Platonic Representations in the Human Brain: Unsupervised Recovery of Universal Geometry

Pablo Marcos-Manchón[1,2] **Rishi Jha**[3] **Lluís Fuentemilla**[1,2,4]

[1]Department of Cognition, Development and Educational Psychology, University of Barcelona
[2]Institute of Neurosciences, University of Barcelona
[3]Department of Computer Science, Cornell University
[4]Bellvitge Institute for Biomedical Research, Spain

Abstract

The Strong Platonic Representation Hypothesis suggests that representational convergence in artificial neural networks can be harnessed constructively: embeddings can be translated across models through a universal latent space without paired data. We ask whether an analogous geometry can be recovered across human brains. Using fMRI data from the Natural Scenes Dataset, we propose a self-supervised encoder that learns subject-specific embeddings from brain data alone by exploiting repeated stimulus presentations. We show that these independently learned spaces can be translated across subjects using unsupervised orthogonal rotations, without paired cross-subject samples or intermediate model representations. Synchronizing pairwise rotations into a single shared latent space further improves cross-subject retrieval. These results provide evidence for a shared geometry in learned representations of human visual responses: subject-specific fMRI embeddings can be aligned across individuals within a common coordinate system.

1 Introduction

The *Platonic Representation Hypothesis* posits that independently trained artificial neural networks converge toward geometrically similar representations by recovering shared latent structure in the world [1–5]. Recent constructive work pushes this hypothesis further: if representations share a common geometry, then embeddings from one model should be translatable into another model's latent space without shared inputs or paired supervision [6, 7]. However, evidence for this strong form of representational convergence has so far come almost entirely from artificial vision and language systems [5, 8, 9]. Whether the same principle extends to biological neural systems remains unknown.

In neuroscience, a long line of work on inter-subject synchrony and representational similarity shows that neural responses, and the geometries they induce, can be preserved across individuals during shared stimulus processing [10–13]. However, existing evidence typically depends on shared stimuli, paired measurements, or external reference spaces. Functional alignment methods for fMRI, such as *hyperalignment*, map neural responses into common spaces by using shared stimuli to establish cross-subject correspondences [14, 15]. Other approaches introduce anchors through model-derived feature spaces [16] or image-to-fMRI encoders [17]. What remains open is whether shared neural geometry can be recovered from independently learned brain representations alone.

In this paper, we extend the Strong Platonic Representation Hypothesis [6] to the human visual cortex. We ask whether subject-specific fMRI embedding spaces, learned independently from neural data,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

can be translated across subjects using only the intrinsic geometry of neural responses. We evaluate this setting on the Natural Scenes Dataset (NSD) [18], a canonical fMRI dataset of subjects viewing complex natural images. Our contributions are:

1. We introduce a self-supervised encoder that learns subject-specific fMRI embeddings from repeated stimulus presentations.
2. We show that independently learned subject embeddings can be translated across subjects using unsupervised orthogonal mappings, recovering accurate instance-level correspondences.
3. We synchronize pairwise rotations into a single shared latent space, improving cross-subject retrieval and showing that independently learned subject spaces are mutually compatible with a common coordinate system.

Our results provide evidence that independently learned fMRI embeddings share a geometric structure that can be recovered without paired cross-subject data, enabling neural models to transfer across individuals.

2 Problem Setup

Let $s \in \{1, \dots, S\}$ represent a subject who provides fMRI responses to visual stimuli drawn from a distribution $\mathcal{D}$. $X^{(s)} \in \mathbb{R}^{n_s \times v_s}$ denotes the subject's neural activity matrix, where rows correspond to image presentations and columns to subject-specific voxel responses. Our goal is to test whether these independently observed fMRI responses can be mapped into a shared latent space $\mathcal{Z}$ using only their intrinsic representational geometry, without paired cross-subject data for learning the mappings.

Subjects observe disjoint image sets independently sampled from $\mathcal{D}$. No shared images or paired cross-subject correspondences are available for learning the translations. A held-out set of shared images, observed by all subjects, is used only for evaluation.

Subject-specific embeddings. For each subject, we learn a mapping $f_s : X^{(s)} \to Z^{(s)}$, with $Z^{(s)} \in \mathbb{R}^{n_s \times d}$, that projects voxel responses into a low-dimensional subject-specific space $\mathcal{Z}^{(s)} \subset \mathbb{R}^d$. The mapping f_s is self-supervised from each subject's neural activity alone, without external model features or cross-subject supervision. Thus, $Z^{(s)}$ is intended to reflect the intrinsic organization of the stimulus responses of subject s.

Unpaired brain-to-brain translation. Given independently learned embeddings $\{Z^{(s)}\}_{s=1}^{S}$, we seek transformations that translate each subject space into a common latent space. Specifically, we learn one transformation $R_s : \mathcal{Z}^{(s)} \to \mathcal{Z}$ per subject, such that embeddings evoked by the same image map to consistent coordinates, i.e., $Z^{(s)}R_s \approx Z^{(t)}R_t$ for subjects s and t on held-out shared images. Since no such paired images are available during training, the transformations must be inferred from the global geometry of the unpaired subject-specific spaces. We evaluate translation quality using cross-subject retrieval on the held-out shared images, following prior unsupervised mapping protocols [6, 7].

Shared geometry hypothesis. We hypothesize that subject-specific spaces $\mathcal{Z}^{(s)}$ are noisy instances of a shared latent geometry. Under this hypothesis, inter-subject differences should be captured by approximately isometric transformations. We therefore restrict transformations to orthogonal maps, $R_s \in \mathcal{O}(d)$, which preserve distances and inner products. This constraint prevents arbitrary geometric warping and provides a strong test of whether independently learned neural representations can be translated into a shared coordinate system using geometry alone.

3 Method

We proceed in three stages. First, for each subject, we learn a subject-specific encoder that maps fMRI responses into a lower-dimensional embedding space. Second, we translate embeddings between subject pairs by estimating unsupervised orthogonal rotations from their geometry. Finally, we synchronize the pairwise rotations to recover one transformation per subject, mapping all embeddings into a shared latent space. An overview is shown in Fig. 1.

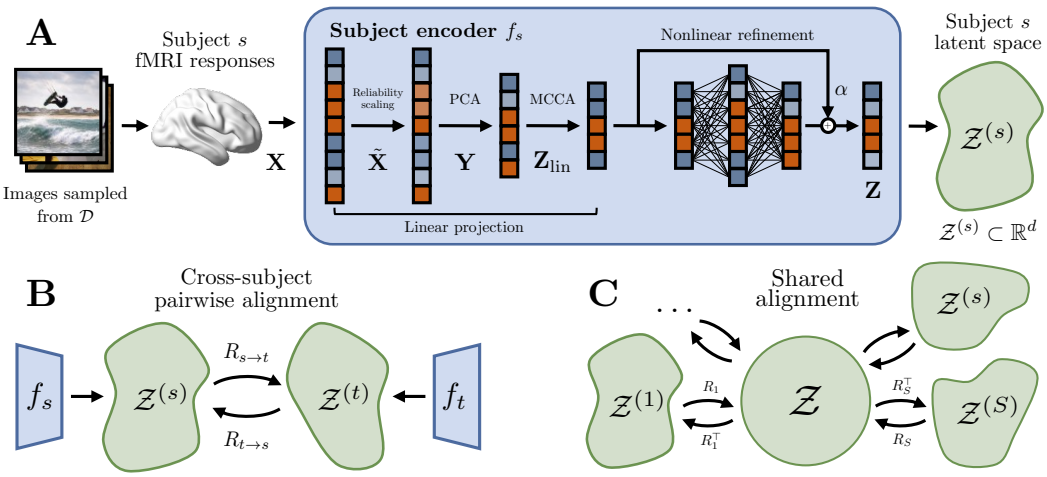


Figure 1: **Method overview. (A) Subject encoder.** For each subject, fMRI responses are mapped into a low-dimensional embedding space using voxel reliability weighting, PCA, and multi-view CCA (MCCA), followed by a residual nonlinear refinement trained from repeated stimulus presentations. **(B) Pairwise brain-to-brain translation.** Independently learned subject embeddings are translated between subject pairs by estimating orthogonal rotations $R_{s\to t}$ from geometry-derived pseudo-correspondences. **(C) Shared latent space.** Pairwise rotations are synchronized to recover one orthogonal transformation R_s per subject, mapping all subject embeddings into a common space.

3.1 Learning repetition-invariant embeddings

For each subject s, we learn a mapping $f_s : \mathbb{R}^{v_s} \to \mathbb{R}^d$ using repeated stimulus presentations as self-supervision. Each image is presented r times, yielding multiple measurements, or views, of the same underlying neural signal. For clarity, we drop the subject index (s) in this subsection. Let $\{X_i\}_{i=1}^r$, with $X_i \in \mathbb{R}^{n_s \times v_s}$, denote view-specific response matrices for the same stimuli, and let $X \in \mathbb{R}^{rn_s \times v_s}$ denote their concatenation. To mitigate temporal drifts, repetitions are randomly assigned to views for each stimulus.

Traditional fMRI denoising averages repeated measurements under independence assumptions [19–22]. In contrast, we treat repetitions as noisy views of a stable latent representation and optimize f_s for repetition invariance, such that for responses x_i and x_j to the same image, $f_s(x_i) \approx f_s(x_j)$.

Voxel reliability weighting. We first reweight voxels by their reliability across repetitions. For each voxel $v \in \{1, \dots, v_s\}$, we compute reliability as the average correlation across repetition pairs:

$$\gamma_v = \frac{2}{r(r-1)} \sum_{1 \le i < j \le r} \rho\big(X_i[:, v],\, X_j[:, v]\big). \tag{1}$$

We then scale each voxel by its reliability, $\tilde{X}_i[:, v] = \max(0, \gamma_v) \cdot X_i[:, v]$, reducing the influence of voxels without stable repetition structure.

