol{\xi}$ is fixed random noise (Ulyanov et al., 2018). Its architecture provides the image prior; only the generator parameters are optimized:

$$\min_{\omega}\ \mathcal{D}_{i,d_{\cos}}(G_\omega(\boldsymbol{\xi})). \tag{34}$$

The feature encoder receives the full generated image with its standard preprocessing. We use 3,000 Adam updates at learning rate 0.01; the generator is not pretrained on images.

Guided diffusion. A frozen unconditional ImageNet ADM (Dhariwal & Nichol, 2021) supplies a learned image prior. Let $\mathbf{x}_t$ denote its noisy image state and $\hat{\mathbf{x}}_0(\mathbf{x}_t, t)$ its clean-image prediction. Following CLIP-guided diffusion,[1] we evaluate feature matching on a clean/noisy blend:

$$\mathbf{u}_t = \rho_t \hat{\mathbf{x}}_0 + (1 - \rho_t)\mathbf{x}_t, \qquad \rho_t = \sqrt{1 - \bar{\alpha}_t^{\mathrm{ADM}}}, \qquad \mathbf{r}_t = (\mathbf{u}_t + 1)/2, \tag{35}$$

[1] https://github.com/afiaka87/clip-guided-diffusion/tree/a631a06b51ac5c6636136fab27833c68862eaa24.

where $\bar{\alpha}_t^{\mathrm{ADM}}$ is the cumulative signal-retention factor and $\mathbf{r}_t$ is the corresponding RGB candidate. Using the same norm as in the main text, define $d_{\mathrm{sph}}(\mathbf{a}, \mathbf{b}) = 2\arcsin^2(\|\,\mathrm{norm}(\mathbf{a}) - \mathrm{norm}(\mathbf{b})\|_2/2)$. The guidance objective combines feature matching with image regularization:

$$\mathcal{L}_{\mathrm{inv},t}(\mathbf{x}_t) = \frac{\lambda_f}{M_c}\sum_{m=1}^{M_c}\mathcal{D}_{i,d_{\mathrm{sph}}}(\mathcal{T}_m\mathbf{r}_t) + \lambda_{\mathrm{TV}}\,\mathrm{TV}(\mathbf{u}_t) + \lambda_{\mathrm{range}}\mathcal{R}(\hat{\mathbf{x}}_0). \tag{36}$$

Here TV averages squared horizontal and vertical pixel differences, and $\mathcal{R}(\mathbf{v}) = \mathrm{mean}[(\mathbf{v} - \mathrm{clip}(\mathbf{v}, -1, 1))^2]$. We use $M_c = 64$ random candidate crops, resized to 224×224, and $(\lambda_f, \lambda_{\mathrm{TV}}, \lambda_{\mathrm{range}}) = (1000, 150, 50)$. The gradient passes through the frozen visual mappings and denoiser to shift the reverse-process mean:

$$\boldsymbol{\mu}_t^{\mathrm{guided}} = \boldsymbol{\mu}_t - \Sigma_t\nabla_{\mathbf{x}_t}\mathcal{L}_{\mathrm{inv},t}(\mathbf{x}_t). \tag{37}$$

Ancestral sampling retains the pretrained reverse covariance Σ_t and uses all 1,000 steps of the linear noise schedule.

Comparison and interpretation. All branches share hyperparameters and sampling noise within each method. Each inversion matches one global feature vector; crops, when used, come only from the candidate image. This adapts DreamSim's analysis to global features rather than reproducing its target-image crop conditioning. The displayed stimuli were selected from eight correctly retrieved images based on the visual quality of their image-conditioned DreamSim inversions; failed EEG inversions remain included. These qualitative examples show that stronger image priors can improve plausibility while changing identity, pose, or layout (Figures A.13–A.15). They do not estimate average reconstruction performance or establish a ranking of the information preserved by different representations.

D.3 Comprehensive Reconstruction Performance

D.3.1 Evaluation and controls

We assess structural fidelity and semantic identification with an established eight-metric suite for brain-to-image reconstruction, also used by MindEye2 (Scotti et al., 2024) and ATM (Li et al., 2024). Scores are averaged over 200 test images, three generations per image, and ten subjects; image-conditioned controls are evaluated once on the shared stimulus set. Deterministic VAE decoding produces one image per stimulus. The CLIP identification metric uses ViT-L/14, distinct from the ViT-H/14 conditioning encoder.

Generator differences. Table A.13 summarizes the systems compared in Table 2. Fine-tuning includes the image-synthesis backbone or its pretrained conditioning interface. AVDE fine-tunes VAR (Dai et al., 2026); MBP keeps the SDXL backbone and VAE frozen while training visual fusion and fine-tuning the IP-Adapter projection and attention modules on ImageNet (Zheng et al., 2026). ATM-S and CC-All learn embedding priors (Li et al., 2024; Zhang et al., 2025), whereas BHA trains brain-side predictors while keeping SDXL-Turbo and IP-Adapter frozen.

Table A.13: **Generator and sampling comparison.** Fine-tuning includes the generation backbone or pretrained conditioning interface; †: settings from released code.

Method	Generator / interface	Fine-tuned?	Resolution	Synthesis steps
ATM-S†	SDXL-Turbo + IP-Adapter	No	512×512	4 denoising steps (released example)
CC-All	SDXL-Turbo + IP-Adapters	No	512×512 †	5 denoising steps
AVDE	VAR + VQ-VAE	Yes	256×256 †	10 autoregressive scales
MBP	SDXL-Turbo + IP-Adapter	Yes†	512×512 †	4 denoising steps
BHA	SDXL-Turbo + IP-Adapter	No	512×512	3 of 5 scheduled steps ($\alpha = 0.75$)

Code-derived settings follow the released implementations.[2] ATM-S uses the released four-step SDXL-Turbo example; the repository also provides other sampling settings, without identifying which

[2] ATM-S, CC-All, AVDE, and MBP (training, sampling).

produced the reported scores. Cross-paper results compare complete systems with different generators, conditioning, and sampling settings; they do not isolate brain-encoder quality. Our guidance and pretraining-source ablations instead hold the reconstruction settings constant (Appendix D.3.2).

Metrics. Following ATM (Li et al., 2024), PixCorr is RGB pixel Pearson correlation and SSIM measures structural fidelity. AlexNet(2)/(5) use post-ReLU features from convolutional layers 2/5 (`features.4`/`features.11`); these, InceptionV3, and CLIP report two-way identification: the fraction of distractors with lower feature correlation than the matched image (chance 0.5). EffNet-B and SwAV report feature correlation distance, $1 - r$. Higher is better except for these two distances.

Image-conditioned references. To assess generation with accurate conditioning, we replace brain predictions with ground-truth OpenCLIP ViT-H/14 image features, VAE latents, or both (Table 2). Semantic-only generation starts from noise; spatial-only generation directly decodes the latent; their combination uses $\alpha = 0.5$. All use the frozen generator with IP-Adapter scale 1.5, classifier-free guidance scale 0, and 512×512 output. Denoising follows a five-step schedule.

Chance level. To test whether reconstruction depends on stimulus-specific predictions, we jointly permute semantic and spatial conditions across the 200 test images within each subject, so that no image retains its own conditions. Reconstructions are evaluated against the original targets, using the same models, sampling settings, and random seeds as the main BHA row.

D.3.2 Reconstruction ablations

Visual target adaptation across tasks. These controls test whether joint brain–visual adaptation can improve angular alignment and retrieval while reducing compatibility with a frozen image generator. With fixed targets, only the brain predictor learns; adding a visual adapter lets both sides adjust their representations under the alignment objective. We compare frozen and adapted OpenCLIP targets under two objectives: $0.9\mathcal{L}_{\mathrm{MSE}} + 0.1\mathcal{L}_{\mathrm{BiCLIP}}$ (Table A.14) and BiCLIP alone (Table A.15). Each comparison evaluates the same predictions on both retrieval and reconstruction, separating angular agreement, image discrimination, and generator compatibility.