Low-dimensional linear projection. We project reliability-weighted responses $\tilde{X}_i$ into a lower-dimensional subspace using PCA, obtaining $Y_i \in \mathbb{R}^{n_s \times d_{\text{PCA}}}$ with $d_{\text{PCA}} \ll v_s$. Let $Y \in \mathbb{R}^{rn_s \times d_{\text{PCA}}}$ denote the concatenation of all views.

To extract components shared across repetitions, we apply multi-view canonical correlation analysis (MCCA) to $\{Y_i\}_{i=1}^r$ [23, 24]. MCCA learns projections $\{U_i\}_{i=1}^r$, with $U_i \in \mathbb{R}^{d_{\text{PCA}} \times d}$, that maximize cross-view correlation:

$$\max_{\{U_i\}} \sum_{i<j} \rho\big(Y_i U_i,\; Y_j U_j\big). \tag{2}$$

To obtain a single target representation, we project each sample in Y through all view-specific mappings and average the projections:

$$\bar{Z}_{\text{lin}} = \frac{1}{r} \sum_{i=1}^{r} YU_i. \tag{3}$$

We then regularize this mapping using ridge regression:

$$W^* = \arg\min_W \ \|YW - \bar{Z}_{\text{lin}}\|_F^2 + \lambda_{\text{reg}} \|W\|_F^2, \tag{4}$$

yielding $Z_{\text{lin}} = YW^* \in \mathbb{R}^{rn_s \times d}$.

Nonlinear residual refinement. We refine the linear embedding with a residual nonlinearity:

$$Z = f_s(X) = Z_{\text{lin}} + \alpha \, g_\theta(Z_{\text{lin}}), \tag{5}$$

where g_θ is a multi-layer perceptron (MLP) and α is a learnable scalar. We freeze the linear projection and optimize only θ and α. Given view embeddings $\{Z_i\}_{i=1}^r$, where $Z_i = f_s(X_i)$, we use a contrastive InfoNCE loss over all view pairs:

$$\mathcal{L}_{\text{NCE}} = \frac{1}{r(r-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{r} \mathcal{L}_{\text{InfoNCE}}(Z_i, Z_j), \tag{6}$$

with in-batch negatives and cosine similarity [25]. We also add a cosine pull term,

$$\mathcal{L}_{\text{pull}} = \frac{2}{r(r-1)} \sum_{1 \leq i < j \leq r} \big(1 - \text{sim}(Z_i, Z_j)\big), \tag{7}$$

where $\text{sim}(Z_i, Z_j)$ is the mean cosine similarity between corresponding samples. The refinement minimizes $(\theta^*, \alpha^*) = \arg\min_{\theta,\alpha} \mathcal{L}_{\text{NCE}} + \lambda_{\text{pull}} \mathcal{L}_{\text{pull}}$.

3.2 Pairwise brain-to-brain translation

To translate embeddings between two subjects, we apply *mini-vec2vec* [7] to neural embeddings. Given subjects s and t, we learn an orthogonal transformation $R_{s \to t} \in \mathcal{O}(d)$ such that $Z^{(s)} R_{s \to t} \approx Z^{(t)}$, without paired cross-subject samples during training.

We average repetitions in embedding space to obtain one representation per image, $Z^{(s)} \in \mathbb{R}^{n_s \times d}$ and $Z^{(t)} \in \mathbb{R}^{n_t \times d}$, reducing trial-level measurement error. We construct pseudo-matched pairs by clustering each space with K-means and matching centroids through their pairwise similarity structure using a quadratic assignment solver. Each embedding in $Z^{(s)}$ is then matched to the average of its nearest neighbors in $Z^{(t)}$ based on relative similarity to these matched anchors, yielding pseudo-parallel pairs $(z^{(s)}, \tilde{z}^{(t)})$. These pairs define the initial orthogonal Procrustes problem:

$$R_{s \to t}^{(0)} = \arg\min_{R \in \mathcal{O}(d)} \|Z^{(s)} R - \tilde{Z}^{(t)}\|_F^2. \tag{8}$$

We refine the translation iteratively using an approach similar to Iterative Closest Point [26]. At iteration k, a subset of source embeddings is transformed by $R_{s \to t}^{(k)}$ and matched to nearest neighbors in the target space. These updated pseudo-targets define a new Procrustes solution $\widehat{R}_{s \to t}^{(k)}$, and the transformation is updated as

$$\tilde{R}_{s \to t}^{(k+1)} = (1 - \beta) R_{s \to t}^{(k)} + \beta \widehat{R}_{s \to t}^{(k)}. \tag{9}$$

After each update, $\tilde{R}_{s \to t}^{(k+1)}$ is projected onto $\mathcal{O}(d)$ by SVD to obtain $R_{s \to t}^{(k+1)}$.

Because the learned mappings are sensitive to clustering initialization, we generate multiple candidate rotations per directed subject pair using different random seeds. We jointly select one candidate per pair by maximizing cycle consistency across subjects: composing the mappings from s to t and from t to u should approximate the direct mapping from s to u. Selection is fully unsupervised and uses

only the candidate rotation matrices, without accessing image correspondences or evaluation scores. Appendix C provides the selection algorithm and its stability analysis.

Finally, we symmetrize the selected pairwise translations to enforce $R_{t\to s} = R_{s\to t}^\top$:

$$R_{s\to t} = \mathrm{Proj}_{\mathcal{O}(d)}\left(\frac{R_{s\to t} + R_{t\to s}^\top}{2}\right). \tag{10}$$

3.3 Shared latent space construction

Given pairwise translations $\{R_{s\to t}\}_{s,t=1}^S$, we construct a shared latent space $\mathcal{Z}$ by solving an orthogonal synchronization problem over $\mathcal{O}(d)$ [27, 28]. The goal is to recover one transformation $R_s \in \mathcal{O}(d)$ per subject such that $R_{s\to t} \approx R_s R_t^\top$, placing all subjects in a common coordinate system, i.e., $Z^{(s)}R_s \approx Z^{(t)}R_t$.

We form a block matrix $B \in \mathbb{R}^{Sd\times Sd}$ whose (s,t)-th block is $R_{s\to t}$ for $s \neq t$ and the identity for $s = t$. In the ideal noise-free case, this matrix factorizes as

$$B = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_S \end{bmatrix} \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_S \end{bmatrix}^\top. \tag{11}$$

We recover a relaxed solution using a spectral method for orthogonal synchronization [27]. We compute the top-d eigenvectors of B, yielding $U \in \mathbb{R}^{Sd\times d}$, whose blocks $\{U_s\}_{s=1}^S$ approximate the subject-specific transformations. Each block U_s is projected onto $\mathcal{O}(d)$ by taking its closest orthogonal matrix.

The resulting transformations define a shared latent space in which each subject embedding is mapped as $Z_{\text{shared}}^{(s)} = Z^{(s)}R_s$. Global synchronization denoises pairwise estimates by enforcing cycle consistency across the subject graph.

4 Experiments

We evaluate our method on the Natural Scenes Dataset (NSD) [18], an fMRI dataset of 8 participants viewing natural images from COCO [29]. Each participant viewed up to 10,000 distinct images, each repeated up to three times. NSD includes subject-specific images, unique to each participant, and a smaller set of images shared across participants. We use only subject-specific, non-shared images to learn subject encoders and unsupervised brain-to-brain translations. Evaluation uses 515 held-out shared images with three repetitions for every subject. Supervised brain-to-brain baselines use additional shared images for alignment training, excluding these 515 evaluation images. See Appendix A for preprocessing details.

We evaluate retrieval in two settings. Within subjects, embeddings from one repetition retrieve the matching image from another repetition among 515 candidates, averaged across repetition pairs. Across subjects, repetitions are first averaged, and translated embeddings from subject s retrieve the matching image from subject t among the same 515 held-out images. In both settings, we report Mean Rank, R@1, and RSA. Mean Rank is the average rank of the correct image (chance $= 258$, optimum $= 1$); R@1 is nearest-neighbor accuracy (chance $= 1/515 \approx 0.002$); RSA is the Pearson correlation between representational dissimilarity matrices.

4.1 Within-subject encoder evaluation

We first evaluate whether the subject-specific encoder extracts stable stimulus representations from noisy repeated fMRI responses, using the within-subject retrieval protocol defined above. Encoder ablations and hyperparameter configurations are provided in Appendix B.

Performance across subjects. Table 1 reports full-encoder performance for each subject. In general, our embeddings are stable: with an average rank of 5.28, within-subject retrieval works well

Table 1: **Per-subject encoder performance.** Within-subject retrieval across repeated presentations, averaged across repetition pairs. Embeddings from one repetition are used to retrieve the matching image from another among 515 candidates. RSA is the Pearson correlation between RDMs across repetitions. Lower is better for Mean Rank (chance $= 258$); higher is better for R@1 (chance $= 0.002$) and RSA.

Method	Metric	S1	S2	S3	S4	S5	S6	S7	S8	Avg
Ours (full)	Mean Rank ↓	1.03	1.12	11.25	4.69	1.77	3.09	7.60	11.67	5.28
	R@1 ↑	0.99	0.98	0.75	0.79	0.87	0.83	0.72	0.62	0.82
	RSA ↑	0.60	0.61	0.47	0.51	0.60	0.52	0.50	0.43	0.53

across subjects. S1 and S2 are nearly perfectly matched across repetitions (Mean Rank ≈ 1; R@1 > 0.98), whereas S3 and S8 show lower retrieval performance. Even for higher-performing subjects, RSA remains around 0.6, indicating that robust instance-level identification does not require exact preservation of the full pairwise geometry.

Comparison with encoder baselines. Table 2 compares our encoder against three baseline families. First, we include direct neural baselines using preprocessed GLMsingle [20] responses and PCA-reduced fMRI responses. Second, we compare against multiview methods trained self-supervised from repeated presentations. Third, we include model-guided baselines, where fMRI responses are linearly regressed to pretrained model embeddings. Direct neural baselines provide limited retrieval, while CCA-based multiview methods improve instance-level matching but yield low RSA. Model-guided baselines preserve stronger geometry, but are weaker at retrieving repeated presentations of the same image. Our full encoder achieves the lowest Mean Rank and highest R@1, with RSA comparable to the strongest model-guided baselines. Thus, repetition-based self-supervision recovers stronger instance-level image information than externally guided encoders, likely because it directly optimizes invariance across neural measurements of the same stimulus rather than fitting an intermediate model space. See Appendix D for implementation details and per-subject results.

Table 2: **Within-subject encoder comparison.** Retrieval and RSA across repeated presentations, averaged across subjects and repetition pairs. Mean Rank and R@1 measure instance-level stability; RSA measures geometry preservation across repetitions. Lower is better for Mean Rank (chance $= 258$); higher is better for R@1 (chance $= 0.002$) and RSA.

Method	Setting	Mean Rank ↓	R@1 ↑	RSA ↑
Preprocessed fMRI (GLMsingle [20])	GLM (repetitions)	144.15	0.07	0.18
PCA	Unsupervised	78.10	0.19	0.27
PCA + SplitAE [30]	SSL Repetitions	93.78	0.07	0.29
PCA + DeepCCA [31]	SSL Repetitions	23.21	0.52	0.19
PCA + Multiview CCA [24]	SSL Repetitions	19.23	0.60	0.21
DINOv2 ViT-S/14 [32] + Ridge	Vision Model Guided	24.82	0.30	0.46
CLIP ViT-B/16 [33] + Ridge	Vision Model Guided	23.23	0.26	0.50
ViT AugReg-L/16 [34] + Ridge	Vision Model Guided	20.30	0.32	0.53
all-MiniLM-L6-v2 [35] + Ridge	Language Model Guided	30.47	0.18	**0.56**
Ours (full)	SSL Repetitions	**5.28**	**0.82**	0.53

4.2 Pairwise brain-to-brain translation

We evaluate brain-to-brain translation through cross-subject image retrieval and representational similarity. Our method learns orthogonal mappings between independently trained subject embeddings using unpaired training images (Subsection 3.2). To address initialization sensitivity, we generate 10 candidate rotations per directed subject pair and jointly select one per pair by maximizing cycle consistency across subjects, using the unsupervised criterion detailed in Appendix C.

Fig. 2 reports retrieval and RSA for each directed subject pair. The selected mappings recover accurate cross-subject correspondences, with nearly perfect retrieval between S1 and S2. Performance varies across pairs, with lower retrieval between these subjects and S5.

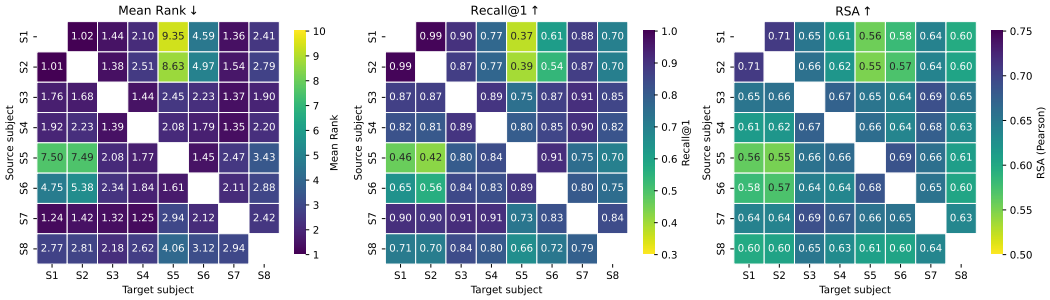


Figure 2: **Pairwise brain-to-brain translation.** Performance for each ordered subject pair (s, t). Embeddings from source subject s are mapped into target subject t's space using the unsupervised orthogonal transformation $R_{s \to t}$ and evaluated on the 515 held-out shared images. Mean Rank (average $\pm$ std: 2.70 ± 1.84, chance $= 258$) and R@1 (average $\pm$ std: 0.78 ± 0.14, chance $= 0.002$) measure image-level retrieval after translation. RSA is the Pearson correlation between subject-specific RDMs (average $\pm$ std: 0.63 ± 0.04). Darker colors indicate better performance.

Table 3: **Pairwise brain-to-brain translation baselines.** Performance is reported as the average across ordered off-diagonal subject pairs on the 515 held-out shared images. Rows compare no-translation controls, optimal-transport matching, external model-space references, and our unpaired orthogonal translation. Lower is better for Mean Rank (chance $= 258$); higher is better for R@1 (chance $= 0.002$) and RSA.

Representation	**Alignment**	**Mean Rank ↓**	**R@1 ↑**	**RSA ↑**
No cross-subject alignment				
fMRI betas + PCA	No alignment	189.28	0.007	0.39
Optimal transport matching on evaluation embeddings (transductive)				
fMRI betas + PCA	Entropic GW	166.04	0.011	0.39
Our embeddings	Entropic GW	80.96	0.465	0.64
External model-defined shared spaces (unpaired training images)				
ViT AugReg-L/16 [34] + Ridge	Vision model guided	5.54	0.480	0.73
CLIP ViT-B/16 [33] + Ridge	Vision model guided	6.60	0.440	0.70
DINOv2 ViT-S/14 [32] + Ridge	Vision model guided	7.00	0.470	0.67
all-MiniLM-L6-v2 [35] + Ridge	Language model guided	7.93	0.350	0.76
Supervised alignment with paired training images				
Our embeddings	Ridge regression (CV)	1.64	0.886	0.67
Our embeddings	Orthogonal Procrustes	1.68	0.882	0.63
Our embeddings	CCA	1.70	0.882	0.62
Our embeddings	Linear regression	3.15	0.772	0.58
Unsupervised alignment with unpaired training images				
Our embeddings	Orthogonal Procrustes	2.70	0.780	0.63

Baseline comparisons. We compare methods with different sources of alignment information (Table 3). The no-alignment control compares PCA-reduced fMRI responses directly across subjects. Entropic Gromov–Wasserstein (GW) matching estimates correspondences from within-subject relational structure, using the complete evaluation embedding sets without their identity correspondences. Model-guided baselines use ridge regression to map each subject's training responses into a pre-trained vision or language model's feature space. They require no shared training images across subjects but rely on an external model to define a common coordinate system. Supervised Procrustes, cross-validated ridge regression, CCA, and linear regression quantify the benefit of paired training images. They use the same neural embeddings and an additional set of shared images, up to 485 per subject pair, while retaining the 515-image evaluation set. The encoders remain unchanged. Appendix D.2 provides implementation details for all baselines.

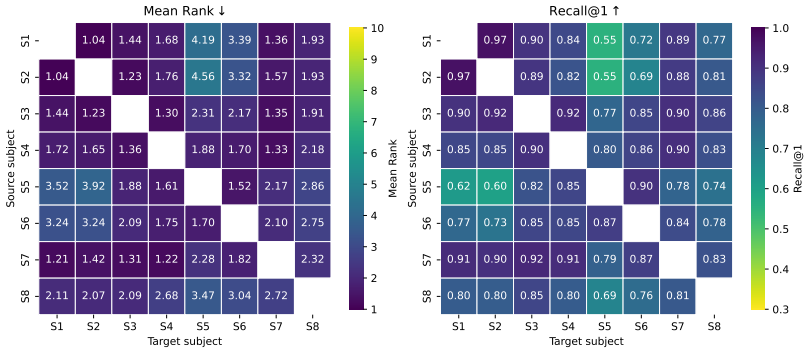


Figure 3: **Shared-space brain-to-brain translation.** Retrieval performance after synchronizing pairwise brain-to-brain rotations into a single shared latent space. Each subject is mapped into the common coordinate system using one orthogonal transformation R_s, and retrieval is evaluated across ordered subject pairs on the 515 held-out shared images. Left: Mean Rank (average $\pm$ std: 2.11 ± 0.84; chance $= 258$). Right: R@1, higher is better (average $\pm$ std: 0.82 ± 0.09; chance $= 0.002$). Darker colors indicate better performance.

Our unsupervised mappings outperform the no-alignment, GW, and model-guided baselines on retrieval, although model-guided spaces achieve higher RSA (Table 3). Supervised Procrustes, ridge, and CCA further improve retrieval, with ridge achieving the strongest pairwise results. This advantage requires paired cross-subject training images that our unsupervised alignment does not use.

4.3 A Platonic brain-to-brain translation layer

Pairwise translations map individual subject pairs, but do not guarantee a coherent coordinate system across subjects. We therefore synchronize the pairwise rotations to recover one orthogonal transformation R_s per subject (Subsection 3.3), mapping all embeddings into a common latent space (Fig. 1C).

Synchronization improves average Mean Rank from 2.70 to 2.11 and R@1 from 0.78 to 0.82 across the 56 directed subject pairs (Fig. 3). Synchronized retrieval exceeded chance in image-identity permutation tests ($p < 10^{-4}$), and remained strong after excluding each subject's evaluation pairs. Appendix E reports these analyses and image-bootstrap confidence intervals. Appendix F illustrates successful retrievals, near misses, and failures across subject pairs.

The improvement after synchronization suggests that enforcing consistency across the subject graph reduces noise in pairwise estimates. One orthogonal transformation per subject is sufficient to recover accurate cross-subject correspondence in a common coordinate system, providing evidence consistent with the Strong Platonic Representation Hypothesis within the evaluated visual domain.

4.4 Model–brain alignment

We next test whether the final-layer representations of four artificial neural networks predict our fMRI embeddings. For each subject, we fit supervised mappings from model to neural embeddings using the same images used to train the subject encoder. We compare semi-orthogonal mappings, which accommodate differences in embedding dimensionality, with ridge regression. We then map model embeddings of the 515 held-out shared images into the subject's neural embedding space and evaluate retrieval and RSA against the corresponding brain embeddings.