For each of ten THINGS-EEG2 subjects, BrainViT is trained from scratch to predict either native OpenCLIP ViT-H/14 features $\mathbf{e}_i = \phi(\mathbf{I}_i)$ or adapted targets $\tilde{\mathbf{e}}_i = V(\mathbf{e}_i)$. OpenCLIP remains frozen and no teacher fusion is used. Following the main-text notation, the visual adapter is $V(\mathbf{e}) = \mathbf{e} + \gamma W_{\mathrm{up}}\,\mathrm{GELU}(W_{\mathrm{down}}\,\mathrm{LN}(\mathbf{e}))$, with width 1,024, bottleneck rank 512, and 1,050,625 trainable parameters. It is initialized near identity ($\gamma = 0.1$ and output weights drawn from $\mathcal{N}(0, 10^{-6})$) and jointly optimized with BrainViT, without an additional geometry regularizer. BiCLIP uses cosine similarity at temperature 0.07; MSE, when present, regresses unnormalized features toward the target used by that model.

Paired models share brain-encoder initialization, image split, minibatch order, and optimization settings: AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 1,024, and 100 epochs. We evaluate the final model; validation is diagnostic only. Retrieval ranks 200 candidates in each model's target space, without CSLS. Reconstruction feeds the same brain prediction directly to frozen SDXL-Turbo and IP-Adapter, with no spatial condition or learned inverse mapping: five denoising steps, IP-Adapter scale 1.5, guidance scale 0, and three matched noise draws per image. Thus any benefit to matching is evaluated alongside the prediction's suitability for the unchanged generator interface.

Under BiCLIP alone, adaptation increases mean paired brain–target cosine similarity from 0.217 to 0.481 and Raw Top-1 retrieval from 30.05% to 37.80% (Table A.15). Reconstruction nevertheless declines: CLIP identification falls from 0.754 to 0.706, and the mean of every other reconstruction metric also worsens in its corresponding direction. The higher cosine similarity is measured against each model's own target, so it indicates improved directional agreement in the learned space, rather than preservation of the native OpenCLIP representation.

Under MSE–BiCLIP supervision, paired cosine increases only from 0.583 to 0.598, while Raw Top-1 drops from 21.10% to 19.00% and CLIP identification from 0.838 to 0.541 (Table A.14). Angular agreement alone therefore does not ensure better discrimination, and the retrieval benefit of adaptation

depends on the objective. Here MSE constrains agreement with the adapted target; it does not anchor that target to the original visual feature.

Together, the controls show that joint adaptation can improve contrastive retrieval while degrading reconstruction through a fixed generator. This is consistent with the distinction between alignment and native-output fidelity in Proposition 2, and motivates BHA's separate predictors for generator-native semantic and spatial conditions. The results establish a compatibility cost, but do not distinguish loss of visual information from a change in feature coordinates or scale that a learned inverse mapping might recover.

Table A.14: **Visual target adaptation under MSE + BiCLIP.** Frozen versus adapted OpenCLIP targets, trained with $0.9\mathcal{L}_{\mathrm{MSE}} + 0.1\mathcal{L}_{\mathrm{BiCLIP}}$. THINGS-EEG2 within-subject means over ten subjects, with SD below each mean. Retrieval is Raw 200-way accuracy (%); reconstruction uses the same frozen generator. See Section D.3.2 for details.

	Retrieval (%)		Reconstruction							
Visual target	Top-1↑	Top-5↑	PixCorr↑	SSIM↑	AlexNet (2)↑	AlexNet (5)↑	Inception V3↑	CLIP↑	EffNet-B↓	SwAV↓
Frozen target	**21.10** ±3.84	**50.30** ±6.26	**0.167** ±0.017	**0.414** ±0.021	**0.830** ±0.012	**0.902** ±0.011	**0.793** ±0.022	**0.838** ±0.019	**0.858** ±0.015	**0.533** ±0.012
Adapted target	19.00 ±5.51	44.45 ±6.19	0.099 ±0.021	0.370 ±0.030	0.657 ±0.011	0.665 ±0.012	0.550 ±0.018	0.541 ±0.016	0.968 ±0.003	0.683 ±0.014

Table A.15: **Visual target adaptation under BiCLIP alone.** Frozen versus adapted OpenCLIP targets, trained with $\mathcal{L}_{\mathrm{BiCLIP}}$. THINGS-EEG2 within-subject means over ten subjects, with SD below each mean. Retrieval is Raw 200-way accuracy (%); reconstruction uses the same frozen generator. See Section D.3.2 for details.

	Retrieval (%)		Reconstruction							
Visual target	Top-1↑	Top-5↑	PixCorr↑	SSIM↑	AlexNet (2)↑	AlexNet (5)↑	Inception V3↑	CLIP↑	EffNet-B↓	SwAV↓
Frozen target	30.05 ±5.07	60.70 ±5.68	**0.168** ±0.016	**0.391** ±0.018	**0.779** ±0.019	**0.823** ±0.023	**0.725** ±0.026	**0.754** ±0.030	**0.900** ±0.013	**0.612** ±0.015
Adapted target	**37.80** ±4.82	**69.50** ±5.37	0.160 ±0.016	0.389 ±0.015	0.764 ±0.024	0.809 ±0.028	0.685 ±0.025	0.706 ±0.027	0.917 ±0.012	0.635 ±0.016

VAE guidance. We test whether the predicted spatial latent can preserve structure throughout denoising by comparing standard sampling with the training-free guidance in Appendix C.5. Guidance settings were selected on Sub-01 and applied unchanged to the other nine subjects. Guidance modestly improves pixel and structural fidelity while leaving CLIP identification largely unchanged (Table 2).

Pretraining-source comparison. To test whether richer pretraining supervision benefits reconstruction, we compare BrainViT trained against frozen OpenCLIP ViT-H/14 features with BrainViT trained against BHA's fused visual targets (Table A.16). Both use the same reconstruction architecture, training schedules, data splits, and model-selection criteria. Reconstruction heads read the pretrained backbone directly. We evaluate ten subjects at $\alpha \in \{0, 0.75, 1\}$, with the generation settings above and three generations per test image.

Fused pretraining improves SSIM at all three strengths, whereas OpenCLIP has higher CLIP identification at $\alpha = 0.75$ and 1. Additional visual supervision therefore benefits some reconstruction properties without uniformly improving semantic identification.

Reconstruction strength. We examine the balance between preserving the predicted spatial latent and allowing semantic denoising by varying $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ with the same predictors and random seeds (Table A.17). The corresponding numbers of denoising updates are $0, 1, 2, 3, 5$; all other settings follow Appendix C.5. Avg and SD summarize ten subjects, and rankings use unrounded scores. Lower strengths favor pixel and structural fidelity, whereas the default $\alpha = 0.75$ achieves the highest mean scores on most feature-based metrics.

Table A.16: **Reconstruction: OpenCLIP vs. Fused VFM pretraining.** Ten-subject means over 200 test images, with three generations per image for $\alpha > 0$ and deterministic decoding for $\alpha = 0$. $\alpha = 0/1$: spatial-/semantic-only conditions.

α	Pretraining	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep.↑	CLIP↑	EffNet↓	SwAV↓
0	OpenCLIP	0.333	0.580	0.604	0.565	0.507	0.514	0.985	0.689
	Fused VFM	0.335	0.583	0.599	0.561	0.507	0.515	0.984	0.687
0.75	OpenCLIP	0.274	0.448	0.873	0.922	0.798	0.841	0.835	0.516
	Fused VFM	0.274	0.459	0.872	0.921	0.794	0.831	0.843	0.515
1	OpenCLIP	0.189	0.413	0.849	0.915	0.799	0.839	0.837	0.519
	Fused VFM	0.185	0.427	0.845	0.910	0.790	0.828	0.844	0.520