Table 4 shows that semi-orthogonal mappings recover moderate instance-level correspondence (best Mean Rank $= 13.33$, R@1 $= 0.29$). Ridge regression improves retrieval and RSA (best Mean Rank $= 5.86$, R@1 $= 0.51$), indicating that model features predict neural embeddings more accurately when more flexible linear transformations are allowed. However, these supervised model-to-brain mappings remain below the shared-space brain-to-brain retrieval results (Mean Rank $= 2.11$, R@1 $= 0.82$).

Table 4: **Supervised model-to-brain alignment.** Final-layer model embeddings are mapped to neural embeddings using each subject's encoder training images and evaluated on the 515 held-out shared images. We compare semi-orthogonal mappings with ridge regression. Values are averaged across subjects. Lower is better for Mean Rank; higher is better for R@1 and RSA.

Source model	Semi-orthogonal mapping			Ridge regression		
	Mean Rank ↓	R@1 ↑	RSA ↑	Mean Rank ↓	R@1 ↑	RSA ↑
all-MiniLM-L6-v2	35.62	0.08	0.31	27.05	0.11	0.45
CLIP ViT-B/16	21.65	0.23	0.23	7.86	0.44	0.58
ViT AugReg-L/16	**13.33**	**0.29**	**0.41**	8.85	0.41	0.57
DINOv2 ViT-S/14	22.71	0.24	0.24	**5.86**	**0.51**	**0.57**

4.5 Cross-subject fMRI reconstruction

Finally, we test whether aligned embeddings preserve information for predicting another subject's voxel responses. For each target subject t, we fit a ridge decoder $D_t(z) = zW_t + b_t$ from its shared-space training embeddings $Z^{(t)}_{\text{train}} R_t$ to its measured voxel responses. We then apply this decoder, without refitting, to another subject's aligned test embeddings:

$$\widehat{X}_{s \to t} = D_t\left(Z^{(s)}_{\text{test}} R_s\right). \tag{12}$$

The rotations are learned unsupervised, and the 515 shared evaluation images are excluded from decoder training.

We train target decoders for S1, S2, S5, and S7, who completed all NSD sessions, and use all eight subjects as embedding sources. Embeddings and voxel responses are averaged across repetitions. We evaluate reconstruction using Pearson correlation across the evaluation images for each voxel, then average correlations across target voxels.

Across the 28 cross-subject evaluations, mean correlation (average $\pm$ SD) was 0.350 ± 0.038, compared with 0.431 ± 0.046 for within-subject reconstruction (Table 5). Cross-subject prediction therefore retained approximately 81% of the within-subject correlation, showing that the shared embeddings contain transferable information about measured neural responses.

Table 5: **Cross-subject fMRI reconstruction.** Mean voxel-wise Pearson correlations between predicted and measured responses across 515 held-out images. Rows denote target subjects and columns denote embedding sources. Bold diagonal entries report within-subject reconstruction. Each target decoder is trained only on that subject's training data and applied without refitting.

Target / Source	S1	S2	S3	S4	S5	S6	S7	S8
S1	**0.433**	0.364	0.350	0.340	0.328	0.334	0.348	0.341
S2	0.372	**0.428**	0.356	0.348	0.336	0.340	0.349	0.346
S5	0.388	0.390	0.410	0.417	**0.489**	0.416	0.406	0.399
S7	0.309	0.309	0.306	0.305	0.298	0.301	**0.376**	0.295

5 Related work

Universal geometry in machine learning. Learned representations in deep neural networks exhibit structured geometries that reflect statistical properties of the data. Empirical work shows that models trained with different architectures, objectives, and modalities often develop similar representational spaces [1–3, 36]. These observations motivate the *Platonic Representation Hypothesis* [5], which proposes that models converge toward a shared latent structure as scale and data coverage increase. This convergence has been studied through representational similarity measures [3] and constructive methods such as model stitching and latent-space mapping [8, 37–39]. Recent work tests a stronger version of this hypothesis by showing that embedding spaces can be translated from intrinsic geometry alone, without shared inputs or paired supervision [6, 7, 40]. In this work, we ask whether this constructive form—the Strong Platonic Representation Hypothesis—also holds for biological neural representations.

Shared representations in neuroscience. Shared neural representations are harder to recover because brain data are noisy, limited, and subject to anatomical and functional variability. Nevertheless, inter-subject synchrony [10, 11] and representational similarity analyses [12, 13] show that neural response structure is partly preserved across subjects [41–43]. Functional alignment methods such as hyperalignment map subjects into common spaces, but typically require shared stimuli to establish correspondences [14, 15]. Other methods introduce anchors through model-derived feature spaces [16] or image-to-fMRI encoders [17], while unsupervised alternatives infer correspondences through structural or distributional matching [44] and optimal transport-based functional alignment [45, 46]. In contrast, we learn subject-specific fMRI embeddings from neural repetitions and test whether they can be translated by orthogonal rotations learned from unpaired subject-specific images.

Model–brain alignment. A complementary literature aligns neural activity with deep neural network representations, revealing systematic correspondences between cortical processing stages and model layers [41, 47–49]. These approaches support encoding, decoding, and benchmarking efforts such as *Brain-Score* [50–55]. Recent work further suggests that brains and models may share low-dimensional representational axes [56]. However, model-mediated approaches impose or evaluate neural representations through an external feature geometry. Our method instead recovers a shared brain space directly from neural data, using model–brain mappings only diagnostically.

6 Discussion

The Strong Platonic Representation Hypothesis [6] posits that the universal latent structure of representations not only exists but can be harnessed to translate representations from one space to another without any paired data or model access. Whereas the original hypothesis was conjectured over artificial embeddings, in this work, we evaluated this prediction in embeddings learned from human fMRI responses to natural images. We found that subject-specific embeddings learned from brain data alone can be translated across individuals using unsupervised orthogonal rotations, without paired cross-subject samples or intermediate model representations. Synchronizing these rotations into a shared latent space further improved retrieval, indicating that the pairwise translations are mutually compatible with a common coordinate system. Together, these results suggest that independently learned subject embeddings share a geometric structure that supports cross-subject translation within the evaluated visual domain.

This shared representational structure has practical implications for cross-subject neural modeling. In our fMRI reconstruction experiment, a target-subject ridge decoder predicts voxel responses from another subject's aligned embeddings without refitting (Table 5). Such transfer could reduce subject-specific calibration for image-to-fMRI encoding, fMRI-to-image decoding, and synthetic neural data generation. Shared spaces could also support training neural models on heterogeneous datasets in which subjects viewed different stimuli.

Our model–brain results further distinguish linear predictability from geometric alignment. Model features predict neural embeddings under supervised linear mappings, with ridge regression outperforming semi-orthogonal mappings. This suggests that current models capture stimulus information relevant to neural responses, while more flexible transformations help align their representational spaces. Brain-derived shared spaces may therefore provide a useful target for investigating which aspects of representation are shared or differ between artificial models and neural embeddings.

Several limitations remain. Our geometric conclusions concern embeddings learned with a common encoder architecture and training objective, whose contributions to the observed alignment remain to be disentangled. Establishing approximate isometry would also require direct tests of distance preservation beyond retrieval performance. The analysis relies on high-quality fMRI data with repeated stimulus presentations to isolate stable stimulus-related signal from trial-level noise. It remains unclear whether comparable alignment will hold in lower-SNR modalities, smaller datasets, or higher-level cognitive tasks. Future work should test whether shared representational structures extend to more distant stimulus distributions, task contexts, and clinical datasets. Finally, subject-agnostic neural spaces may improve decoding and transfer learning, but related methods raise neural privacy concerns if applied to sensitive brain data without consent; future use should require explicit consent, careful data governance, and safeguards against misuse.

Acknowledgments and Disclosure of Funding

This work was supported by the Spanish Ministerio de Ciencia, Innovación y Universidades through the Agencia Estatal de Investigación (AEI), under project PID2022-140426NB-I00, co-funded by the European Regional Development Fund (ERDF, "A way to build Europe"). We thank the CERCA Programme/Generalitat de Catalunya for institutional support. R.J. was supported by the Digital Life Initiative Doctoral Fellowship at Cornell Tech.

References

[1] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751. Association for Computational Linguistics, 2013.

[2] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6078–6087, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[3] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, volume 97, pages 3519–3529, 2019.

[4] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine Hermann, Kerem Oktar, Klaus Greff, Martin N Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus Robert Muller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=Hiq7lUh4Yn`.

[5] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024.

[6] Rishi Dev Jha, Collin Zhang, Vitaly Shmatikov, and John Xavier Morris. Harnessing the universal geometry of embeddings. In *Advances in Neural Information Processing Systems*, 2025.

[7] Guy Dar. mini-vec2vec: Scaling universal geometry alignment with linear transformations, 2026. URL `https://arxiv.org/abs/2510.02348`.

[8] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=ak06J5jNR4`.

[9] Jianglin Lu, Hailing Wang, Yi Xu, Yizhou Wang, Kuo Yang, and Yun Fu. Representation potentials of foundation models for multimodal alignment: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16669–16684. Association for Computational Linguistics, 2025. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.843.

[10] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664):1634–1640, 2004. doi: 10.1126/science.1089506.

[11] Samuel A Nastase, Valeria Gazzola, Uri Hasson, and Christian Keysers. Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, 14(6):667–685, 2019. doi: 10.1093/scan/nsz037.

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

[13] Baihan Lin and Nikolaus Kriegeskorte. The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, 121(42):e2317881121, 2024. doi: 10.1073/pnas.2317881121. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2317881121`.

[14] James V. Haxby, J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, M. Ida Gobbini, Michael Hanke, and Peter J. Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011. doi: 10.1016/j.neuron.2011.08.026.

[15] J. Swaroop Guntupalli, Michael Hanke, Yaroslav O. Halchenko, Andrew C. Connolly, Peter J. Ramadge, and James V. Haxby. A model of representational spaces in human cortex. *Cerebral Cortex*, 26(6): 2919–2934, 2016. doi: 10.1093/cercor/bhw068.

[16] Qiongyi Zhou, Changde Du, Shengpei Wang, and Huiguang He. CLIP-MUSED: CLIP-guided multi-subject visual neural information semantic decoding. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lKxL5zkssv.

[17] Navve Wasserman, Roman Beliy, Roy Urbach, and Michal Irani. Functional brain-to-brain transformation without shared stimuli. *NeuroImage*, 327:121741, 2026. doi: 10.1016/j.neuroimage.2026.121741.

[18] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962-x.

[19] Thomas T. Liu. Noise contributions to the fMRI signal: An overview. *NeuroImage*, 143:141–151, 2016. doi: 10.1016/j.neuroimage.2016.09.008.

[20] Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick N. Kay. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, 11, 2022. doi: 10.7554/eLife.77599.

[21] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, Sep 2023. doi: 10.1038/s41598-023-42891-8.

[22] Paul Steven Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Cohen Ethan, Aidan James Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=rwrblCYb2A.

[23] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971. doi: 10.1093/biomet/58.3.433.

[24] Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, 2011.

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

[26] Paul J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2), 1992. doi: 10.1109/34.121791.

[27] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36, 2011. doi: 10.1016/j.acha.2010.02.001.

[28] Lanhui Wang and Amit Singer. Exact and stable recovery of rotations for robust synchronization. *Information and Inference: A Journal of the IMA*, 2:145–193, 10 2013. doi: 10.1093/imaiai/iat005.

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[30] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1083–1092, 2015. URL https://proceedings.mlr.press/v37/wangb15.html.

[31] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1247–1255, 2013. URL https://proceedings.mlr.press/v28/andrew13.html.

[32] 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. URL `https://openreview.net/forum?id=a68SUt6zFt`.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[34] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. URL `https://openreview.net/forum?id=4nPswr1KcP`.

[35] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

[36] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings of Machine Learning Research*, pages 196–212. PMLR, 2015.

[37] David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1214.

[38] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with Wasserstein Procrustes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1880–1890, 2019.

[39] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=SrC-nwieGJ`.

[40] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=H196sainb`.

[41] Pablo Marcos-Manchón and Lluís Fuentemilla. Shared representations in brains and models reveal a two-route cortical organization during scene perception. *Communications Biology*, May 2026. doi: 10.1038/s42003-026-10169-0.

[42] Janice Chen, Yuan Chang Leong, Christopher J. Honey, Chung H. Yong, Kenneth A. Norman, and Uri Hasson. Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, 20(1):115–125, 2017. doi: 10.1038/nn.4450.

[43] Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):1–33, 2017. doi: 10.1371/journal.pcbi.1005508.

[44] Daiki Nakamura, Shizuo Kaji, Ryota Kanai, and Ryusuke Hayashi. Unsupervised method for representation transfer from one brain to another. *Frontiers in Neuroinformatics*, Volume 18 - 2024, 2024. doi: 10.3389/fninf.2024.1470845.

[45] T. Bazeille, H. Richard, H. Janati, and B. Thirion. Local optimal transport for functional brain template estimation. In Albert C. S. Chung, James C. Gee, Paul A. Yushkevich, and Siqi Bao, editors, *Information Processing in Medical Imaging*, pages 237–248. Springer International Publishing, 2019.

[46] Thomas Bazeille, Elizabeth DuPre, Hugo Richard, Jean-Baptiste Poline, and Bertrand Thirion. An empirical evaluation of functional alignment using inter-subject decoding. *NeuroImage*, 245:118683, 2021. doi: 10.1016/j.neuroimage.2021.118683.

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

[48] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Computational Biology*, 10(11):1–29, 2014. doi: 10.1371/journal.pcbi.1003915.

[49] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, 2016. doi: 10.1038/srep27755.

[50] Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020. doi: 10.1016/j.neuron.2020.07.040.

[51] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022. doi: 10.1038/s42003-022-03036-1.

[52] Talia Konkle and George A. Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, 13(1):491, 2022. doi: 10.1038/s41467-022-28091-4.

[53] Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, 2023. doi: 10.1038/s42256-023-00753-y.

[54] Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, pages 1220–1234, 2025. doi: 10.1038/s42256-025-01072-0.

[55] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, 2024. doi: 10.1038/s41467-024-53147-y.

[56] Zirui Chen and Michael F. Bonner. Universal dimensions of visual representation. *Science Advances*, 11 (27):eadw7697, 2025. doi: 10.1126/sciadv.adw7697.

[57] Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. doi: 10.1038/nature18933.

[58] Liang Wang, Ryan E.B. Mruczek, Michael J. Arcaro, and Sabine Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, 25(10):3911–3931, 2015. doi: 10.1093/cercor/bhu277.

[59] Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015. doi: 10.1007/s10107-015-0892-3. URL `https://doi.org/10.1007/s10107-015-0892-3`.

[60] Ross Wightman. PyTorch image models, 2019.

[61] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

[62] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024. URL `https://arxiv.org/abs/2212.03533`.

[63] Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, Kate Soule, Arafat Sultan, and Radu Florian. Granite embedding models, 2025. URL `https://arxiv.org/abs/2502.20204`.

[64] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, December 2022. doi: 10.18653/v1/2022.emnlp-main.669.

[65] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Appendix

A fMRI preprocessing details

For all main analyses, we use the Natural Scenes Dataset (NSD) [18] volumetric 1 mm GLM beta estimates[20], version 3, as provided by the dataset authors[1]. For each subject, we extract voxels from the `nsdgeneral` ROI, a reliability-based mask that includes visually responsive voxels across visual and high-level associative cortex. This yields approximately 85,000 voxels per subject. The resulting responses are organized into a trial-by-voxel matrix $X \in \mathbb{R}^{n_{\text{trials}} \times v_s}$ for each subject s.

Before fitting the encoder, we apply a simple normalization procedure to reduce the effect of extreme beta values. For each subject, voxel values are clipped at the 0.05th and 99.95th percentiles and stored as the input response matrix for subsequent preprocessing. We additionally remove acquisition-related confounds by fitting a linear model of session/run structure on the training trials and subtracting the predicted confound component from all responses. All train/test splits are defined at the stimulus level before fitting any preprocessing transformations.

A.1 Confound removal

NSD data were acquired over multiple scanning sessions, with each subject completing up to 40 sessions and multiple runs per session. We observed that trial-by-trial correlation matrices contained visible block structure aligned with run/session order, indicating acquisition-related components in the raw fMRI responses (Fig. S1A–B). To reduce this structure, we residualized each subject's voxel responses with respect to run/session confounds.

Let $X \in \mathbb{R}^{n_{\text{trials}} \times v_s}$ denote the preprocessed beta matrix and let $C \in \mathbb{R}^{n_{\text{trials}} \times c}$ denote a confound design matrix encoding the run/session membership of each trial. We fit a ridge-regularized linear confound model using only training trials:

$$B^* = \arg\min_B \|X_{\text{train}} - C_{\text{train}} B\|_F^2 + \lambda_{\text{conf}} \|B\|_F^2.$$

The fitted confound coefficients are then applied to all trials, and residualized responses are computed as

$$X_{\text{res}} = X - CB^*.$$

Thus, nuisance parameters are estimated exclusively from the training set and then applied to both training and held-out trials, avoiding leakage from the evaluation responses into the fitted residualization model.

Residualization partially mitigates the run/session block structure visible in trial-wise correlation matrices (Fig. S1). All experiments and baselines in the main text use these residualized fMRI responses as input.

[1] `https://registry.opendata.aws/nsd/` (accessed May 2026).

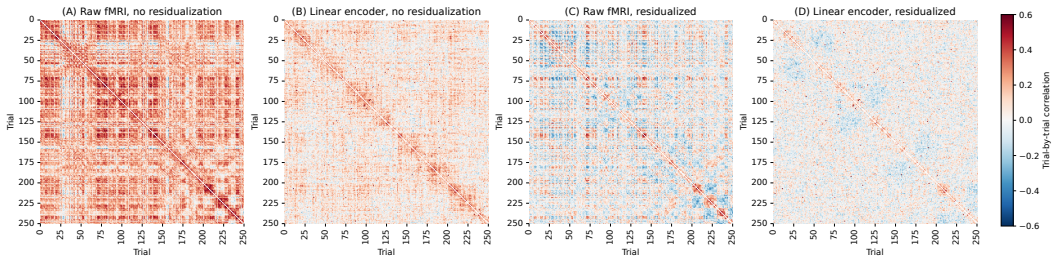


Figure S1: **Residualization mitigates acquisition-related trial structure.** Trial-by-trial correlation matrices for the first 250 trials from the first session of subject S1, shown for raw fMRI responses and linear encoder embeddings before and after residualizing run/session confounds. Trials are ordered by acquisition order; therefore, block-like off-diagonal structure reflects correlations due to scanning confounds rather than stimulus identity. Raw fMRI responses show pronounced run/session structure before residualization, which is partially mitigated after confound regression. The linear encoder further attenuates acquisition-related correlations. Diagonal entries are omitted for visualization.

A.2 Brain region selection

For the main analyses, we extract neural responses from the NSDGeneral ROI provided by the NSD authors [18]. This mask includes visually responsive voxels across visual and high-level associative cortex selected based on stimulus-related reliability. We use the volumetric preprocessing for the main experiments.

Table S1 evaluates the effect of cortical region and preprocessing space using the same encoder configuration as in the main analyses (see Section B). Encoder quality is measured with within-subject retrieval across repeated presentations. Performance improves as the cortical mask expands from V1 to V1–V4 and NSDGeneral, with NSDGeneral and Visual Streams yielding the best retrieval and RSA. Volumetric and surface-based NSDGeneral preprocessing give similar performance, so we use volumetric NSDGeneral responses for all main reported analyses.

Table S1: **Effect of brain region and preprocessing on encoder performance.** Within-subject retrieval across repeated presentations for different cortical regions (V1, V1–V4, NSDGeneral, Kastner, and visual streams; combining both hemispheres) and preprocessing spaces (volumetric and surface-based). Results are averaged across subjects and repetition pairs. Lower is better for Mean Rank (chance $= 258$); higher is better for R@1 (chance=0.002) and RSA.