Table A.17: **Reconstruction strength sweep on THINGS-EEG2.**

Subject	α	**PixCorr↑**	**SSIM↑**	**AlexNet(2)↑**	**AlexNet(5)↑**	**InceptionV3↑**	**CLIP↑**	**EffNet-B↓**	**SwAV↓**
sub-01	0	0.3514	**0.5831**	0.6086	0.5713	0.4969	0.5106	0.9864	0.6859
sub-01	0.25	**0.3889**	0.5739	0.7526	0.8240	0.7031	0.7322	0.8956	0.5794
sub-01	0.5	0.3498	0.5226	0.8693	0.9060	0.7589	0.8098	0.8559	0.5285
sub-01	**0.75 (Ours)**	0.2983	0.4607	**0.8919**	**0.9226**	0.7788	**0.8229**	0.8425	**0.5216**
sub-01	1	0.2073	0.4293	0.8612	0.9119	**0.7814**	0.8212	**0.8392**	0.5228
sub-02	0	0.3510	**0.5846**	0.5953	0.5582	0.5110	0.5106	0.9844	0.6885
sub-02	0.25	**0.3877**	0.5739	0.7633	0.8448	0.7440	0.7477	0.8887	0.5735
sub-02	0.5	0.3304	0.5125	0.8692	0.9220	0.7901	0.8265	0.8619	0.5210
sub-02	**0.75 (Ours)**	0.2669	0.4393	**0.8804**	**0.9319**	0.7922	**0.8294**	**0.8576**	**0.5173**
sub-02	1	0.1866	0.4038	0.8610	0.9241	**0.8043**	0.8251	0.8578	0.5218
sub-03	0	0.3266	**0.5801**	0.5945	0.5601	0.5211	0.5163	0.9829	0.6865
sub-03	0.25	**0.3762**	0.5730	0.7519	0.8198	0.7182	0.7540	0.8917	0.5862
sub-03	0.5	0.3345	0.5200	0.8520	0.9087	0.7850	0.8263	0.8498	0.5258
sub-03	**0.75 (Ours)**	0.2766	0.4458	**0.8640**	**0.9234**	**0.8092**	**0.8382**	**0.8427**	**0.5202**
sub-03	1	0.1841	0.4060	0.8351	0.9088	0.7879	0.8299	0.8490	0.5220
sub-04	0	0.2902	**0.5790**	0.5697	0.5443	0.5089	0.5123	0.9853	0.6842
sub-04	0.25	**0.3386**	0.5679	0.7289	0.8315	0.7653	0.7799	0.8747	0.5654
sub-04	0.5	0.2967	0.5108	0.8361	0.8960	0.7918	0.8365	0.8462	0.5170
sub-04	**0.75 (Ours)**	0.2485	0.4381	**0.8584**	**0.9172**	0.8021	0.8509	0.8395	**0.5115**
sub-04	1	0.1766	0.4020	0.8276	0.9123	**0.8050**	**0.8570**	**0.8380**	0.5154
sub-05	0	0.3226	**0.5816**	0.5821	0.5535	0.4978	0.5127	0.9857	0.6866
sub-05	0.25	**0.3627**	0.5697	0.7394	0.8006	0.7127	0.7326	0.8977	0.5817
sub-05	0.5	0.3145	0.5100	0.8332	0.8757	0.7359	0.7885	0.8778	0.5381
sub-05	**0.75 (Ours)**	0.2657	0.4420	**0.8476**	0.8923	0.7260	**0.7941**	0.8717	0.5343
sub-05	1	0.1815	0.4107	0.8273	**0.8972**	**0.7417**	0.7918	**0.8668**	**0.5331**
sub-06	0	0.3559	**0.5829**	0.6021	0.5702	0.5005	0.5137	0.9823	0.6857
sub-06	0.25	**0.3927**	0.5730	0.7707	0.8260	0.7221	0.7424	0.8883	0.5748
sub-06	0.5	0.3407	0.5181	0.8657	0.9109	0.7707	0.8114	0.8579	0.5279
sub-06	**0.75 (Ours)**	0.2812	0.4443	**0.8841**	**0.9225**	**0.7827**	0.8178	**0.8528**	**0.5240**
sub-06	1	0.2138	0.4037	0.8484	0.9114	0.7757	**0.8191**	0.8536	0.5247
sub-07	0	0.3183	**0.5801**	0.5969	0.5636	0.5063	0.5133	0.9826	0.6877
sub-07	0.25	**0.3590**	0.5722	0.7541	0.8231	0.7249	0.7417	0.8884	0.5803
sub-07	0.5	0.3182	0.5237	0.8559	0.8968	**0.7957**	0.8080	0.8519	0.5303
sub-07	**0.75 (Ours)**	0.2614	0.4594	**0.8709**	**0.9154**	0.7886	0.8114	**0.8489**	**0.5230**
sub-07	1	0.1578	0.4206	0.8259	0.9000	0.7938	**0.8142**	0.8506	0.5266
sub-08	0	0.3627	**0.5846**	0.6236	0.5875	0.5078	0.5219	0.9812	0.6854
sub-08	0.25	**0.3955**	0.5759	0.7814	0.8441	0.7488	0.7768	0.8756	0.5685
sub-08	0.5	0.3551	0.5291	0.8719	0.9225	0.7985	0.8597	0.8281	0.5041
sub-08	**0.75 (Ours)**	0.3052	0.4757	**0.8908**	**0.9379**	0.8157	**0.8625**	**0.8157**	**0.4974**

Continued on next page

Subject	α	PixCorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	InceptionV3↑	CLIP↑	EffNet-B↓	SwAV↓
sub-08	1	0.2174	0.4473	0.8557	0.9237	**0.8213**	0.8532	0.8222	0.5049
sub-09	0	0.3705	**0.5833**	0.6243	0.5639	0.4930	0.5102	0.9827	0.6851
sub-09	0.25	**0.4105**	0.5759	0.7562	0.8217	0.7260	0.7287	0.8797	0.5713
sub-09	0.5	0.3649	0.5314	0.8578	0.9054	0.7755	0.8000	0.8573	0.5262
sub-09	**0.75 (Ours)**	0.3008	0.4698	**0.8738**	**0.9232**	0.7778	**0.8052**	0.8542	**0.5187**
sub-09	1	0.1919	0.4302	0.8382	0.9090	**0.7786**	0.8037	**0.8533**	0.5235
sub-10	0	0.3306	**0.5827**	0.6177	0.5727	0.5136	0.5158	0.9856	0.6919
sub-10	0.25	**0.3619**	0.5750	0.7448	0.8340	0.7687	0.7753	0.8747	0.5675
sub-10	0.5	0.3220	0.5248	0.8614	0.9129	0.8213	0.8570	0.8291	0.5032
sub-10	**0.75 (Ours)**	0.2702	0.4666	**0.8783**	**0.9324**	0.8351	**0.8670**	**0.8217**	**0.4971**
sub-10	1	0.1762	0.4324	0.8548	0.9203	**0.8356**	0.8562	0.8281	0.5076
Avg	0	0.3380	**0.5822**	0.6015	0.5645	0.5057	0.5137	0.9839	0.6867
SD		0.0246	0.0019	0.0177	0.0118	0.0086	0.0035	0.0018	0.0022
Avg	0.25	**0.3774**	0.5730	0.7543	0.8270	0.7334	0.7511	0.8855	0.5749
SD		0.0216	0.0026	0.0151	0.0128	0.0222	0.0196	0.0087	0.0069
Avg	0.5	0.3327	0.5203	0.8573	0.9057	0.7823	0.8224	0.8516	0.5222
SD		0.0207	0.0074	0.0135	0.0138	0.0236	0.0235	0.0148	0.0112
Avg	**0.75 (Ours)**	0.2775	0.4542	**0.8740**	**0.9219**	0.7908	**0.8300**	**0.8447**	**0.5165**
SD		0.0187	0.0139	0.0142	0.0125	0.0291	0.0245	0.0166	0.0117
Avg	1	0.1893	0.4186	0.8435	0.9119	**0.7925**	0.8271	0.8459	0.5202
SD		0.0187	0.0157	0.0143	0.0091	0.0262	0.0223	0.0138	0.0086

E SUPPLEMENTARY EXAMPLES

These galleries complement the representation and reconstruction analyses in Appendices D.2 and D.3.

E.1 NEURON CONCEPT ASSOCIATIONS

Figures A.7–A.12 show six training-selected neurons for each teacher-specific expert, the shared expert, and the gated BHA output. Each row pairs a five-concept word cloud with the five test images that elicit the highest mean activation across ten subjects. Neuron and concept selection use training images only; word sizes express relative scores within a neuron. These pooled associations do not establish correspondence between neurons across subjects. Section D.2.2 specifies the selection and pooling procedure; subject-wise quantitative scores are reported in Section D.2.2.



Figure A.7: **OpenCLIP-aligned student expert: concept associations.** Each row shows one of six training-selected neurons: its five highest-scoring concepts (left) and five highest-activating test images (right). Activation values are ten-subject means. See Section D.2.2 for more details.