Brain region	Preprocessing	Region size	Mean Rank ↓	R@1 ↑	RSA ↑
V1 [57]	Volumetric 1mm	19,820 voxels	23.42	0.54	0.37
V1–V4 [57]	Volumetric 1mm	51,197 voxels	8.20	0.74	0.47
Kastner2015 [58]	Volumetric 1mm	56,743 voxels	6.05	0.80	0.51
NSDGeneral [18]	Volumetric 1mm	84,848 voxels	5.28	0.82	0.53
Streams [18]	Volumetric 1mm	152,370 voxels	5.18	0.82	0.53
NSDGeneral [18]	Native surface	68,464 vertices	5.59	0.81	0.52

B Encoder details and ablations

B.1 Encoder implementation details

All subject encoders use the same hyperparameters. For the linear stage, we use $d_{\mathrm{PCA}} = 768$ PCA components and an embedding dimensionality of $d = 128$. The nonlinear residual refinement is a one-hidden-layer MLP with hidden size $d_{\mathrm{hidden}} = 768$. The residual scaling parameter α is learned jointly with the MLP parameters and converges to $\alpha \approx 0.4$ across subjects.

The MLP is optimized with Adam for 2000 training steps using the combined objective described in the main text, with $\lambda_{\mathrm{NCE}} = 1$ and $\lambda_{\mathrm{pull}} = 0.5$. Hyperparameters were selected by within-subject retrieval across repeated presentations, balancing Mean Rank and R@1. Unless stated otherwise, all experiments use these settings, yielding 128-dimensional subject embeddings.

Table S2: **Encoder component ablation**. Within-subject retrieval across repeated presentations after removing components of the encoder. Results are averaged across subjects and repetition pairs. Lower is better for Mean Rank; higher is better for R@1 and RSA.

Configuration	Reliability	PCA	MCCA	Nonlinear	Mean Rank ↓	R@1 ↑	RSA ↑
Reliability + PCA	✓	✓	–	–	116.52	0.09	0.13
PCA + MCCA	–	✓	✓	–	77.31	0.20	0.27
Reliability + PCA + MCCA	✓	✓	✓	–	11.73	0.73	0.23
PCA + nonlinear	–	✓	–	✓	14.05	0.60	0.55
PCA + MCCA + nonlinear	–	✓	✓	✓	14.05	0.60	0.54
Reliability + PCA + nonlinear	✓	✓	–	✓	21.89	0.43	0.51
Full encoder	✓	✓	✓	✓	**5.28**	**0.82**	**0.53**

B.2 Encoder component ablation.

We ablated the main components of the encoder by removing reliability weighting, MCCA, and the nonlinear residual refinement. Table S2 shows that the full encoder gives the best retrieval performance. The nonlinear refinement contributes strongly to RSA and retrieval, while the full combination of reliability weighting, PCA, MCCA, and nonlinear refinement yields the best Mean Rank and R@1.

B.3 Encoder dimensionality sensitivity

We evaluated within-subject retrieval across a hyperparameter grid varying PCA dimensionality $d_{\mathrm{PCA}} \in \{512, 768, 1024, 1280\}$, embedding dimensionality $d \in [64, 384]$, hidden size $d_{\mathrm{hidden}} \in \{128, 256, 512, 768, 1024\}$, and MLP depth $\in \{1, 2, 3\}$. Fig. S2 summarizes the search by plotting Mean Rank as a function of d, with separate curves for d_{PCA} and panels for MLP depth. Since d_{hidden} is not shown, each point reports the best Mean Rank across hidden sizes for that combination of d, d_{PCA}, and depth. Overall, a single hidden layer performs best, and the strongest stable configurations use $d_{\mathrm{PCA}} = 768$ or 1024 with embedding dimensionality around $d = 100$–150.

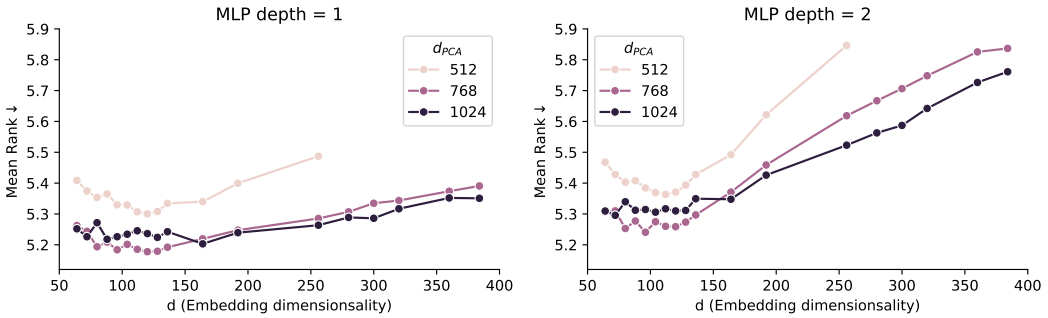


Figure S2: **Encoder dimensionality sensitivity.** Mean Rank across encoder dimensionalities. Lower is better. Each point reports the best value across d_{hidden} sizes. Mean Rank averaged across subjects.

C Unsupervised selection

The pairwise translation procedure (Section 3.2) is sensitive to its initialization. The pseudo-correspondences obtained through centroid matching depend on the K-means initialization[7]. Different random seeds produce different initial rotations and converge to different local solutions, some of which correspond to poor alignments.

To address this sensitivity without using paired samples or evaluation labels, we run the pairwise translation procedure with C random seeds. For every ordered subject pair (s, t), this produces a set of candidate rotations

$$\mathcal{R}_{s\to t} = \left\{ R^{(c)}_{s\to t} \right\}_{c=1}^{C}, \qquad R^{(c)}_{s\to t} \in O(d). \tag{S1}$$

We select one candidate per subject pair using cycle consistency. For three distinct subjects s, t, p, a coherent set of rotations should satisfy

$$R_{s\to t} R_{t\to p} \approx R_{s\to p}. \tag{S2}$$

We measure agreement between two orthogonal transformations A and B using their normalized Frobenius inner product,

$$s_{\mathrm{rot}}(A, B) = \frac{1}{d} \operatorname{tr}(A^\top B). \tag{S3}$$

This score is 1 for identical transformations and is expected to be close to 0 for unrelated rotations. Let $c_{st} \in \{1, \dots, C\}$ denote the candidate selected for pair (s, t), and let $\mathbf{c} = \{c_{st}\}_{s \neq t}$ denote a complete assignment of candidates to all ordered subject pairs. We define the global cycle-consistency objective as

$$J(\mathbf{c}) = \frac{1}{|\mathcal{T}|} \sum_{(s,t,p) \in \mathcal{T}} s_{\mathrm{rot}} \left(R_{s \to t}^{(c_{st})} R_{t \to p}^{(c_{tp})}, R_{s \to p}^{(c_{sp})} \right), \tag{S4}$$

where $\mathcal{T}$ contains all ordered triplets of distinct subjects.

We optimize this discrete objective using coordinate ascent [59]:

1. Randomly select one candidate for every ordered subject pair.
2. Visit each ordered pair and select the candidate that maximizes J, while keeping all other selections fixed.
3. Repeat these sweeps until no selection changes.
4. Repeat steps 1–3 from M random initializations and retain the selection with the largest value of J.

The criterion depends exclusively on the candidate rotation matrices and does not use training embeddings, evaluation embeddings, image identities, or retrieval performance. The resulting rotations, $\left\{ R_{s \to t}^{(c^\star_{st})} \right\}_{s \neq t}$, are used as the pairwise translations in all subsequent analyses.

C.1 Unsupervised selection stability

To quantify the stability of the unsupervised selection algorithm, we first ran the pairwise translation procedure with $C = 10$ random seeds, obtaining 10 candidate rotations for each ordered subject pair. Keeping these candidate sets fixed, we ran the coordinate-ascent selection algorithm 100 times with different global random seeds, using $M = 100$ random restarts in each run. We evaluated Mean Rank and Recall@1 for the selected pairwise rotations, both before and after the spectral synchronization described in Section 3.3. The evaluation metrics were used only to characterize the resulting solutions and were not accessed by the selection algorithm. The same selection of 56 directed rotations was recovered in 63 of the 100 runs. All main results use this most frequently recovered selection.

Table S3: Stability of the unsupervised rotation-selection algorithm over 100 runs. Values in the first row report the mean and standard deviation across runs.

Statistic	Mean Rank (pairwise) ↓	Recall@1 (pairwise) ↑	Mean Rank (synchronized) ↓	Recall@1 (synchronized) ↑
Mean ± std	**2.7046 ± 0.0245**	**0.7800 ± 0.0003**	**2.1038 ± 0.0064**	**0.8213 ± 0.0004**
Median	2.7002	0.7801	2.1080	0.8215
Most frequent solution (63/100)	2.7002	0.7801	2.1080	0.8215
Minimum	2.6694	0.7794	2.0799	0.8203
Maximum	2.7725	0.7804	2.1080	0.8215

D Experimental details

This section reports additional details to reproduce the evaluation and main results.

D.1 Within-subject encoder baseline details

To assess whether encoder improvements were consistent across subjects, Table S4 reports Mean Rank for each within-subject baseline separately for all subjects (extending Table 2).

We evaluated model-guided baselines by fitting a linear ridge regression from PCA-reduced fMRI responses (768 components) to intermediate model representations, and then measuring within-subject retrieval across repeated presentations. Vision embeddings were extracted from the last

Table S4: **Per-subject within-subject retrieval across encoder baselines.** Mean Rank for within-subject retrieval across repeated presentations, reported separately for each subject and averaged across repetition pairs. Lower is better; chance Mean Rank is 258.

Method	S1	S2	S3	S4	S5	S6	S7	S8	Avg.
Preprocessed fMRI (GLMsingle [20])	114.48	122.62	155.37	154.53	118.40	150.04	155.55	182.19	144.15
PCA	54.28	35.95	98.65	83.27	68.03	86.60	93.09	104.93	78.10
PCA + SplitAE [30]	60.99	58.23	104.29	101.37	65.04	105.20	113.08	142.11	93.78
PCA + DeepCCA [31]	1.94	4.72	31.67	22.27	15.75	22.20	30.93	56.18	23.21
PCA + Multiview CCA [24]	1.40	2.79	27.95	19.78	9.57	17.71	19.72	54.92	19.23
DINOv2 ViT-S/14 [32] + Ridge	11.43	16.42	28.99	26.76	13.12	20.75	29.60	51.52	24.82
CLIP ViT-B/16 [33] + Ridge	12.30	16.88	27.1	24.86	11.71	18.1	28.25	46.66	23.23
ViT AugReg-L/16 [34] + Ridge	7.93	12.9	24.51	22.98	10.22	15.4	24.37	44.12	20.30
all-MiniLM-L6-v2 [35] + Ridge	18.67	24.80	35.63	31.97	18.24	25.45	34.00	55.00	30.47
Ours (linear)	1.09	1.20	20.55	13.4	4.54	8.42	15.38	29.26	11.73
Ours (full)	**1.03**	**1.12**	**11.25**	**4.69**	**1.77**	**3.09**	**7.60**	**11.67**	**5.28**

layer of pretrained models applied to the NSD images using the TIMM library[60], while language embeddings were extracted from the COCO captions associated with each image using sentence-transformers[35]. Language models follow those used in [7], and vision models were selected from [41]. A complete list of model identifiers is provided as a Hugging Face collection[2]. The main text reports representative models from the main model families in Table 2. Table S5 provides the detailed per-subject evaluation for the larger set of model representations tested.

Table S5: **Per-subject model-guided ridge baselines.** Mean Rank for within-subject retrieval across repeated presentations using ridge mappings from fMRI responses to pretrained vision- and language-model embeddings. Results are reported separately for each subject and averaged across subjects. Lower is better; chance Mean Rank is 258.

Model	S1	S2	S3	S4	S5	S6	S7	S8	Avg.
Vision models									
DINOv2 ViT-S/14 [32]	11.43	16.42	28.99	26.76	13.12	20.75	29.60	51.52	24.82
DINOv2 ViT-B/14	19.11	24.55	35.68	38.05	20.33	32.07	40.18	65.86	34.48
DINOv2 ViT-L/14	18.57	24.35	38.20	39.97	19.70	32.75	42.43	73.26	36.15
DINOv2 ViT-G/14	26.51	30.49	44.01	46.19	24.07	39.55	49.98	77.67	42.31
AugReg ViT-T/16 [34]	14.33	17.77	29.29	27.42	14.32	20.53	30.00	48.51	25.27
AugReg ViT-S/16	11.87	17.66	27.84	26.57	12.26	19.31	29.49	48.57	24.20
AugReg ViT-B/16	8.63	13.95	25.15	23.07	10.99	16.27	25.25	43.62	20.87
AugReg ViT-L/16	**7.93**	**12.90**	**24.51**	**22.98**	**10.22**	**15.40**	**24.37**	**44.12**	**20.30**
MAE ViT-B/16 [61]	18.67	25.38	40.77	35.04	20.81	30.20	41.84	59.57	34.03
MAE ViT-L/16	12.63	17.72	32.39	27.66	14.20	22.66	33.31	50.08	26.33
MAE ViT-H/14	13.23	18.17	32.98	27.89	14.12	23.25	34.80	49.73	26.77
CLIP ViT-B/16 FT [33]	17.97	23.00	34.80	34.59	17.65	26.87	36.85	61.08	31.60
CLIP ViT-L/14 FT	14.08	20.36	31.19	30.23	14.18	23.15	33.49	56.79	27.93
CLIP ViT-H/14 FT	16.70	23.27	35.24	33.82	16.89	26.59	38.52	62.65	31.71
CLIP ViT-B/16 [33]	12.30	16.88	27.10	24.86	11.71	18.10	28.25	46.66	23.23
CLIP ViT-L/14	12.78	17.99	28.57	25.85	12.36	19.14	29.21	48.99	24.36
CLIP ViT-H/14	10.86	15.74	26.51	23.78	11.36	17.54	26.88	46.87	22.44
Language models									
E5-small [62]	24.19	30.91	42.11	38.52	21.75	31.60	42.15	65.67	37.11
E5	22.92	30.18	41.78	38.04	20.71	31.02	41.41	65.26	36.41
Stella	22.34	29.00	39.95	36.62	19.86	29.19	39.28	63.24	34.93
Granite Embedding-small [63]	22.50	27.69	38.77	34.96	19.04	29.05	38.55	60.79	33.92
Granite Embedding	21.78	27.77	38.80	34.80	19.38	28.58	37.63	61.37	33.76
GTR [64]	19.49	24.90	36.21	33.78	16.95	26.84	35.83	58.75	31.59
all-MiniLM-L6-v2	18.67	24.80	35.63	31.97	18.24	25.45	34.00	55.00	30.47

[2]https://huggingface.co/collections/pablomm/platonic-representations-in-the-human-brain

D.2 Cross-subject alignment baselines

All baselines for the cross-subject alignment task use repetition-averaged representations and are evaluated on the same 515 shared images. Results are averaged across the 56 directed subject pairs.

PCA without alignment. We independently fit 768-component PCA representations for each subject and compare them directly, with components ordered by decreasing explained variance. No cross-subject mapping is fitted.

Optimal transport. We use entropic Gromov–Wasserstein matching from the Python Optimal Transport (POT) library [65]. The method computes within-subject distance matrices and estimates a transport plan that matches their relational structure across subjects. We apply it to the complete evaluation sets of PCA representations and learned embeddings, without providing the true image correspondences.

Model-guided ridge regression. For each subject, ridge regression maps neural responses to model features of its own training images. Cross-subject retrieval is then evaluated in the common model feature space. These baselines use unpaired training images but require external image or caption representations, extracted as described above.

Supervised alignment protocol. The following methods use our fixed 128-dimensional embeddings and up to 485 additional shared training images per subject pair. The 515 evaluation images are excluded from fitting these mappings, and the encoders remain trained only on non-overlapping images.

Orthogonal Procrustes (supervised). We fit an orthogonal mapping that minimizes squared error between paired source and target embeddings, using singular value decomposition. The mapping is then applied to source evaluation embeddings.

Cross-validated ridge regression (supervised). We fit an affine source-to-target mapping with an ℓ_2 penalty on its weights. Regularization is selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.2, 0.5, 1\}$ by five-fold cross-validation minimizing mean squared error, with shuffled folds and random seed 42. The selected model is refitted on all paired training images.

Canonical correlation analysis (CCA) (supervised). We use regularized CCA from `mvlearn`, with 128 components and regularization parameter 0.3. Separate subject projections are fitted to paired training embeddings, and retrieval is evaluated between their projected evaluation embeddings.

Linear regression. We fit an unregularized affine source-to-target mapping by least squares and apply it to source evaluation embeddings.

D.3 Compute requirements

All experiments were run on a single workstation with an NVIDIA A40 GPU (48 GB) and an AMD Ryzen 9 CPU. The experiments are not computationally intensive: training the linear encoder for one subject takes approximately 2 minutes on CPU, training the nonlinear residual MLP takes approximately 1 minute on GPU, and estimating one pairwise brain-to-brain rotation takes approximately 5 minutes on CPU.

E Statistical support and robustness

We assess statistical significance and uncertainty in cross-subject retrieval (Fig. 3) using image-identity permutation tests, image bootstrapping, and subject-deletion sensitivity. All analyses use the synchronized rotations defining the shared space reported in Section 4.3, with encoders and rotations held fixed. We evaluate 515 held-out shared images across 56 directed pairs of eight NSD participants and average retrieval metrics across pairs.

E.1 Image-identity permutation tests

We tested retrieval against chance using $10{,}000$ random permutations of image identities on one side of each comparison. Each permutation was shared across all directed subject pairs, and metrics were averaged across pairs. No permutation yielded a Mean Rank as low as observed or an R@1 as high

Table S6: **Image-level statistical support for synchronized retrieval.** Global metrics are averaged across 56 directed subject pairs. Null and bootstrap summaries report mean ± standard deviation across 10,000 permutations or resamples. Confidence intervals are percentile image-bootstrap intervals.

Metric	Observed	Permutation null	Bootstrap	95% CI
R@1 ↑	0.8215	0.00192 ± 0.00163	0.8215 ± 0.0068	$[0.8079, 0.8350]$
Mean Rank ↓	2.108	258.011 ± 5.191	2.107 ± 0.118	$[1.893, 2.357]$

as observed. Using the one-sided permutation p-value $(k+1)/(B+1)$, where k counts equally or more extreme results, both metrics gave $p < 10^{-4}$ (Table S6).

E.2 Subject-deletion sensitivity

We assessed sensitivity to individual participants by excluding each subject's evaluation pairs and averaging metrics over the remaining 42 directed pairs, with the fitted pipeline held fixed. We computed approximate 95% jackknife intervals using subject-deletion standard errors and a t distribution with seven degrees of freedom. Mean Rank ranged from 1.909 to 2.261 and R@1 from 0.8032 to 0.8492, showing strong retrieval after each exclusion (Table S7).

Table S7: **Subject-deletion sensitivity.** Metrics exclude each subject's evaluation pairs in turn, keeping the fitted eight-subject shared space fixed. Intervals are approximate 95% jackknife intervals.

Metric	Full result	Subject-deletion range	95% interval
R@1 ↑	0.8215	0.8032–0.8492	$[0.7337, 0.9093]$
Mean Rank ↓	2.108	1.909–2.261	$[1.342, 2.874]$

E.3 Image-bootstrap confidence intervals

To quantify uncertainty due to image sampling, we computed $10{,}000$ bootstrap resamples of 515 query-image indices with replacement, sharing sampled indices across subject pairs while keeping the candidate set fixed. The 95% percentile intervals were $[1.893, 2.357]$ for Mean Rank and $[0.8079, 0.8350]$ for R@1 (Table S6).