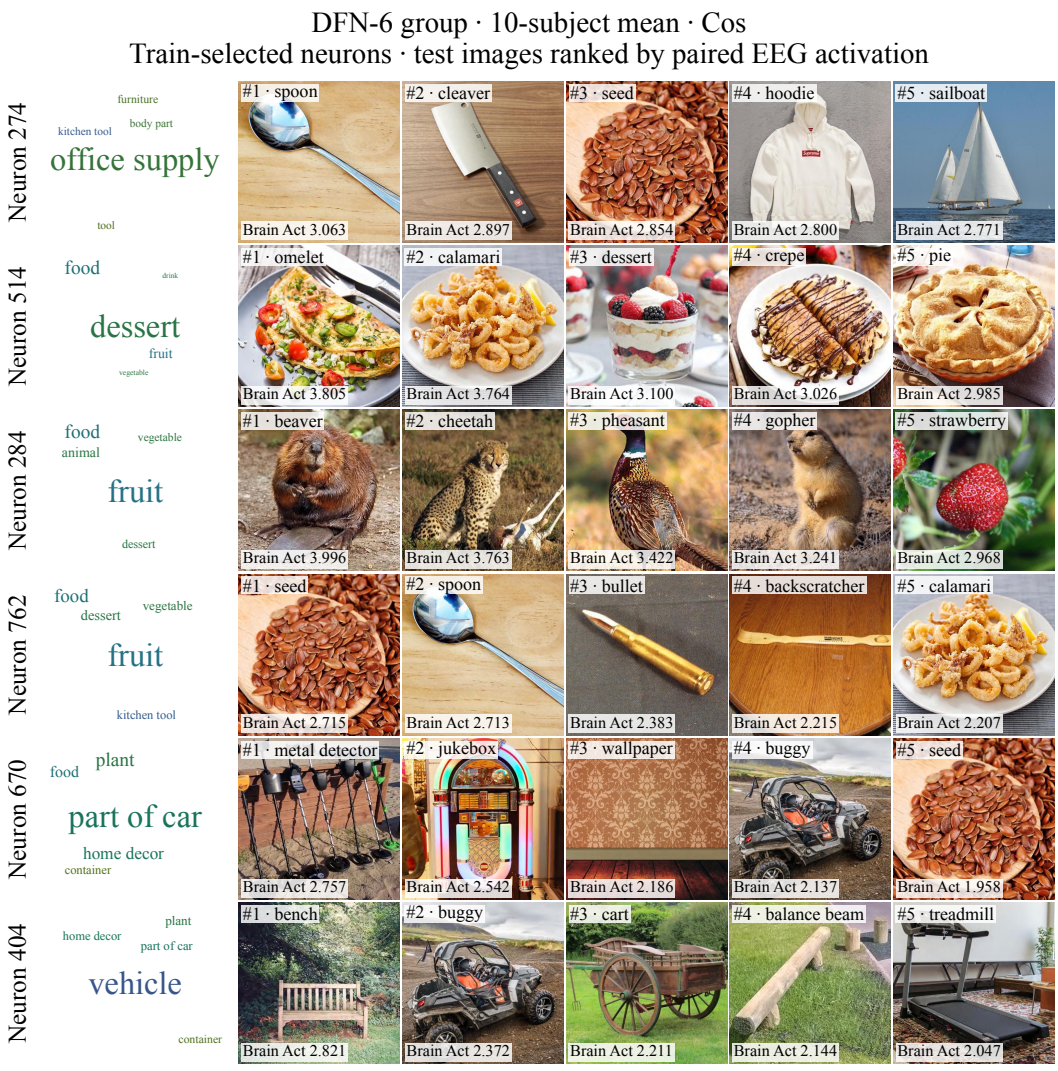


Figure A.8: **DFN-6-aligned student expert: concept associations.** Each row shows one of six training-selected neurons: its five highest-scoring concepts (left) and five highest-activating test images (right). Activation values are ten-subject means. See Section D.2.2 for more details.



Figure A.9: **MAE-18-aligned student expert: concept associations.** Each row shows one of six training-selected neurons: its five highest-scoring concepts (left) and five highest-activating test images (right). Activation values are ten-subject means. See Section D.2.2 for more details.

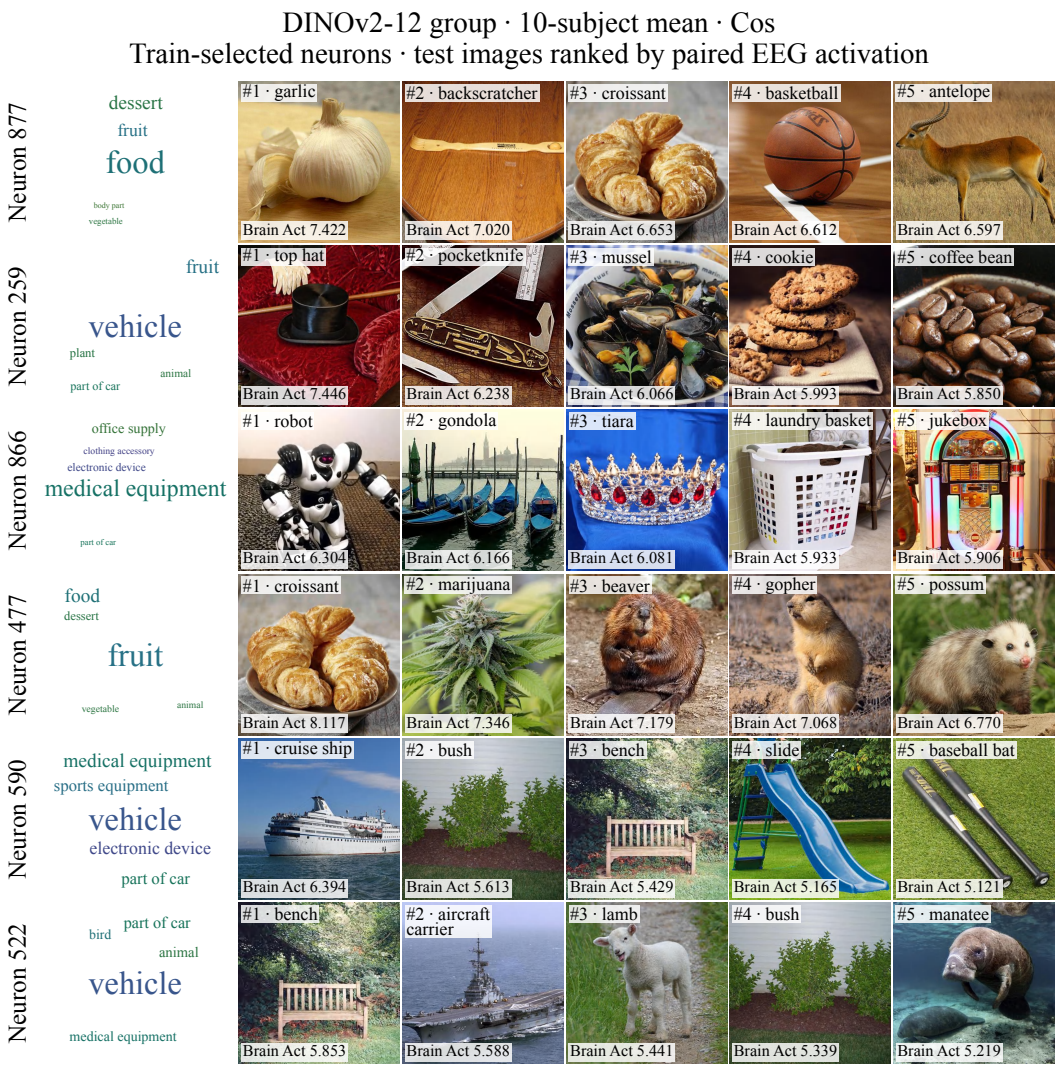


Figure A.10: **DINOv2-12-aligned student expert: concept associations.** Each row shows one of six training-selected neurons: its five highest-scoring concepts (left) and five highest-activating test images (right). Activation values are ten-subject means. See Section D.2.2 for more details.



Figure A.11: **Shared student expert: concept associations.** Each row shows one of six training-selected neurons: its five highest-scoring concepts (left) and five highest-activating test images (right). Activation values are ten-subject means. See Section D.2.2 for more details.



Figure A.12: **BHA gated student output: concept associations.** Each row shows one of six training-selected neurons: its five highest-scoring concepts (left) and five highest-activating test images (right). Activation values are ten-subject means. See Section D.2.2 for more details.

E.2 Feature Inversion Examples

Figures A.13–A.15 compare direct pixel optimization, a deep image prior, and guided diffusion for the same EEG-predicted and image-derived targets. They illustrate how the image prior affects visual plausibility and stimulus fidelity; Section D.2.3 defines the objectives and example selection.

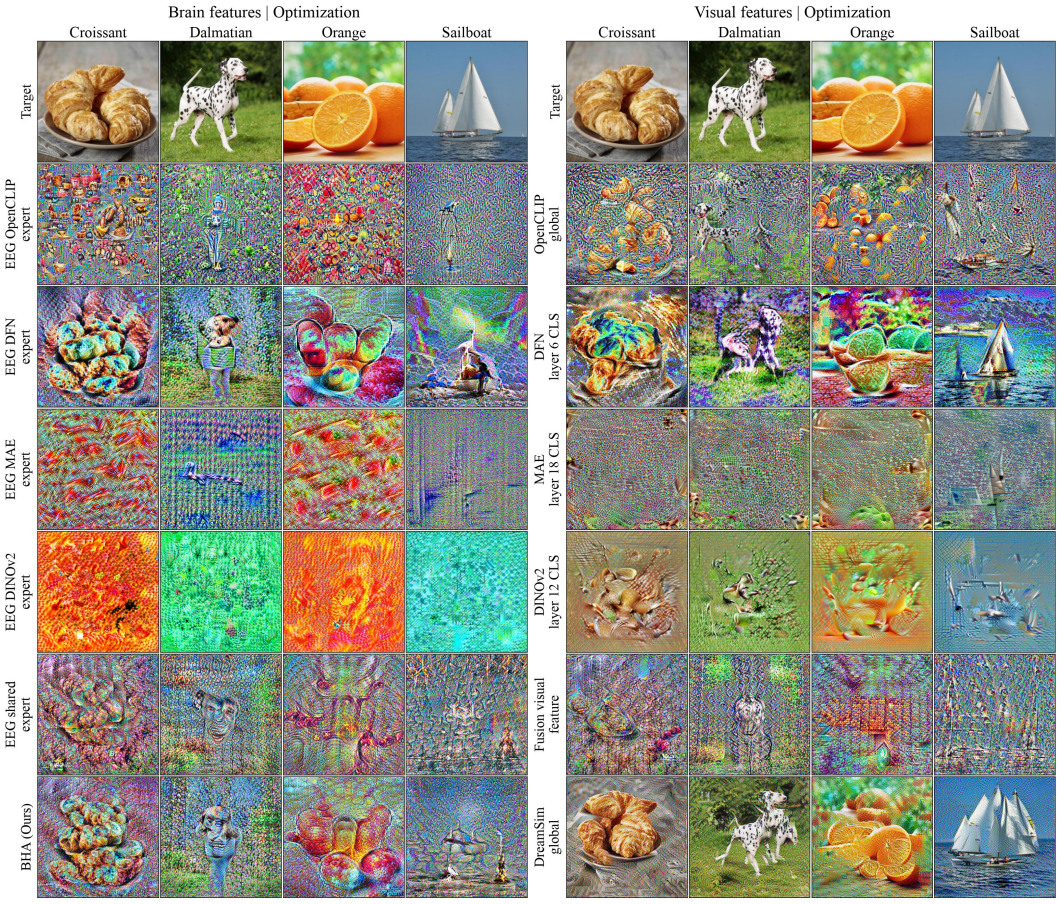


Figure A.13: **Direct pixel inversion.** Left: EEG-predicted features from subject 08 (80-trial averages before encoding); right: GT-image features. Columns: *croissant, dalmatian, orange, and sailboat*, with GT images above. Bottom rows: gated BHA (left) and an independent image-derived DreamSim reference (right). See Section D.2.3 for more details.

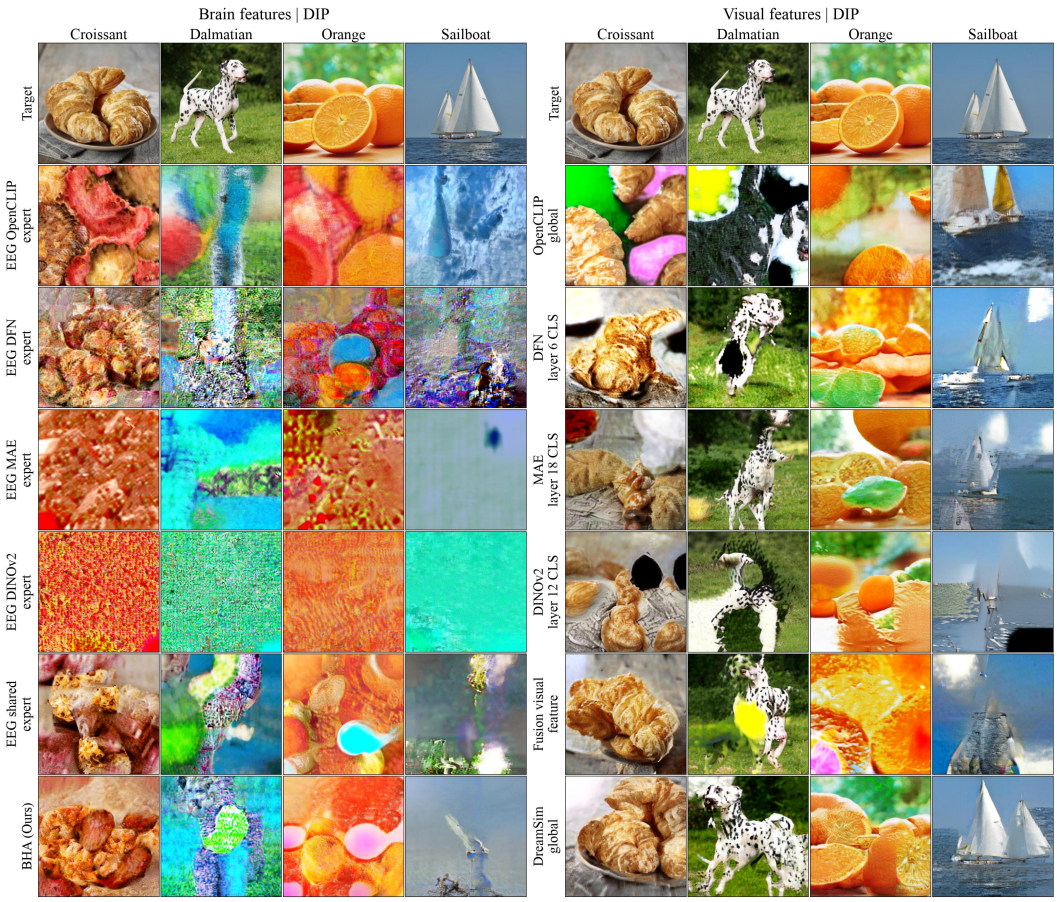


Figure A.14: **Feature inversion with a deep image prior.** A randomly initialized generator is fitted to each feature target. Left: EEG-predicted features from subject 08 (80-trial averages before encoding); right: GT-image features. Columns: *croissant, dalmatian, orange, and sailboat*, with GT images above. Bottom rows: gated BHA (left) and the image-derived DreamSim reference (right). See Section D.2.3 for more details.



Figure A.15: **Feature inversion with a pretrained diffusion prior.** A frozen unconditional ADM supplies the image prior. Left: EEG-predicted features from subject 08 (80-trial averages before encoding); right: GT-image features. Columns: *croissant, dalmatian, orange, and sailboat*, with GT images above. Bottom rows: gated BHA (left) and the image-derived DreamSim reference (right). See Section D.2.3 for more details.

E.3 RECONSTRUCTION EXAMPLES

Figures A.16–A.25 show reconstructions across mixing strengths for all ten subjects. Columns correspond to $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ and the target; $\alpha = 0.75$ is the default.

For each subject, we rank 200 test images by paired OpenCLIP ViT-H/14 cosine similarity, averaged equally across the five strengths using generation 0. The top 30%, middle 40%, and bottom 30% define the Best, Median, and Worst partitions. We display the three highest-scoring images, the three middle-partition images closest to the overall median, and the three lowest-scoring images. This selection illustrates variation in reconstruction fidelity rather than estimating average test performance.



Figure A.16: **Subject-01: reconstruction examples.** Columns: $\alpha = 0, 0.25, 0.5, 0.75, 1$ and target. Rows: score-selected Best/Median/Worst examples (Appendix E.3).