Table S8 reports 95% image-bootstrap confidence intervals for each directed subject pair. Pairwise Mean Rank ranged from 1.04 to 4.56, and all intervals remained below the chance value of 258.

Table S8: **Pairwise cross-subject retrieval uncertainty.** Each cell reports Mean Rank above its pointwise 95% image-bootstrap confidence interval. Lower Mean Rank is better; chance is 258.

Subject	S1	S2	S3	S4	S5	S6	S7	S8
S1	—	1.04 [1.02, 1.06]	1.44 [1.18, 1.81]	1.68 [1.44, 1.98]	4.19 [3.44, 5.02]	3.39 [2.35, 4.70]	1.36 [1.22, 1.52]	1.93 [1.65, 2.25]
S2	1.04 [1.02, 1.06]	—	1.23 [1.15, 1.32]	1.76 [1.48, 2.09]	4.56 [3.69, 5.54]	3.32 [2.52, 4.25]	1.57 [1.31, 1.90]	1.93 [1.54, 2.52]
S3	1.44 [1.18, 1.84]	1.23 [1.14, 1.32]	—	1.30 [1.17, 1.44]	2.31 [1.82, 2.93]	2.17 [1.64, 2.81]	1.35 [1.20, 1.54]	1.91 [1.46, 2.60]
S4	1.72 [1.44, 2.06]	1.65 [1.39, 1.96]	1.36 [1.22, 1.53]	—	1.88 [1.60, 2.21]	1.70 [1.42, 2.02]	1.33 [1.20, 1.46]	2.18 [1.74, 2.72]
S5	3.52 [2.92, 4.20]	3.92 [3.20, 4.77]	1.88 [1.52, 2.37]	1.61 [1.40, 1.85]	—	1.52 [1.22, 1.99]	2.17 [1.80, 2.58]	2.86 [2.16, 3.85]
S6	3.24 [2.23, 4.48]	3.24 [2.45, 4.21]	2.09 [1.64, 2.64]	1.75 [1.46, 2.13]	1.70 [1.39, 2.11]	—	2.10 [1.60, 2.75]	2.75 [2.18, 3.38]
S7	1.21 [1.12, 1.33]	1.42 [1.21, 1.68]	1.31 [1.15, 1.50]	1.22 [1.13, 1.32]	2.28 [1.88, 2.74]	1.82 [1.43, 2.36]	—	2.32 [1.71, 3.11]
S8	2.11 [1.74, 2.57]	2.07 [1.61, 2.77]	2.09 [1.58, 2.84]	2.68 [2.08, 3.39]	3.47 [2.64, 4.51]	3.04 [2.40, 3.74]	2.72 [2.05, 3.59]	—

F Qualitative Examples

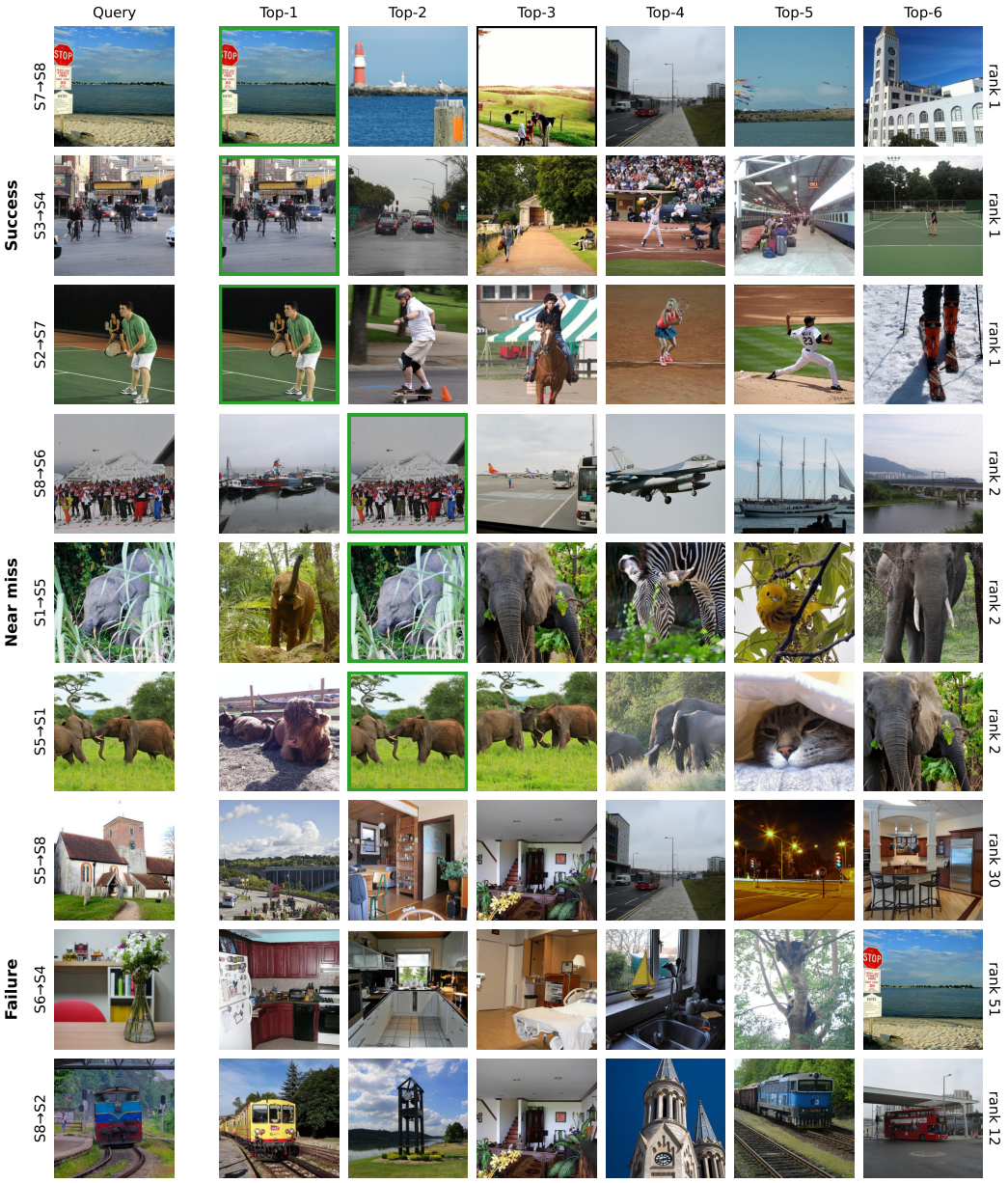


Figure S3: **Cross-subject retrieval examples in the shared space.** Each row shows a query image and the top six images retrieved from 515 candidates for the indicated source→target subject pair. The correct image is outlined in green when shown; its rank appears on the right. Examples were sampled uniformly within three strata across all 56 directed pairs using a fixed random seed: successes (rank 1; 82.2%), near misses (ranks 2–3; 10.4%), and failures (rank $>$ 10; 2.6%).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the paper's main contributions: subject-specific fMRI embedding learning, unsupervised orthogonal brain-to-brain translation, and synchronization into a shared latent space. These claims are directly evaluated in the experimental sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Discussion describes the main limitations, including reliance on high-quality fMRI data with repeated stimulus presentations, uncertainty about generalization to lower-SNR modalities or other cognitive domains, and sensitivity of the unpaired translation procedure, based on mini-vec2vec, to initialization and seed selection.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical and does not present new theoretical results, theorems, or formal proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The Methods and Experiments describe the encoder, pairwise translation, shared-space synchronization, evaluation protocol, dataset splits, metrics, and baselines used for the main results. Additional preprocessing details, ablations, and seed-selection details are provided in the Supplementary Material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use the publicly available Natural Scenes Dataset. We include supplementary code with the submission to reproduce the main analyses at https://github.com/memory-formation/platonic-representations-fmri.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The Experiments section specifies the train/test split, evaluation set, metrics, and main hyperparameters used for the reported results. Additional preprocessing, implementation, and hyperparameter details are provided in the Supplementary Material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We report standard deviations across subjects or directed subject pairs, with per-subject results in the Supplementary Material. Appendix E provides image-identity permutation tests, image-bootstrap confidence intervals, and subject-deletion sensitivity analyses for synchronized cross-subject retrieval.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The Supplementary Material provides details on the CPU/GPU resources used. All experiments were run on a single workstation and are not compute-intensive relative to current deep learning models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The human fMRI data are from the publicly available Natural Scenes Dataset, which was collected under appropriate ethical approval and consent procedures, and our work uses only the released anonymized data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential positive impacts for cross-subject neural modeling, decoding, transfer learning, and brain–model comparison, as well as risks related to neural privacy and misuse of brain decoding methods. We emphasize that the current work is basic research on anonymized fMRI data and is not a deployed system.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk model, scraped dataset, or deployed system. The experiments use anonymized public fMRI data, and released code is intended only to reproduce the research analyses.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used in the paper, including the fMRI dataset, pretrained model representations, and software packages, are credited through the appropriate citations. Dataset and preprocessing details, including access information, are provided in the paper and Supplementary Material.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide code with the submission to reproduce the main analyses. The code is available at https://github.com/memory-formation/platonic-representations-fmri

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not collect new crowdsourcing or human-subject data. We use the publicly released Natural Scenes Dataset, whose participant instructions, consent procedures, and acquisition details are documented in the original dataset publication.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not collect new human-subject data. We use the publicly released Natural Scenes Dataset, whose ethical approval, consent procedures, and participant-risk information are documented in the original dataset publication.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development and experiments do not use LLMs as an important, original, or non-standard component. Any LLM use was limited to writing, editing, or formatting assistance and did not affect the scientific methodology or results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.