Figure A.17: **Subject-02: reconstruction examples.** Columns: $\alpha = 0, 0.25, 0.5, 0.75, 1$ and target. Rows: score-selected Best/Median/Worst examples (Appendix E.3).

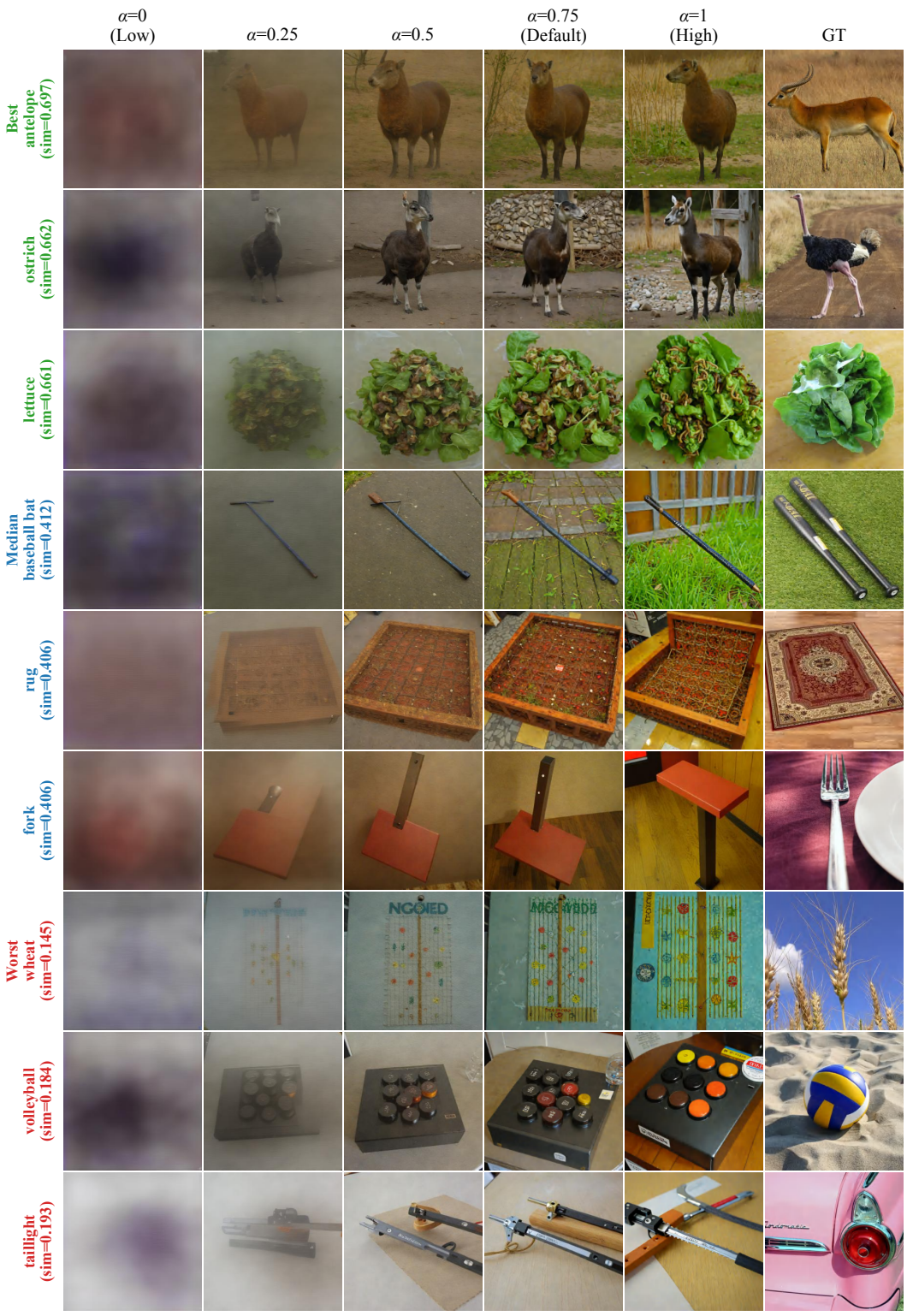


Figure A.18: **Subject-03: reconstruction examples.** Columns: $\alpha = 0, 0.25, 0.5, 0.75, 1$ and target. Rows: score-selected Best/Median/Worst examples (Appendix E.3).



Figure A.19: **Subject-04: reconstruction examples.** Columns: $\alpha = 0, 0.25, 0.5, 0.75, 1$ and target. Rows: score-selected Best/Median/Worst examples (Appendix E.3).

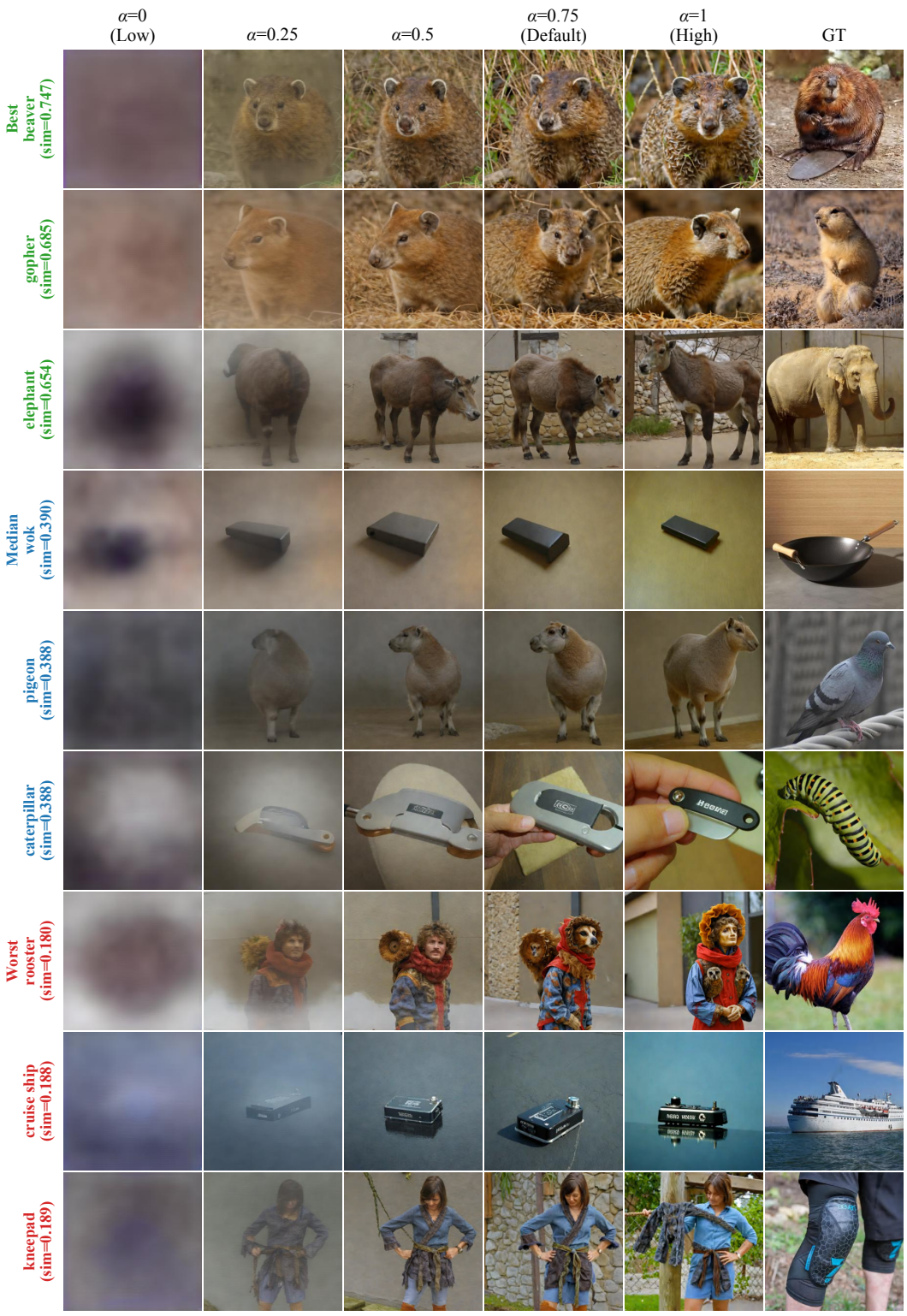


Figure A.20: **Subject-05: reconstruction examples.** Columns: $\alpha = 0, 0.25, 0.5, 0.75, 1$ and target. Rows: score-selected Best/Median/Worst examples (Appendix E.3).

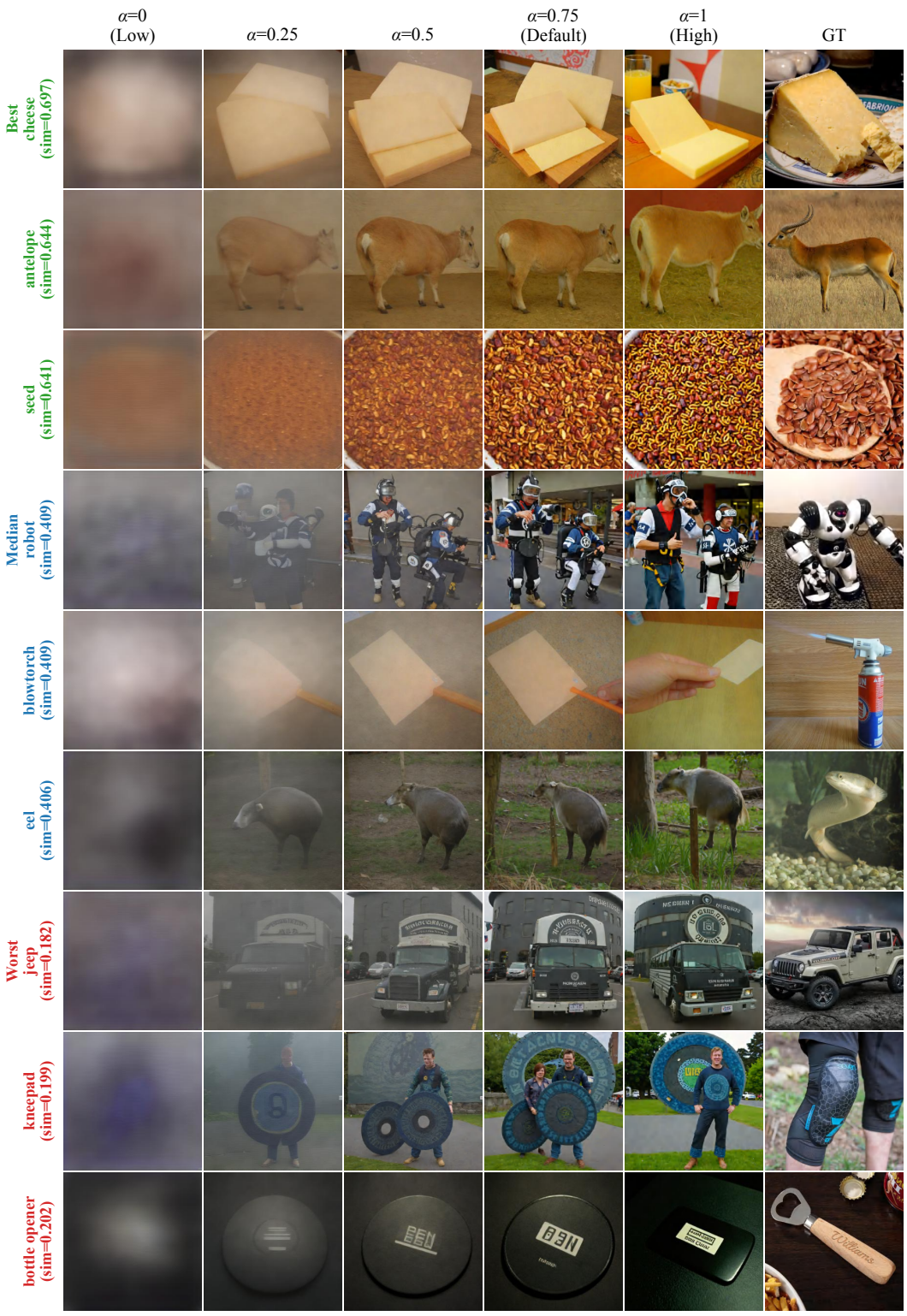


Figure A.21: **Subject-06: reconstruction examples.** Columns: $\alpha = 0, 0.25, 0.5, 0.75, 1$ and target. Rows: score-selected Best/Median/Worst examples (Appendix E.3).



Figure A.22: **Subject-07: reconstruction examples.** Columns: $\alpha = 0, 0.25, 0.5, 0.75, 1$ and target. Rows: score-selected Best/Median/Worst examples (Appendix E.3).

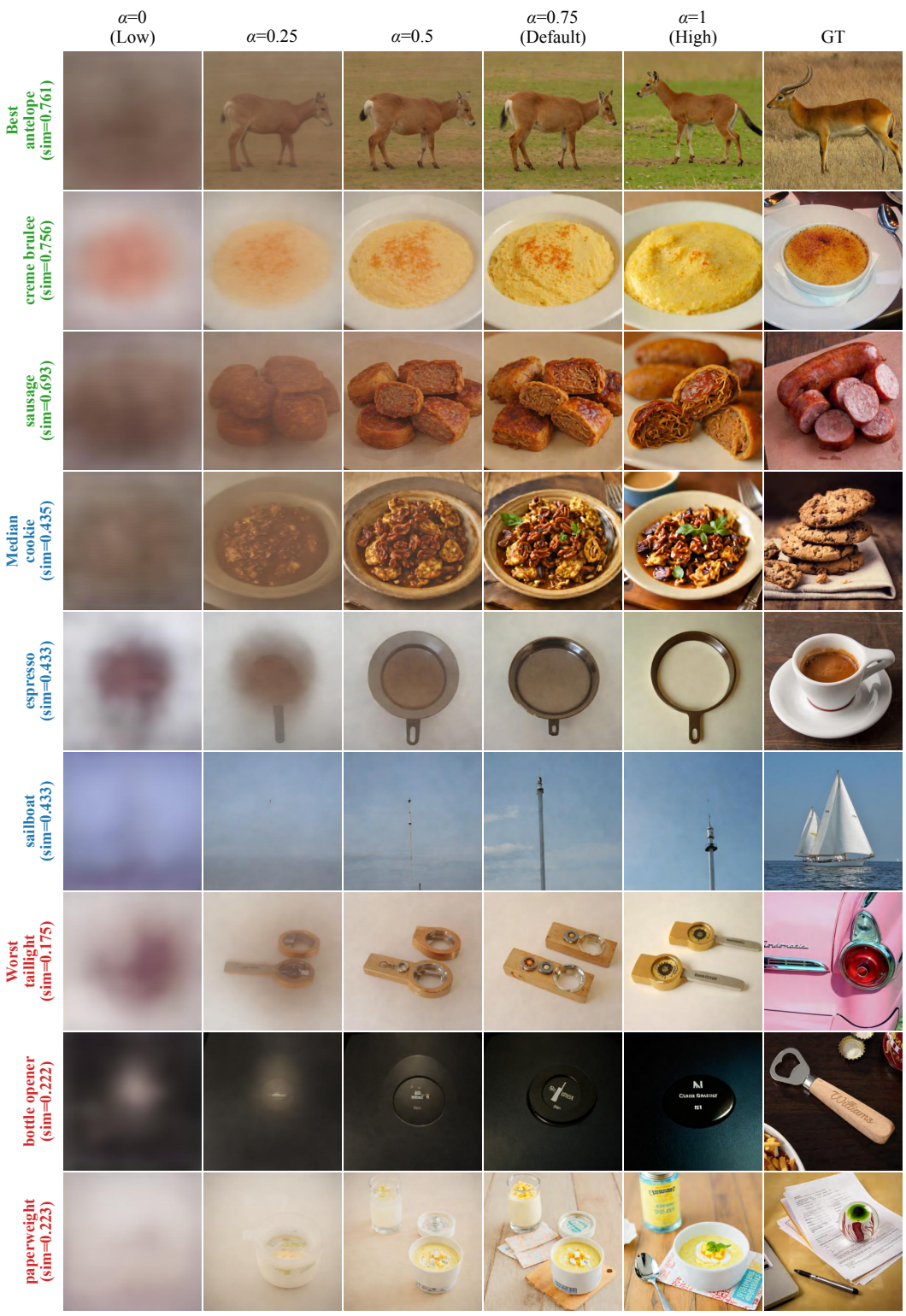


Figure A.23: **Subject-08: reconstruction examples.** Columns: $\alpha = 0, 0.25, 0.5, 0.75, 1$ and target. Rows: score-selected Best/Median/Worst examples (Appendix E.3).

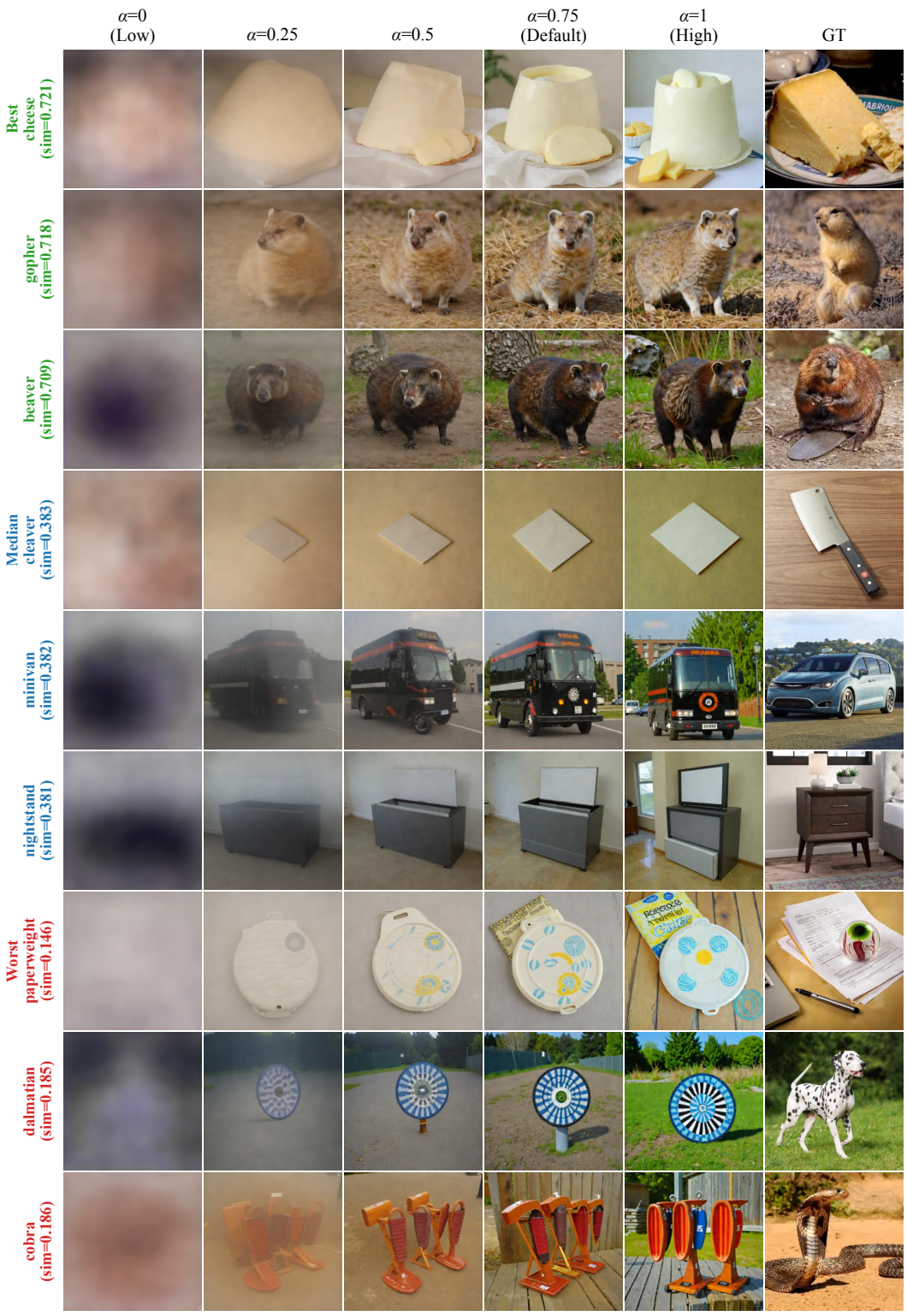


Figure A.24: **Subject-09: reconstruction examples.** Columns: $\alpha = 0, 0.25, 0.5, 0.75, 1$ and target. Rows: score-selected Best/Median/Worst examples (Appendix E.3).

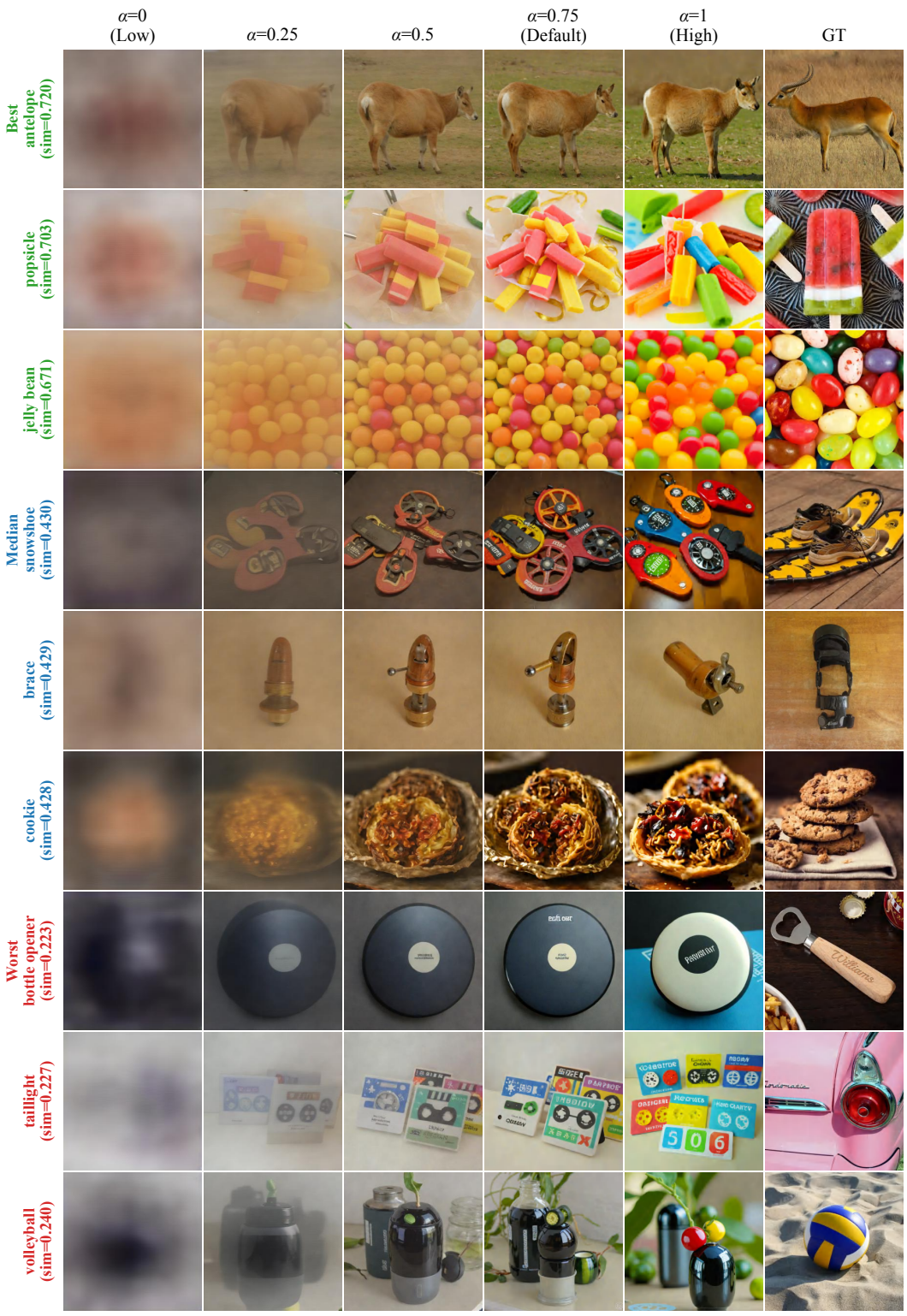


Figure A.25: **Subject-10: reconstruction examples.** Columns: $\alpha = 0, 0.25, 0.5, 0.75, 1$ and target. Rows: score-selected Best/Median/Worst examples (Appendix E.3).